Pós-Graduação em Ciência da Computação

MARCELO IURY DE SOUSA OLIVEIRA

A Metadata Curation Framework for Data Ecosystems

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2019

MARCELO IURY DE SOUSA OLIVEIRA

A Metadata Curation Framework for Data Ecosystems

A Ph.D. Thesis presented to the Center for Informatics of Federal University of Pernambuco in partial fulfillment of the requirements for the degree of Philosophy Doctor in Computer Science.

Concentration Area: Database

Supervisor: Bernadette Farias Lóscio

Recife
2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

O48m Oliveira, Marcelo Iury de Sousa
A metadata curation framework for data ecosystems / Marcelo Iury de Sousa Oliveira. – 2019.
287 f.: il., fig., tab.

Orientadora: Bernadette Farias Lóscio.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2019.
Inclui referências e apêndices.

1. Banco de dados. 2. Metamodelos. 3. Gestão de metadados. I. Lóscio, Bernadette Farias (orientadora). II. Título.

025.04 CDD (23. ed.) UFPE- MEI 2019-070

Marcelo Iury de Sousa Oliveira

A Metadata Curation Framework for Data Ecosystems

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Aprovado em: 22/02/2019.

_____

Orientadora: Profa. Dra. Bernadette Farias Lóscio

BANCA EXAMINADORA

_____

Profa. Dra. Ana Carolina Brandão Salgado
Centro de Informática /UFPE

_____

Profa. Dra. Carina Frota Alves
Centro de Informática /UFPE

_____

Profa. Dra. Sandra de Albuquerque Siebra
Departamento de Ciência da Informação / UFPE

_____

Profa. Dra. Cláudia Maria Fernandes Araújo Ribeiro
Instituto Federal do Rio Grande do Norte / Campus Natal

_____

Prof. Dr. José Viterbo Filho
Instituto de Computação / UFF

I dedicate this Thesis to my wife Vanessa, my mother Neuma and my father Ilom, who have always supported me and taught me to never give up my dreams.

ACKNOWLEDGMENTS

Que bela jornada foi esse doutorado. Foram cinco anos de muito trabalho, luta e também de amadurecimento. Aprendi bastante através da interação com colegas e professores maravilhosos que tive a felicidade de conhecer durante esses anos. Tenho muitas pessoas para agradecer. Todas contribuíram direta ou indiretamente com o meu trabalho das formas mais variadas possíveis.

Agradeço a minha família e, em especial, aos meus pais, Dona Neuma e Dr. Ilom, pela educação, pelo apoio, pelo carinho, pelo amor,..., enfim por um mundo de cuidados e educação que contribuíram para me tornar a pessoa que sou hoje. Nunca me esquecerei da correria e disponibilidade de minha mãe para levar-me a aulas de inglês, olimpíadas de biologia e química, feiras de ciências, estudo em casas de amigos (sempre fui um pouco nerd, né. :-)) e também para shows e micaretas (a gente não quer só comida; quer diversão, balé). Espero poder retribuir por toda essa dedicação. Obrigado Mãezinha e Paizinho!

À minha linda esposa Vanessa, por todo o seu amor, carinho, ajuda e compreensão durante a realização deste trabalho. Ela topou e me incentivou nas várias aventuras surgidas durante este doutorado, inclusive ficar nove meses distante da família e de nossos gatinhos. Viver estes anos ao seu lado tem sido maravilhoso, e você teve e tem um papel fundamental na minha vida. Espero seguirmos juntos até fim de minha vida.

Também tenho plena gratidão a minha orientadora, professora Bernadette, pelo apoio e direcionamento concedido a mim. Além de uma cearense arretada, é um exemplo de profissional, professora e guru. Sou grato pelas inúmeras oportunidades de aprendizado e de trabalho proporcionadas por ela ao longo desta jornada. Durante esse tempo, firmamos uma parceria que vai além da pesquisa e que contribuiu de forma geral para o meu crescimento profissional.

Durante o doutorado, convivi com outros professores que contribuíram para a minha formação. Em particular, cito a professora Ana Carolina Salgado, famosa Carol. Uma das pessoas mais educadas e finas que conheci. Tanto interagi com a professora Carol em grupos de estudo quanto fui seu aluno em uma disciplina do doutorado. Há inclusive diversas tentativas minha de ser considerado um de seus ex-alunos. Um dia conseguirei, quem sabe? :-)

Também agradeço aos professores Kiev Gama e Genoveva Vargas. Kiev foi meu co-orientador no início do doutorado. Por incompetência minha, não consegui manter essa relação. Mas, durante esse breve período aprendi bastante com Kiev. Seu pragmatismo foi um dos motivadores para o desenvolvimento do Waldo. Já a professora Genoveva, tive o prazer de ser supervisionado por ela durante o meu doutorado sanduíche em Grenoble. Ela parecia o Buckaroo Banzai. Dona de muitos talentos e conhecimentos. Sempre encontrava um caminho para inovação.

Aos professores Ana Carolina Salgado, Cláudia Ribeiro, Sandra Siebra, Carina Frota e José Viterbo, agradeço pela disponibilidade em participar da banca de defesa do doutorado e pelas contribuições somadas a este trabalho.

Ao Lairson, um grande amigo que tive a sorte de conhecer durante o doutorado. Uma pessoa de coração enorme! Sempre pronto para ajudar quem quer que seja. Ele colaborou diretamente comigo em diversos trabalhos, além de ter participado em longos e frequentes momentos de debates e discussões acadêmicas. Tal qual a professora Bernadette, espero continuar essa nossa parceria e amizade por muitos anos.

Aos integrantes do grupo Aladin, que sempre me brindaram com suas opiniões, críticas e sugestões. Em especial, sou grato aos integrantes da primeira geração do grupo, dos quais cito Helton, Glória, Wilker, Karina e Rayele. Helton e Glória trabalharam comigo mais de perto na construção da base teórica necessária para a construção do Louvre, e através da ajuda deles pude desenvolver o meu trabalho com mais rapidez e firmeza. Muito obrigado.

Também tenho sinceros agradecimentos aos outros grupos de pesquisa que participei durante o doutorado. Alguns foram extintos, mas através de todos pude estender meu conhecimento através da interação com pessoas brilhantes. Assim, agradeço aos colegas da Toca dos Dados, dos quais cito (Super) Diego, Danusa, Everaldo Neto, Paulo Orlando, Priscilla, Natacha e Gabrielle Canalle. Também agradeço aos colegas do projeto Basic, em especial a Herbertt Diniz e Emanuel Carneiro. Herbertt foi um parceirão, tendo me acompanhado em viagens a congressos e ajudado nas avaliações das soluções desenvolvidas durante o doutorado.

Dentre os integrantes desses grupos de pesquisa, em especial, sou muito grato a Gabriele Canalle, uma paranaense maravilhosa que cuidou dos meus gatinhos pelo tempo que estive fora do Brasil. Nos tornamos grande amigos, e, quem sabe, em um futuro próximo, serei seu padrinho de casamento. Até hoje, os meus gatinhos "perguntam" pela Tia Gaby. Gabrielle também me ajudou a conduzir a última avaliação deste trabalho. Valeu, guria.

Aos meus amigos Daniel Nº 2 e Suzanny que trouxeram motivação e apoio diante alguns obstáculos que surgiram durante o doutorado. Não posso esquecer também do casal maravilhoso Paulo e Mauricea, pais de Suzanny, que também me apoiaram durante alguns momentos do doutorado.

Aos vizinhos que tive sorte de conviver durante o doutorado. Em Recife, foram Manoel Santos e Olívia. Em Grenoble, foram Pascal e Eunice. Ambas as famílias me deram muito apoio durante adaptação em Recife e Grenoble. Sempre socorrendo nos aperreios e ajudando a descontrair nos momentos de cansaço e stress. Inclusive, Olívia e Manoel ajudaram a cuidar da minha casa e dos meus gatinhos sempre que precisei me ausentar de Recife.

Na temporada europeia, também recebi ajuda de minhas tias Rejane e Ray. Duas tias lindas que me deram carinho quando estive muito carente da família, além de terem me

recebido nas suas casas.

Também não posso esquecer de outras pessoas importantes da minha família, como meus irmãos Izabelle e Ilom, minha sogra "Franchesca" Gelba, e vários primos, tios e outros parentes que me deram suporte nos mais variados momentos de minha vida acadêmica.

Também sou grato aos participantes das pesquisas empíricas que realizei, a quem não devo aqui nomear por acordo de sigilo de suas identidades. Muito obrigado pela disponobilidade e colaboração com meu trabalho. Sem ajuda de vocês não teria conseguido validar minhas soluções.

Aos meus colegas professores do curso de Bacharelado em Sistemas de Informação da Unidade Acadêmica de Serra Talhada que seguraram a barra enquanto estive afastado. Em especial, agradeço ao professor Leonardo Mendes que me auxiliou na análise estatística de algumas avaliações.

Também agradeço a várias outras pessoas que me ajudaram de alguma forma durante o doutorado. Em especial, ao professor Carlos Texeira que me deu suporte quando tive que viajar para assistir aulas em Recife. Ao Cleidson e à família Brito que davam toda atenção e companhia a minha esposa quando tinha que viajar. À Edileide e Danielle por cuidarem da nossa casa e dos nossos gatinhos, e serem nossas "anjas" protetoras em Serra Talhada.

Conheci algumas pessoas durante a temporada recifense que certamente tornaram para a minha adaptação à cidade mais fácil. Em especial, cito Renan (Bibiu) Paiva, Wanessa Botelho, Seu Bolinha, Thaiz Burity, Jorge Correa Neto, Marcos Antonio Eugênio, Filipe, Jean, Luiz, Lewson, Klissiomara, e ao casal Jô de Marquinhos e Marquinhos de Jô. Alguns desses também me hospedaram em suas casas enquanto ainda enfrentava a viagem semanal Serra-Recife.

Ao Maguila, à Nina e ao Palito, muito obrigado por fazerem parte de minha vida.

Finalmente, a vários outros que certamente posso ter esquecido neste momento, meus sinceros agradecimentos.

ABSTRACT

A Data Ecosystem can be defined as a complex socio-technical network that enables collaboration between autonomous actors in order to explore data. Such ecosystems provide an environment for creating, managing and sustaining data sharing initiatives. While Data Ecosystems are thus arguably gaining importance, several ecosystems are not sustainable and consequently the effort spent by their actors end up not being properly used or forgotten. A comprehensive and meaningful description of all Data Ecosystem resources is needed. The increasing recognition of metadata as an essential asset had motivated an increased demand for metadata curation solutions. Metadata curation covers the creation and harvesting of metadata, appraisal and selection of metadata, quality assurance, preservation of metadata and other lifecycle stages, and also involves a number of IT systems. While important, in general, the current initiatives on metadata curation are a confusing mixture of activities, standards, terms and vocabularies, methods and tools. Referential guidelines would provide a basis to choose standard terms and definitions, processes and practices, roles and deliverables for metadata curation practitioners. In this context, this thesis aims to propose a framework, called Louvre, which offers a wide range of processes for aiding to curate metadata in Data Ecosystems. Each process describes a coherent set of engineering and management activities related to metadata curation. The Louvre structure is flexible and may be adapted to the needs of the actors interested in curating Data Ecosystem metadata. In this sense, processes are organized in functional dimensions, enabling modularization of the framework. Louvre also provides a set of best practices aligned with principles of agile and open collaboration for managing curation work through the collaborative effort of self-organizing actors. Finally, the framework is based on state-of-the-art in the area. This research also contributes to the Data Ecosystem area, by mapping the state-of-the-art of Data Ecosystems. In addition, it contributes also to the understanding of several issues related to the Data Ecosystems creation and maintenance. Also noteworthy is the definition, formalization and modelling of essential constructs related to Data Ecosystems.

Keywords: Data Ecosystem. Meta-model. Metadata Management. Metadata Curation.

RESUMO

Um Ecossistema de Dados pode ser definido como uma rede sociotécnica complexa que permite a colaboração entre atores autônomos para explorar dados. Esses ecossistemas fornecem um ambiente para criar, gerenciar e sustentar iniciativas de compartilhamento de dados. Embora os Ecossistemas de Dados estejam ganhando importância, vários ecossistemas não são sustentáveis e, consequentemente, o esforço despendido por seus atores acaba não sendo adequadamente usado ou esquecido. É necessária uma descrição abrangente e significativa dos recursos do Ecossistema de Dados. O crescente reconhecimento dos metadados como um ativo essencial motivou uma demanda crescente por soluções de curadoria de metadados. A curadoria de metadados abrange a criação e a coleta de metadados, avaliação e seleção de metadados, garantia de qualidade, preservação de metadados e outras etapas do ciclo de vida. Assim como, a curadoria de metadados também envolve o uso de vários sistemas e ferramentas de gestão e preservação de metadados. Embora importantes, as atuais iniciativas de curadoria de metadados são uma mistura confusa de atividades, padrões, termos e vocabulários, métodos e ferramentas. As guidelines e modelos de referencia poderiam forcener uma base para escolher termos e definições padrão, processos e práticas, papéis e resultados para os profissionais de curadoria de metadados. Neste contexto, esta tese tem como objetivo propor um framework, denominado Louvre, que oferece uma ampla gama de processos para auxiliar na organização de metadados em Ecossistemas de Dados. Cada processo descreve um conjunto coerente de atividades de engenharia e gerenciamento relacionadas à curadoria de metadados. A estrutura do Louvre é flexível e pode ser adaptada às necessidades dos atores interessados em realizar a curadoria de metadados. Nesse sentido, os processos são organizados em dimensões funcionais, possibilitando a modularização do framework. O Louvre também fornece um conjunto de práticas recomendadas alinhadas com princípios de desenvolvimento ágil e colaboração aberta para gerenciar o trabalho de curadoria através do esforço colaborativo de atores auto-organizados. Esta pesquisa também contribui para a área de Ecossistemas de Dados, mapeando o estado da arte da área. Além disso, este trabalho ainda contribui para o entendimento de várias questões relacionadas à criação e manutenção de Ecossistemas de Dados. Destaca-se também a definição, formalização e modelagem de constructos essenciais relacionados a Ecossistemas de Dados.

Palavras-chaves: Ecossistemas de Dados. Metamodelos. Gestão de Metadados. Curadoria de Metadados.

LIST OF FIGURES

Figure 1 – Research Methodology Steps. Source: Author 26
Figure 2 – Design Science Research Lifecycle. Source: Author 29
Figure 3 – Metadata Curation Environment. Source: Author 29
Figure 4 – DCC Curation Lifecycle Model. Source:(HIGGINS, 2008) 34
Figure 5 – DDI Combined Lifecycle Model. Source:(DDI, 2012) 35
Figure 6 – DataONE Lifecycle Model. Source:(ALLARD, 2012) 37
Figure 7 – OAIS Functional Mode. Source:(LEE, 2010) 39
Figure 8 – DMBOK functions. Source:(MOSLEY et al., 2010) 42
Figure 9 – Systematic literature mapping process. Adapted from: (LOPEZ-HERREJON; LINSBAUER; EGYED, 2015) 49
Figure 10 – Keyword-based query used to automate for search studies. Source: Author 51
Figure 11 – Distribution of the number of studies published regarding Data Ecosystems between 2011 and 2016. Source: Author 55
Figure 12 – Research design: construct-evaluate cycle. Adapted from:(SANTANA; ALVES, 2016) 80
Figure 13 – Meta-model for Data Ecosystems. Source: Author 93
Figure 14 – Meta-model main classes. Source: Author 94
Figure 15 – Meta-model: Actor, Relationship and related classes. Source: Author . 95
Figure 16 – Meta-model: Role and related classes. Source: Author 96
Figure 17 – Meta-model: Resource and related classes. Source: Author 97
Figure 18 – The Twin Peaks model. Adapted from: (NUSEIBEH, 2001) 114
Figure 19 – Louvre Structure. Source: Author 118
Figure 20 – Louvre Framework Elements. Source: Author 123
Figure 21 – Louvre Dimensions. Source: Author 134
Figure 22 – Multiple iterative loops of a PDCA method. Source:(ROSER, 2016) . . . 141
Figure 23 – Experts Educational Degree. Source: Author 154
Figure 24 – Experts Working Context. Source: Author 154
Figure 25 – Experts Experience on Data Ecosystems. Source: Author 154
Figure 26 – Experts Knowledge about Data Ecosystems. Source: Author 154
Figure 27 – Experts Knowledge about Metadata Curation. Source: Author 155
Figure 28 – Experts Rating about Metadata Importance for Data Ecosystems. Source: Author 155
Figure 29 – Experts Classification of their Metadata Curation Practice. Source: Author 155
Figure 30 – Actors and Roles Evaluation Source: Author 173

Figure 31 – Agile Practices Evaluation
Source: Author 175
Figure 32 – Metadata Curation Analysis and Planning Dimension Evaluation
Source: Author 176
Figure 33 – Metadata Acquisition Dimension Evaluation
Source: Author 177
Figure 34 – Metadata Quality Management Dimension Evaluation
Source: Author 178
Figure 35 – Metadata Curation Monitoring and Controlling Dimension Evaluation
Source: Author 179
Figure 36 – Metadata Curation Monitoring and Controlling Dimension Evaluation
Source: Author 180
Figure 37 – Metadata Curation Platform Administration Dimension Evaluation
Source: Author 181
Figure 38 – Model created in the Case Study presented at Chapter 4. Source: Author 240

LIST OF TABLES

Table 2 – Design Science Research Artifacts Classification 24
Table 3 – Relation among data curation models 41
Table 4 – Sources and number of studies 52
Table 5 – Rationale for excluding papers 52
Table 6 – List of Selected Studies 54
Table 7 – Studies classified according to whether or not they offer an explicit definition for a Data Ecosystem. 55
Table 8 – Previous papers most cited by the studies selected 56
Table 9 – Roles of Data Ecosystem actors 61
Table 10 – Studies that describe a form of Data Ecosystem organizational structure 63
Table 11 – Theoretical foundations adopted by Data Ecosystems studies 72
Table 12 – Benefits expected from Data Ecosystems 73
Table 13 – Barriers and challenges for Data Ecosystems 75
Table 14 – Data Ecosystem Characteristics 83
Table 15 – Data Ecosystem Main Constructs 84
Table 16 – Overall average of evaluation results of information quality assessment questionnaire 103
Table 17 – Detailed average of evaluation results of information quality assessment questionnaire 104
Table 18 – Overall average of evaluation results of cognitive quality assessment questionnaire 105
Table 19 – Detailed average of evaluation results of cognitive quality assessment questionnaire 106
Table 20 – Main Elements, Influences and Inspirations 117
Table 21 – Louvre Framework - Alpha Version 120
Table 22 – Louvre Framework - Beta Version 121
Table 23 – Louvre Dimensions and Processes 135
Table 24 – Evaluation Results 156
Table 25 – Metadata Curation Analysis and Planning Dimension Evaluation 159
Table 26 – Metadata Acquisition Dimension Evaluation 161
Table 27 – Metadata Quality Management Dimension Evaluation 162
Table 28 – Metadata Preservation and Dissemination Dimension Evaluation 164
Table 29 – Metadata Curation Monitoring and Controlling Dimension Evaluation . 166
Table 30 – Metadata Curation Platform Administration Dimension Evaluation . . . 168
Table 31 – Focus Group Participants Characterization 171
Table 32 – Focus Group - General Evaluation of Louvre Framework 183

LIST OF ABBREVIATIONS AND ACRONYMS

AIMQ	Assessment Information Methodology Quality
DAMA	Data Management International
DCC	Digital Curation Centre
DDI	Data Documentation Initiative
DMBOK	Data Management Body of Knowledge
DWMS	Data on the Web Management System
EC	Exclusion Criteria
EMF	Eclipse Modeling Framework
GMF	Graphical Modeling Framework
ICT	Information and Communication Technologies
IoT	Internet of Things
ISO	International Organization for Standardization
MOF	Meta Object Facility
OAIS	Open Archival Information System
OGD	Open Government Data
OMG	Object Management Group
PDCA	Plan, Do, Check and Act
SME	Small and Medium Enterprises

CONTENTS

1 INTRODUCTION 18
1.1 MOTIVATION AND PROBLEM STATEMENT 21
1.2 OBJECTIVES 22
1.3 RESEARCH METHOD 23
1.3.1 Philosophical Stance 23
1.3.2 Research Objective 25
1.3.3 Research Variables Nature 25
1.3.4 Research Methodology Steps 26
1.4 OUT OF SCOPE 29
1.5 ORGANIZATION OF THE DOCUMENT 30

2 THEORETICAL BACKGROUND 31
2.1 BACKGROUND ON METADATA CURATION 31
2.2 DATA CURATION MODELS 33
2.2.1 DCC Curation Lifecycle Model 33
2.2.2 DDI Combined Lifecycle Model 35
2.2.3 DataONE Lifecycle Model 36
2.2.4 OAIS Reference Model 38
2.2.5 Summary of Data Curation Models 39
2.3 DATA MANAGEMENT FRAMEWORKS 41
2.3.1 DMBOK 42
2.3.2 ISO 8000 44
2.4 CONSIDERATIONS OF THE CHAPTER 46

3 STATE OF THE ART ON DATA ECOSYSTEMS 48
3.1 REVIEW PROTOCOL 49
3.1.1 Research Questions 49
3.1.2 Data Sources and Search Strategy 50
3.1.3 Selection of Studies 51
3.1.4 Data Extraction and Synthesis 53
3.2 FINDINGS 53
3.2.1 Data Ecosystems Research Evolution 53
3.2.2 Data Ecosystem Terminology 54
3.2.3 Data Ecosystem Characterization 57
3.2.4 Data Ecosystem Elements 58
3.2.5 Data Ecosystem Actors and Roles 59

3.2.6 Data Ecosystem Structure . . . 63
3.2.7 Data Ecosystems Creation . . . 65
3.2.8 Data Ecosystems Infrastructure . . . 66
3.2.9 Data Ecosystems Value Generation . . . 67
3.2.10 Data Ecosystem Management . . . 68
3.2.11 Data Ecosystems Privacy . . . 70
3.2.12 Data Ecosystems Theoretical Foundations . . . 71
3.2.13 Data Ecosystems Benefits and Barriers . . . 73
3.3 CONSIDERATIONS OF THE CHAPTER . . . 76

4 A META-MODEL FOR DATA ECOSYSTEMS . . . 78
4.1 META-MODEL DEVELOPMENT METHOD . . . 80
4.2 RUNNING EXAMPLE . . . 81
4.3 META-MODEL KNOWLEDGE ACQUISITION . . . 82
4.4 META-MODEL CONCEPTUALIZATION . . . 83
4.4.1 Resource . . . 85
4.4.2 Role . . . 86
4.4.3 Actor . . . 87
4.4.4 Relationship . . . 88
4.4.5 Data Ecosystem Properties . . . 89
4.5 META-MODEL FORMALIZATION . . . 90
4.5.1 Modelling practices and conventions . . . 90
4.5.2 Data Ecosystem Meta-model . . . 91
4.6 META-MODEL EVALUATION METHOD . . . 98
4.7 META-MODEL EVALUATION . . . 100
4.7.1 Case Study Design . . . 101
4.7.2 Case Study Context . . . 102
4.7.3 Evaluation Results . . . 103
4.8 DISCUSSION . . . 107
4.9 CONSIDERATIONS OF THE CHAPTER . . . 108

5 THE LOUVRE FRAMEWORK . . . 111
5.1 MOTIVATIONAL SCENARIO . . . 112
5.2 FRAMEWORK DEVELOPMENT METHOD . . . 113
5.3 LOUVRE: A METADATA CURATION FRAMEWORK . . . 115
5.3.1 Principles, Values, and Mission . . . 115
5.3.2 Structure . . . 117
5.3.3 Louvre Framework Versions . . . 119
5.4 OVERVIEW OF THE LOUVRE FRAMEWORK ELEMENTS . . . 122
5.4.1 Actors and Roles . . . 122

5.4.1.1 Contributors and Teams . . . 122
5.4.1.2 Roles . . . 124
5.4.2 Agile Practices . . . 125
5.4.2.1 User Story . . . 126
5.4.2.2 Persona . . . 127
5.4.2.3 Backlog . . . 127
5.4.2.4 Sprints . . . 128
5.4.2.5 Coordination Meetings . . . 129
5.4.2.6 Continuous Integration . . . 130
5.4.2.7 Continuous Refactoring . . . 131
5.4.2.8 Automated Tests . . . 131
5.4.2.9 Collective Ownership . . . 132
5.4.2.10 Burndown Chart . . . 133
5.4.3 Dimensions and Processes . . . 133
5.4.3.1 Metadata Curation Analysis and Planning Dimension . . . 135
5.4.3.2 Metadata Acquisition Dimension . . . 136
5.4.3.3 Metadata Quality Management Dimension . . . 137
5.4.3.4 Metadata Preservation and Dissemination Dimension . . . 137
5.4.3.5 Metadata Curation Coordination Dimension . . . 138
5.4.3.6 Metadata Curation Platform Administration Dimension . . . 139
5.4.4 Deployment Method . . . 140
5.5 REVISITING THE MOTIVATIONAL SCENARIO . . . 143
5.5.1 Metadata Curation Platform Implementation . . . 145
5.5.2 Metadata Creation . . . 146
5.5.3 Metadata Ingestion . . . 147
5.5.4 Summary of Motivational Scenario . . . 147
5.6 CONSIDERATIONS OF THE CHAPTER . . . 148

6 EVALUATION OF THE LOUVRE FRAMEWORK . . . 149
6.1 SURVEY BASED ON EXPERTS OPINION . . . 149
6.1.1 Survey Questionnaire . . . 150
6.1.2 Survey Population and Sample . . . 151
6.1.3 Survey Data Collection . . . 152
6.1.4 Survey Results . . . 153
6.1.4.1 Experts Characterization . . . 153
6.1.4.2 Louvre Evaluation . . . 155
6.1.5 Summary of the findings . . . 169
6.2 FOCUS GROUP . . . 170
6.2.1 Focus Group Protocol . . . 170
6.2.2 Focus Group Results . . . 173

6.2.2.1 Actors and Roles Evaluation . . . 173
6.2.2.2 Agile Practices Evaluation . . . 174
6.2.2.3 Metadata Curation Analysis and Planning Dimension Evaluation . . . 175
6.2.2.4 Metadata Acquisition Dimension Evaluation . . . 176
6.2.2.5 Metadata Quality Management Dimension Evaluation . . . 178
6.2.2.6 Metadata Preservation and Dissemination Dimension Evaluation . . . 179
6.2.2.7 Metadata Curation Monitoring and Controlling Dimension Evaluation . . . 180
6.2.2.8 Metadata Curation Platform Administration Dimension Evaluation . . . 181
6.2.2.9 General evaluation of the Louvre framework . . . 182
6.2.3 Summary of the findings . . . 183
6.3 CONSIDERATIONS OF THE CHAPTER . . . 184

7 CONCLUSION AND FUTURE WORKS . . . 185
7.1 RESEARCH CONTRIBUTIONS . . . 186
7.2 LIMITATIONS . . . 187
7.3 FUTURE WORK . . . 188

REFERENCES . . . 191

APPENDIX A – META-MODEL SPECIFICATION . . . 207

APPENDIX B – EXAMPLE OF MODEL DERIVED FROM THE META-MODEL FOR DATA ECOSYSTEM . . . 239

APPENDIX C – LOUVRE FRAMEWORK SPECIFICATION . . . 241

APPENDIX D – SURVEY QUESTIONNAIRE . . . 264

APPENDIX E – FOCUS GROUP QUESTIONNAIRE . . . 279

APPENDIX F – PUBLICATIONS . . . 285

1 INTRODUCTION

Highly rapid development of networks, Internet of Things (IoT), and Web related technologies opens up new possibilities for capturing, storing, publishing, and analyzing data (MADHAVAN et al., 2007; BARNAGHI; SHETH; HENSON, 2013; CHEN; MAO; LIU, 2014). Governments, research institutions, and individuals are producing and making available large amounts of data on a variety of platforms (*e.g.*, the Web, sensor-based applications and social networks) (MADHAVAN et al., 2007; CHEN; MAO; LIU, 2014). An increasing number of individuals are recognizing the importance of data and as consequence setting up platforms for publishing, trading or even selling data (BARBOSA et al., 2014; OLIVEIRA et al., 2016a).

According to Pollock (2011), in the majority of these cases, the current basic model for the provision and usage of data is a one-way street. There is no feedback loop between data users and data consumers; *i.e.*, data users do not share data and knowledge back to their data producers. In an ideal scenario, data users should share back their cleaned and integrated data, for example. Data users should also be able to give their contribution by flagging errors, or by submitting corrections themselves. Indeed, all data users and data producers should be able to collaborate. In order to unlock the potential benefits of sharing data, a Data Ecosystem needs to be established (UBALDI, 2013).

In this context, a Data Ecosystem may be defined as a complex socio-technical network that enables collaboration between autonomous actors in order to explore data (POLLOCK, 2011; UBALDI, 2013; ZUIDERWIJK; JANSSEN; DAVIS, 2014; LEE, 2014). Such ecosystems provide an environment for creating, managing and sustaining data sharing initiatives (UBALDI, 2013; ZUIDERWIJK; JANSSEN; DAVIS, 2014; LEE, 2014; HARRISON; PARDO; COOK, 2012), such as Smart Cities (ABU-MATAR, 2016), Open Data (LEE, 2014) and Scientific Data Communities (LINDMAN; KINNARI; ROSSI, 2016). Moreover, Data Ecosystems are built on collaboration and coordination between various stakeholders, including public and private organizations, development partners and end users (OLIVEIRA; LIMA; LÓSCIO, 2019; OLIVEIRA; LÓSCIO, 2018).

While Data Ecosystems are thus arguably gaining importance, several ecosystems are not sustainable and consequently the effort spent by their actors end up not being properly used or forgotten. The lack of communication and cooperation between data producers and consumers, is one of the main obstacles moving towards sustainable Data Ecosystems (GAMA; LÓSCIO, 2014). Moreover, designing, developing and further maintaining systems for Data Ecosystems are not trivial.

For instance, no one can responsibly consume data without accompanying information that explains how the data are created, where it is located, details about the structure and meaning of the data, and how to collect, integrate, and analyze the data. A comprehensive

and meaningful description of all Data Ecosystem resources is needed. As suitable means for such a description constitute the corresponding metadata (SHANKARANARAYANAN; EVEN, 2006; STOCK; WINTER, 2011).

The most concise definition of metadata is "data about data" or "information about information"(SEN, 2004). It is high-level abstract data used to represent information in a broad range of subject areas (SHANKARANARAYANAN; EVEN, 2006; SEN, 2004; VNUK; KORONIOS; GAO, 2012). Metadata are the foundation for harnessing the vast and diverse amounts of data before they become unmanageable (DINTER; SCHIEDER; GLUCHOWSKI, 2015). When metadata are available, the objects (*e.g.*, datasets, systems and services) they describe can be rapidly located and accessed (DINTER; SCHIEDER; GLUCHOWSKI, 2015; OLIVEIRA et al., 2016b). In this sense, metadata need to be maintained and preserved as well as to be highly available over long-time in order to be properly discovered and used.

The increasing recognition of metadata as an essential asset had motivated an increased demand for metadata curation solutions. Metadata curation covers the creation and harvesting of metadata, appraisal and selection of metadata, quality assurance, preservation of metadata and other lifecycle stages, and also involves a number of IT systems. While important, in general, the current initiatives on metadata curation are a confusing mixture of activities, standards, terms and vocabularies, methods and tools (DALLAS, 2016). Referential guidelines would provide a basis to choose standard terms and definitions, processes and practices, roles and deliverables for metadata curation practitioners.

Traditional digital curation models could be adapted to curate metadata. However, in practice, these curation models have had limited impact even in the maintenance of data (MAUTHNER; PARRY, 2013). Qin, Ball e Greenberg (2012), Dallas (2016) and Kouper et al. (2013) argued that metadata curation should reduce effort demanded for curation work to meet the requirements for metadata creation, quality assurance, and management. The goal is to find a middle ground that establishes a sufficient process for curation of metadata without abandoning quality and systematic curation practices.

Moreover, the Data Ecosystem landscape brings new challenges to the metadata curation. For instance, the distributed nature of Data Ecosystem requires the creation of a kind of distributed curation environment, in which actors need to collaborate to curate metadata. Such a distributed environment also requires a great deal of communication between actors who exchange numerous ideas, information, and resources through different tools and different formats.

This new way of curating metadata also presents challenges, due to temporal, geographic and sociocultural distances and the lack of communication between the now dispersed teams. Some of the problems are: lack of face-to-face interactions, lack of experience in metadata curation, difficulty in managing the cooperation among the collaborators, limited technological infrastructure, cultural differences, time zone differences, among others.

In order to overcome the presented challenges, solutions that enable the collaboration during the metadata curation combined with lightweight curation practices should be envisioned. A promising alternative is combining the principles of Agile Software Development and Open Collaboration to construct a collaborative and flexible metadata curation model.

The Agile philosophy was popularized by the Manifesto for Agile Software Development (BECK et al., 2001). It was originated in software development but spread to other applications such as project management and IT governance (LUNA, 2009). The Agile philosophy defends, among other principles, small and highly motivated teams; active and continuous communication; informal methods; minimal software engineering artifacts and, above all, simplicity in overall development. In metadata curation, these principles could be used to light the burden of over documentation and micro-management of curation work. Moreover, the agile practices may contribute to address the lack of centralized organizations responsible for the curation work as well as the lack of dedicated actors to curate metadata.

In its turn, Open Collaboration is a dynamic process through which (professional or volunteer) collaborators collective produce a product, service or artifact (FORTE; LAMPE, 2013). Open Collaboration can be viewed as a specialized instance of collaborative systems, in which the participants are loosely coordinated and everyone can join. In Data Ecosystems, actors may join efforts in a collaborative approach to share the work of curating metadata. Therefore, actors can share the costs, risks, and technical challenges while benefiting from the wisdom of the community. The collaborative metadata curation also allows participants to have stronger and more lasting relationships.

In this context, this thesis aims to propose a framework, called Louvre[1]. A framework can be defined as structure for supporting or enclosing something else, especially a skeletal support used as the basis for something being constructed (Oxford Dictionary, 2019b). In this sense, the Louvre framework offers a wide range of processes for aiding to curate metadata in Data Ecosystems. Each process describes a coherent set of engineering and management activities related to metadata curation.

The Louvre structure is flexible and may be adapted to the needs of the actors interested in curating Data Ecosystem metadata. In this sense, processes are organized in functional dimensions, enabling modularization of the framework. The Louvre also provides a set of best practices aligned with principles of agile and open collaboration for managing curation work through the collaborative effort of self-organizing actors.

Besides fulfilling a research gap, the Louvre framework is meant to be used as a model to aid curators understand the processes, roles and responsibilities involved in successful metadata curation, and develop curation and preservation methodologies for their Data

[1] Museums and libraries have been applying curating practices for a long time to manage and care for their collections. In this way, we took inspiration from one of the most well-known museums in the world to name the proposed framework.

Ecosystems. It also help curators identify the possible risks that current practice places on their metadata, and the steps required to mitigate these risks. Like DCC curation lifecycle (HIGGINS, 2008), it can be used as an organisational planning tool to metadata curation activities, to build frameworks of standards and technologies, to ensure that processes and policies are adequately documented. The Louvre framework also contributes to enrich Data Ecosystems by producing useful, quality metadata, while preserving them to current and future use.

Moreover, this research also contributes to the Data Ecosystem area, by mapping the state-of-the-art of Data Ecosystems. It also contributes to the understanding of several issues related to the Data Ecosystems creation and maintenance. Also noteworthy is the definition, formalization and modelling of essential constructs related to Data Ecosystems.

1.1 MOTIVATION AND PROBLEM STATEMENT

There is a general consensus as to the crucial role metadata can play in Data Ecosystems (SEN, 2004; SHANKARANARAYANAN; EVEN, 2006; DINTER; SCHIEDER; GLUCHOWSKI, 2015; GRUNZKE et al., 2014; RUSSOM, 2013). In fact, different studies have found that a robust metadata management process is not only necessary but it is a fundamental requirement for a successful data consumption process (STOCK; WINTER, 2011; DINTER; SCHIEDER; GLUCHOWSKI, 2015; XIAO et al., 2015). In fact, most of the researches on Data Ecosystems assume the existence of catalogues or repositories that are used to store and to manage metadata. However, in most cases, the metadata management is underspecified, if not unaddressed at all (BELHAJJAME et al., 2008; MISSIER et al., 2007; SHANKARANARAYANAN, 2004; GRUNZKE et al., 2014; LUDÄSCHER et al., 2006; DINTER; SCHIEDER; GLUCHOWSKI, 2015).

The absence of works in the metadata management for Data Ecosystems is motivated by the incipient state of the research in Data Ecosystems. In addition, the digital curatorial area, although somewhat more mature, had its application more restricted to communities of librarianship and research data management. In addition, it should be noted that metadata, even with its recognized importance, is still treated as a second-class citizen. Hence, the lack of a solution for organizing, preserving, and provisioning metadata for the long term has resulted in valuable information becoming lost or discarded (WITT, 2008; RUSSOM, 2013; DINTER; SCHIEDER; GLUCHOWSKI, 2015).

A promising solution is the employment of a well conceived, efficient curation strategy for metadata. Metadata curation is the continuous process of managing, improving and enhancing the metadata and its use (FREITAS; CURRY, 2016; ABBOTT, 2013; RUSBRIDGE et al., 2005). Furthermore, metadata curation process aims to ensure that the metadata meets a defined set of quality requirements, such as security and privacy rules, integrity constraints or metadata availability expectations. Without proper curation metadata may deteriorate in terms of its quality and integrity over time. One of the major challenges

towards achieving efficient and continuous metadata curation is to create a methodology to structure the curation process as well as to provide a set of tools to support the curation process.

Through the literature, a broad variety of models and frameworks used for data curation were proposed, such as the works (HIGGINS, 2008; LEE, 2010; DDI, 2012; PATEL, 2011; BURTON; TRELOAR, 2009; BISHOP; GRUBESIC, 2016; JONES, 2011; CROWSTON; QIN, 2011; CORTI et al., 2014). These works focus on the specification of stages, steps and/or methods related to the curation of data. Despite important, none of the presented works are fully devoted to the curation of metadata. It is, therefore, unsurprising that metadata curation in a generalist context is still an immature discipline. Moreover, very often data users have relied on *ad-hoc* methods (spreadsheets, documents or descriptive files) to manage and maintain metadata (DINTER; SCHIEDER; GLUCHOWSKI, 2015; DALLAS, 2016). As the amount of metadata to be treated may be very large, because the potential growing number of data in a Data Ecosystems, *ad-hoc* curation methods do not scale at all.

Even, metadata curation solutions in the literature are still not consolidated. Several works remain only to the conceptual model and do not give any details about the implementation level of their solutions. As a consequence, there is a little conceptual base and few concrete metadata curation solutions that can be used as reference. Thus, the metadata curation problem presents as an opportunity for outstanding contributions in the academic aspect as well as to the Data Ecosystem actors themselves. This research aims to contribute to a better understanding of curation of metadata in Data Ecosystems and how these concepts can be applied in practice.

Based on this context, our research question is: **How metadata curation in Data Ecosystems can be structured and systematized?**

1.2 OBJECTIVES

The main goal of this research is to propose a framework to aid in performing metadata curation in Data Ecosystems.

In order to achieve the overall objective above, the following specific objectives were defined:

- Review the state-of-the-art related to the central themes of the research, involving mainly: management and curation of metadata; Data Ecosystems;
- Consolidate the theoretical basis, through the execution of systematic literature mappings that raise the state of art on Data Ecosystems and Data Published and Consumed on the Web;
- Analyze and identify the elements that compose a Data Ecosystem;

- Construct a meta-model for describing Data Ecosystems, which will be used to create metadata;
- Identify practices to manage metadata mentioned in literature;
- Structure the knowledge acquired through the definition of a framework for metadata curation;
- Evaluate the proposed framework.

1.3 RESEARCH METHOD

According to Easterbrook et al. (2008), a research method is a set of organizing principles around which empirical data is collected and analyzed. A variety of methods can be applied to any research problem, and it is often necessary to use a combination of methods to fully understand the problem. Marconi e Lakatos (2010) affirm that the research method must be directly related to the problem to be studied. That is, a research must be rigorously analyzed even before its actual execution. Aiming to meet the central objective of this work, some principles, procedures and techniques were defined. These principles, procedures and techniques are detailed in the following sections.

1.3.1 Philosophical Stance

Since different research methodologies serve different purposes, one of the first steps is choosing an appropriate research philosophical paradigm (EASTERBROOK et al., 2008). The philosophical paradigm concerns with the source, nature and development of knowledge (CRESWELL, 2010). In general, stating the research philosophy involves being aware and formulating the research beliefs and assumptions. In this sense, this thesis proposal is based on a pragmatic philosophical paradigm. According to Easterbrook et al. (2008) and Creswell (2010), this philosophical stance is characterized by accepting different concepts to support the research. Instead of focusing on the methods, the problem is more important, and researchers use all means to understand the problem. The pragmatism paradigm aims not find truth or reality, but to facilitate human problem-solving, by seeking the application of "whatever works" to solve the problem and includes a combination of different research strategies (CRESWELL, 2010). Hence, pragmatism has a strong philosophical foothold in the mixed methods or methodological pluralism camps.

This research is also based on the Design Science Research paradigm, which is aimed at building and evaluating artifacts to solve concrete classes of relevant problems by using rigorous scientific methods (HEVNER et al., 2008). According to Alan et al. (2004), Design Science Research is typically applied to categories of artifacts including languages, models, methods, and instantiations (implementations and prototypes of systems). The Design Science Research paradigm is aligned with the pragmatic philosophical stance.

The Design Science Research paradigm is fundamentally a problem solving methodology aimed at creating and evaluating a purposeful IT artifacts to address an important organizational problem. It must be described effectively, enabling its implementation and application in an appropriate domain (ALAN et al., 2004). According to Alan et al. (2004), an IT artifact include not only systems but also the constructs, models, and methods applied in the development and use of information systems. Table 2 details this vision by classifying the intended results of this thesis as artifacts types mentioned by (ALAN et al., 2004) as possible results of a research-design.

Table 2 – Design Science Research Artifacts Classification

Central Research Objective	**Artifact Type**
Construct a framework for aiding metadata curation in Data Ecosystems	Methods
Create an understanding about Data Ecosystems	Constructs (vocabulary and symbols)
Identify conceptual elements that define and describe Data Ecosystems	Constructs (vocabulary and symbols)
Construct a meta-model for describing Data Ecosystem	Models (abstractions and representations)
Identify activities and practices to systemize metadata curation	Not applicable

The objective of research in information systems is to acquire knowledge and understanding that enable the development and implementation of technology-based solutions to heretofore unsolved and important business problems. We argued that the solutions to the research problem (detailed in Section 1.2 and Table 2) are in the research-design application category, since:

- Metadata curation is still an emerging theme and lacking of effective solution;
- Data Ecosystem is also an emerging theme and its understanding is characterized by considerable conceptual imprecision;
- There are few reports of empirical research detailing how to curate metadata in distributed environments like Data Ecosystems.

This research is also aligned with other Design Science Research guidelines. For instance, design-science research requires the application of rigorous methods in both the construction and evaluation of the designed artifact (ALAN et al., 2004). All the artifacts produced were constructed using an iterative and incremental method. In particular, the artifacts were constructed through repeated cycles (iterative) and in smaller portions at a time (incremental), allowing to take advantage of what was learned during development

of earlier artifact version. Moreover, in each iteration, the artifact version were evaluated in order to verify the utility, quality, and efficacy of a proposed artifact.

Design-science research requires the application of rigorous methods in both the construction and evaluation of the designed artifact (ALAN et al., 2004). We have employed multiple empirical methods to evaluate the proposed solutions. Moreover, we provided clear contributions in the areas for both metadata curation and Data Ecosystems. Most often, the contribution of design-science research is the artifact itself (in our case, the artifacts presented in Table 2. But, we also contributed to understand and advance research in Data Ecosystem theory.

1.3.2 Research Objective

This work aims to investigate and develop a metadata curation framework for Data Ecosystems. Both Data Ecosystems and metadata curation are emerging research themes. From its objective perspective, this research is both exploratory and descriptive.

Exploratory studies are aimed at providing a better understanding on the proposed problem (YIN, 2013). Typically, an exploratory research is applied in order to allow researchers create a greater proximity to the universe of the study object. Through the exploratory research, it is also possible to obtain an explanation of the phenomena that were not initially accepted by other researchers, discover new phenomena, and formulate new ideas and hypotheses (YIN, 2013).

A descriptive research aims to describe a particular individual, group or phenomena (GIL, 2008; YIN, 2013). Different from exploratory research, a descriptive study stands out because it represents attempts to explore and explain a given topic by providing additional information about it. This is where the research aims to describe what is happening in more detail, filling in the missing parts and expanding our understanding. To do this, collect as much information as possible, rather than making assumptions or models designed to predict the future (GIL, 2008; YIN, 2013).

1.3.3 Research Variables Nature

With regards the nature of the variables, this research can be classified as both quantitative and qualitative. Qualitative variables are not measured in the form of numbers. These variables are adequate to develop a detailed understanding of the meanings and characteristics presented by study subjects. This means that qualitative variables are used to study things in their natural settings, attempting to make sense of, or interpret, phenomena or feelings in terms of the meanings people bring to them.

In its turn, quantitative variables gather data in a numerical form. Experiments typically yield quantitative data, as they are concerned with measuring things. However, other research methods, such as controlled observations and questionnaires can produce quantitative information (MCLEOD, 2017). In fact, pragmatic research typically employ

mixed research procedures associated with both quantitative and qualitative data (EASTERBROOK et al., 2008).

1.3.4 Research Methodology Steps

This research was planned based on the research processes used by Almeida (2015), Luna (2009), Santana (2015). In summary, twelve macro research stages were performed, as shown in Figure 1. Several research procedures were employed, including *ad-hoc* literature review, systematic mappings, survey, case study and focus group. In summary, this research was structured in two phases: an exploratory (Dark Boxes) and a descriptive one (Light Boxes).

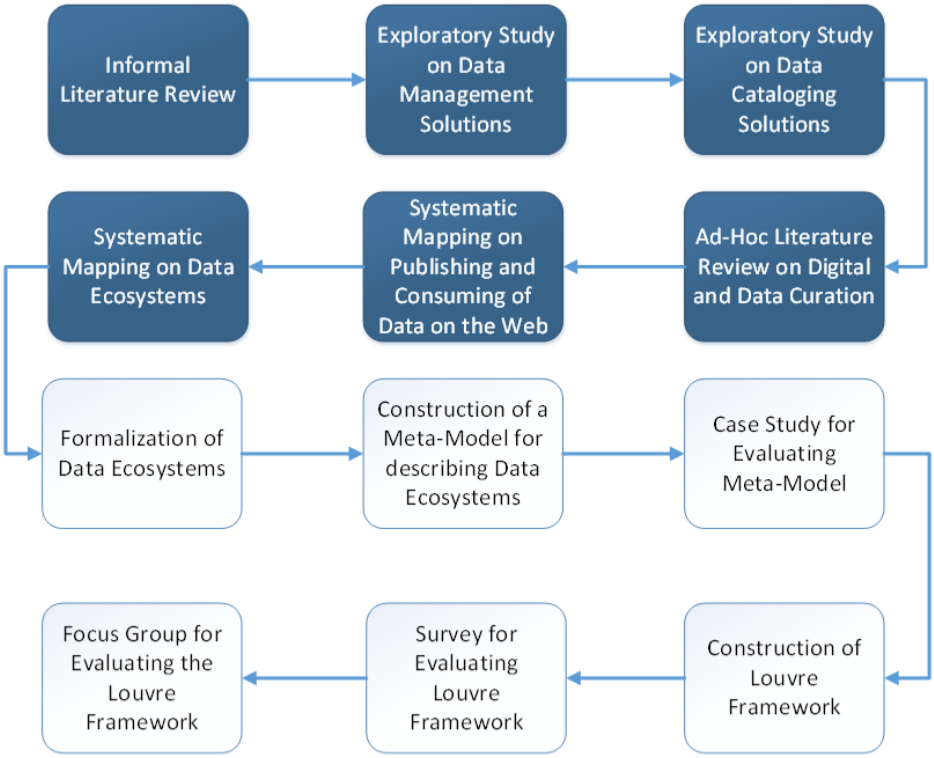

Figure 1 – Research Methodology Steps. Source: Author

The first phase presents a more exploratory characteristic and is composed of a set of six stages. This phase had as main objective the construction of a solid theoretical basis to support the later stages of this research work. An *ad-hoc* literature review (consisting of an initial bibliographic research), two exploratory studies and two systematic mappings of the literature were carried out.

At first, an *ad-hoc* literature review was carried out to gather an overview of the research problem. An ***ad-hoc* literature review** is an informal search and evaluation of the available literature in a given subject or chosen topic area. In particular, an *ad-hoc* literature review employs informal or subjective methods to collect and interpret studies. The aim of this review was to obtain the main concepts of the area. This stage

was important as it formed the initial theoretical background for the continuity of this research. Furthermore, the ad-hoc review created a basis that allowed preparing a protocol to formally investigate the state-of-the-art of management and curation of metadata.

In the next stages, two literature exploratory studies were then performed. The first study aimed to identify and analyze data management solutions. The second study aimed to identify solutions for cataloging datasets. The results of these stages of the research resulted in a set of publications (SILVA et al., 2015; OLIVEIRA; OLIVEIRA; LÓSCIO, 2017).

In addition, in order to consolidate the theoretical basis, two systematic literature mappings ((i) data published and consumed on the Web and (ii) Data Ecosystems) were performed in this work. A systematic mapping study is a protocol-driven review and synthesis of data focusing on a topic or on related key questions. The protocol specifies the methods used to guide the selection of the corpus of studies, the extraction of data, and the analysis and interpretation of the results. Unlike *ad-hoc* literature review, systematic mapping is more suitable for dealing with broad and poorly-defined areas (PETERSEN et al., 2008; KITCHENHAM; CHARTERS, 2007). In addition, a systematic mapping is also more readily able to answer broader exploratory questions (*e.g., What do we know about topic "T"?*) (ARKSEY, 2005).

This first systematic literature mapping aimed at identifying and analyzing how data have been published and consumed on the Web. It also verified what the existing solutions for the metadata management are. From the analyzed publications (studies) was possible to identify a gap in the research related to metadata management. The results were published in (SANTOS et al., 2018). In its turn, the second systematic literature mapping aimed at defining and analyzing how Data Ecosystems are defined, structured and organized. It also analyzed how a Data Ecosystem is characterized and how it works. The results of this study are described in details in Chapter 3. The results were also published in (OLIVEIRA; LIMA; LÓSCIO, 2019).

In the second phase, the research presents a more descriptive perspective. This phase aimed to construct and continuously improve the artifacts proposed in this thesis. In a first stage, we aimed to construct a more precise and formal terminology regarding Data Ecosystems. In this sense, it was identified the conceptual elements of Data Ecosystems, based on the definitions and other wider conceptualization of the field (*e.g.*, characteristics, features and properties) derived from relevant papers found in the related systematic mapping. The results were published in (OLIVEIRA; LÓSCIO, 2018). This formal definition was then used as a starting point for the creation of a meta-model. Such meta-model defines the Data Ecosystem fundamental concepts and their inter-relationships for enabling analysis and description of Data Ecosystems. The meta-model, besides allowing the creation of a shared vocabulary, also serves as a basis for the creation of descriptive metadata of Data Ecosystems.

In the next stage, a case study was performed in which the meta-model could be

evaluated in a real Data Ecosystem related to an open data program coordinated by a Brazilian public university. Yin (2003) defines a *case study* as "*an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident*". A case study is one of the recommended strategies for evaluating artifacts produced with a design science research approach (PRIES-HEJE; BASKERVILLE; VENABLE, 2008). A case study is conducted in a real environment, which gives it the power to perform an appropriate evaluation of a study object in a concrete reality. The results were published in (OLIVEIRA et al., 2018).

In the final stages, an initial design of the metadata curation framework was created. This initial design was then subjected to various verification and refinement cycles. As a result, an alpha version of the framework was developed. In the next stage, a survey study was performed to evaluate the alpha version of the proposed framework. A survey is a common research procedure used to collect the opinions, beliefs and feelings of selected groups of individuals, often chosen for demographic or community sampling (KITCHENHAM; PFLEEGER, 2002). It encompasses a series of questions to gather respondents' social, professional or familiar situation, their opinions, their expectations, their knowledge, evaluation or conscience level about an product, event or problem, for instance. The feedback collected from this study was used as a basis for the generation of a beta version of the proposed framework. The results obtained in this stage of the research generated a publication (OLIVEIRA; LÓSCIO, 2017) and an article submitted to the Brazilian Symposium on Information Systems (SBSI) (OLIVEIRA; LÓSCIO, 2019).

In the next stage, another evaluation was carried out using the focus group method. A focus group can be defined as a group of individuals reunited to evaluate products, systems, concepts or to evidence problems (KONTIO; LEHTOLA; BRAGGE, 2004). In this study, the focus group was aimed at evaluating the beta version of the proposed framework. Based on the results obtained in this study, a new version (version 1.0) of the proposed framework was created and then presented in this thesis.

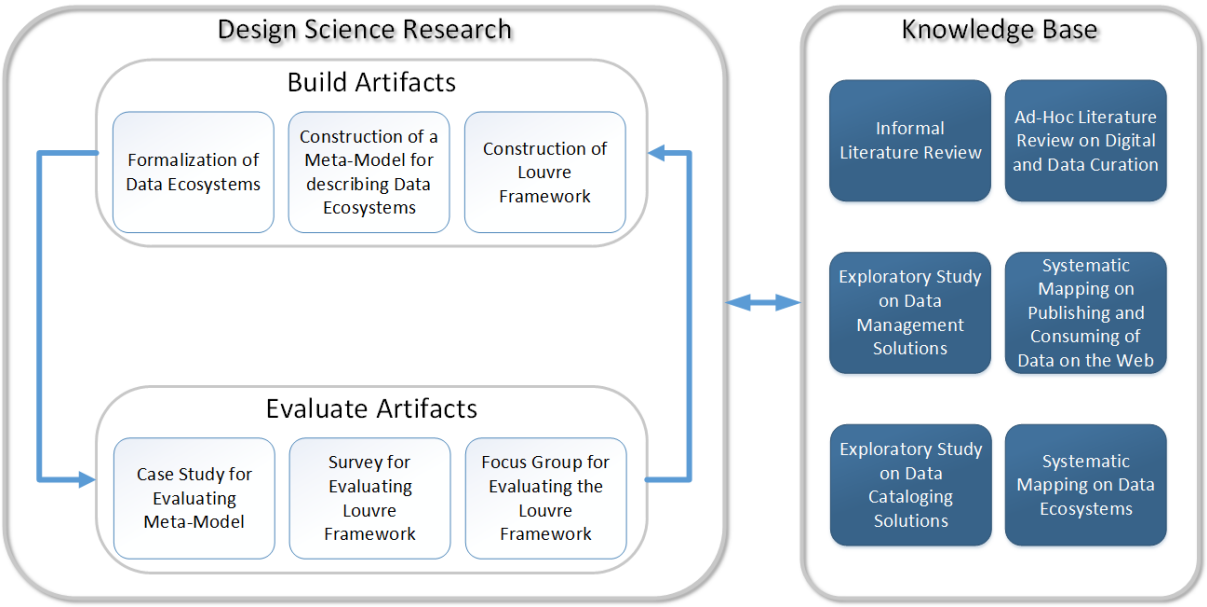

Figure 2 – Design Science Research Lifecycle. Source: Author

These steps also is aligned with Design Science Research lifecycle. Figure 2 borrows the Design Science research lifecycle found in (HEVNER, 2007) and overlays the previously presented steps. The dark boxes connect the research activities with the knowledge base of scientific foundations, experience, and expertise that informs the research project. In its turn, light boxes represent the central Design Cycle, which iterates between the core activities of building and evaluating the design artifacts and processes of the research.

1.4 OUT OF SCOPE

The proposed framework is part of a broader context, which involves the creation and management of a metadata curation environment as presented in Figure 3.

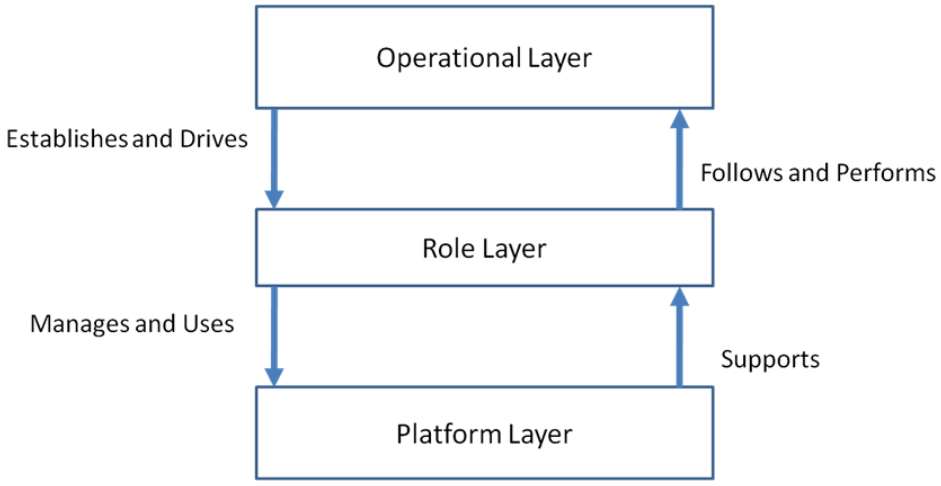

Figure 3 – Metadata Curation Environment. Source: Author

The top layer, called Operational Layer, represents the set of procedures, practices, processes, activities, and techniques used to guide the curation tasks to be performed on metadata throughout its life. The middle layer, called Role Layer, represents the people part of curation work. It includes organizational structures, roles, and responsibilities involved in performing and supervising the curation tasks. The bottom layer, called Platform Layer, represents a set of instruments produced to assist those responsible for metadata curation. These instruments facilitate the curation of metadata during the metadata lifecycle, including metadata repositories, metadata extraction tools, data cleansing tools, and quality assessments systems.

In particular, the Louvre encompasses elements of the top and middle layer. Despite important, the platform layer is left out of Louvre scope. Even with the use of the above four solutions, considerable efforts should be carried out to the design and development of a metadata curation platform. Such platform can be as complex as the framework proposal, involving the definition of architectures, infrastructure, while addressing interoperability and reliability issues. For instance, there are several metadata storage solutions available (*e.g.*, NoSQL databases, cloud-based storage services or cataloguing solutions), as a consequence, selecting the most suitable storage solution is not a trivial task. It must be efficient regarding latency, storage overhead, and it also needs to meet diverse requirements. Nonetheless, it must also be scalable enough to allow multiple actors to access metadata simultaneously. Another primary function is to integrate different metadata.

1.5 ORGANIZATION OF THE DOCUMENT

The remainder of this document is organized as follows. In Chapter 2 we discuss the state of the art on Metadata Curation and some related works. In Chapter 3, we describe the concepts and relations of a Data Ecosystem. In Chapter 4, we present the proposed metamodel. In Chapter 5, we present a general overview of the proposed framework. Chapter 6 presents the evaluation of the framework. Chapter 7 concludes with final remarks about this research, its main contributions, and future work.

2 THEORETICAL BACKGROUND

In this work, we aim to provide a metadata curation framework for Data Ecosystems. A well-curated metadata would potentially enable the sustainability of Data Ecosystems by sharing and improving knowledge of and about the ecosystem as a whole.

Metadata curation is a term that seems to have acquired multiple meanings depending on the author and their particular perspective. These varying meanings can easily lead to confusion, especially in cross-disciplinary discussions. Metadata curation involves the creation of models to describe the metadata and its curation operations as well as developing metadata management platforms interacting with different storage technologies.

Meanwhile, there is a myriad of work dealing with some of these aspects long before the rise of contemporary Data Ecosystem. Historically, metadata was either implicitly managed by data solutions or co-manage alongside data assets. Regarding to the proposal of metadata management processes, the first attempt were performed by some Data Management frameworks, which had dedicated a specific process to deal with the metadata management issues. With the advent of data curation, efforts were also made to manage metadata more explicitly.

This Chapter presents the background concepts of this thesis proposal and the state of the related work on metadata curation and data management.

2.1 BACKGROUND ON METADATA CURATION

Metadata curation and data curation are relatively a new research field. They are often accomplished by archivists, librarians, scientists and historians. Enterprises and research institutions are starting to use curation to improve the quality of their data and knowledge within their operational and strategic processes (CURRY; FREITAS; O'RIÁIN, 2010).

Metadata curation is a crucial process for several areas, including Data Ecosystems. However, we have not found any definition for this term. Most of the studies mention metadata curation as "the process whereby metadata are curated". Since metadata curation can be viewed as a specialized approach for digital curation, we may use its terminology to develop an understanding about the concerns related to the metadata curation.

For historical background, in the early 1990s, the practices of curation were almost completely focus on data, so called data curation. Over time the field of curation has grown and digital curation has become recognised as a broad practice, which curate resources ranging from data objects to multimedia content (*e.g.*, images and videos)(ABBOTT, 2013; RUSBRIDGE et al., 2005). For instance, the Digital Curation Centre (DCC) defines digital curation as "the processes of digital archiving and digital preservation, but it also includes all the processes needed for good data creation and management, and the capacity to add

value to data to generate new sources of information and knowledge" (RUSBRIDGE et al., 2005). More recently, data-focused curation efforts become more popular again due to explosion of digital data.

According to Lord e Macdonald (2003b), Digital curation is the "activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use.". Digital curation refers to "the actions needed to maintain and utilize digital data and research results over their entire lifecycle for current and future generations of users" (COMMITTEE et al., 2003). Pennock (2007) also use the lifecycle concept to define digital curation. But, they add the notion of importance for current and future use. According to them, "digital curation, broadly interpreted, is about maintaining and adding value to a trusted body of digital information for current and future use; specifically, we mean the active management and appraisal of data over the lifecycle of scholarly and scientific material."

Another definition was proposed by Lord e Macdonald (2003a), who define digital curation as the process of curating, archiving and preserving data to ensure the data is fit for contemporary purpose and available for further discovery and re-use. All the three activities (*i.e.*, curation, preservation and archiving) are involved in digital curation over the time. However, both definitions were proposed for digital objects in general. Nevertheless, according to Ball (2012) in some contexts there is a tendency to consider data and digital object as the same thing. In fact, most of the digital curation solutions have been used in practice to curate data and datasets.

Based on these definitions, in this Thesis, metadata curation is defined as the set of maintenance tasks related to organization and use of metadata from its point of creation, to ensure it fits quality criteria and is available for current and future use. For Data Ecosystems, the metadata curation process enhances the long-term value of its resources by making the knowledge acquired available for further actors.

According to Goble et al. (2008), the lack of a metadata curation leads to the ignorance of the availability of resources, the inadequate information to construct, choose or operate data-based systems and poor adoption of the data and supporting systems. Data consumers (*i.e.*, individuals or organizations who process, analyze, filter, and/or aggregate data under specific conditions along a time interval) simply do not know what is available and hence reinvent solutions. In several cases, the Data Ecosystem resources are unknown, unused, poorly used or misused as a result (GOBLE et al., 2008). The lack of metadata describing these assets is a major obstacle to sustainable ecosystems.

The metadata is invaluable for creation, functioning, reuse and comprehension of Data Ecosystems. Hence, its proper curation involves the capture, organization and storage of metadata about all the important aspects of a Data Ecosystem, plus its appraisal, processing and delivery. In this sense, metadata curation is a broad practice that encompasses a number of data disciplines, including data modeling, data storage, data integration, data

quality, data governance, content management, database administration, and so on.

2.2 DATA CURATION MODELS

Through the literature, a broad variety of curation process models were proposed (BALL, 2012). These works focus on the specification of stages, steps and/or methods related to the curation of digital assets, mainly data. They also present different definitions of curation process, such as lifecycle, vocabulary, guidelines and methodologies. Despite these works do not focus on metadata curation, they cover some aspects related to metadata curation in their processes and/or lifecycles.

In the following, some of the main digital curation models available in the literature are described. These models were chosen because they are most closely aligned with the intended goals of this Thesis.

2.2.1 DCC Curation Lifecycle Model

The DCC Curation Lifecycle Model (HIGGINS, 2008) proposes a full lifecycle of data curation tasks, intended as a planning tool for data producers, curators and data consumers. The key element of the DCC Curation Lifecycle Model is data, which represents any information in binary digital form, including:

- Digital Objects: simple digital objects (discrete digital items such as text files, image files or sound files, along with their related identifiers and metadata) or complex digital objects (discrete digital objects made by combining a number of other digital objects, such as websites)
- Databases: structured collections of records or data stored in a computer system

The Figure 4 presents the curation stages and actions defined by the DCC model. Different from other curation models, DCC lifecycle model provides a non-linear view of digital curation. There are several concentric rings in its structure. One or more rings represents multiple sets of related actions. The Full Lifecycle Actions are represented by Description and Representation Information, Preservation Planning, Community Watch and Participation, and Curate and Preserve. These actions apply to every stage in the lifecycle.

Description and Representation Information stage represents the creation, collection, preservation and maintenance of sufficient metadata to enable the data to be used and re-used for as long as they have value to justify continued curation. Preservation Planning stage defines the strategies, policies and procedures for all curation actions. Community Watch and Participation stage comprises the observation of the target community of the data, in order to track changes in their requirements for the data, and participation in the development of standards, tools and software relevant for the data. Curate and Preserve

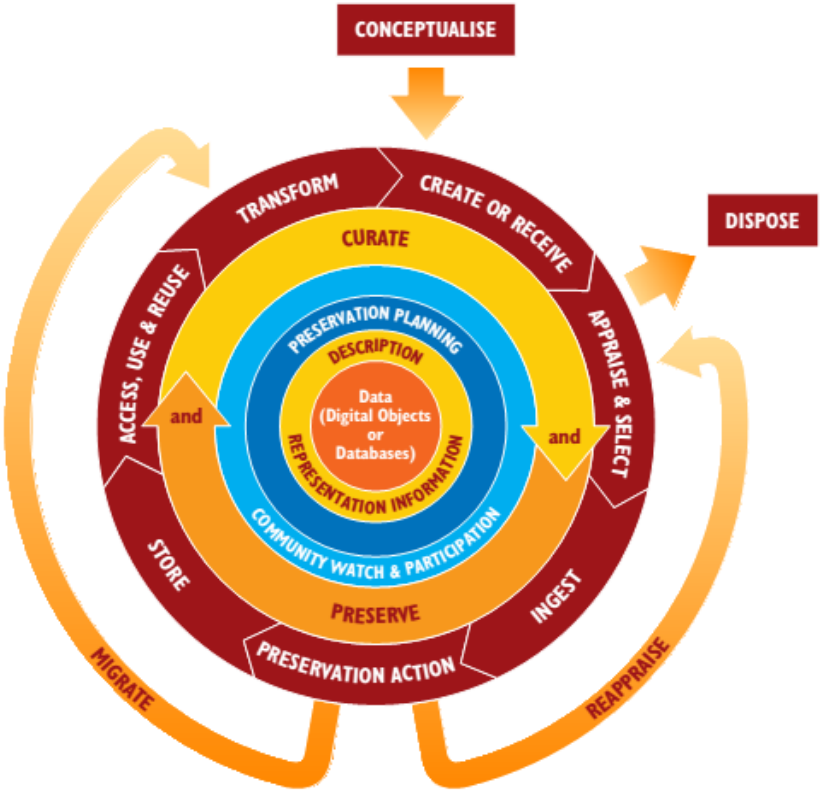

Figure 4 – DCC Curation Lifecycle Model. Source:(HIGGINS, 2008)

stage properly describes most of the curation actions in the DCC model, which are used to represent the execution of the planned management and administrative actions supporting curation.

The Sequential Actions in the outer ring represents the key actions needed to curate data, ranging from creation to their use and re-use. Moreover, these actions are not carried out once only. They may be repeated for long as the digital objects are being curated. The sequential actions are not exclusively concerned with curation, but rather represent stages of the data lifecycle which ought to have a curation component. The sequential actions are:

- Conceptualise: planning stage of the data generation and collection activities;
- Create or Receive: data collection , where ‘create’ refers to original data generated and recorded by researchers, and ‘receive’ refers to pre-existing data collected from other sources;
- Appraise and Select: evaluating data and selecting for long-term curation and preservation;
- Ingest: transferring data to an archive, repository, data centre or other custodian;

- Preservation Action: undertaking actions to ensure long-term preservation and retention of the authoritative nature of data;
- Store: storing the data in a secure manner adhering to relevant standards;
- Access, Use and Reuse: ensuring that data is accessible to both designated users and reusers, on a day-to-day basis;
- Transform: creating new data from the original.

The ultimate set of actions is represented by Occasional Actions, which may occur in face of specific conditions. Furthermore, these actions are not required for all data. The Dispose Action aims to remove from archive repository data that has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. The Reappraise Action aims to return data which fails validation procedures for further appraisal and re-selection. The Migrate Action aims to migrates data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence.

Various steps of DCC filecycle model are on some level associated with the metadata curation. The Description and Representation Information stage and Create or Receive action create administrative, descriptive, structural, technical and other metadata to enable the data to be used and re-used. Once the data have completed, in the Preservation Action, preservation metadata may also be created as well as they are properly maintained in the archive repository. Meanwhile, the Preservation Planning stage also may include strategies, policies and procedures for metadata curation.

2.2.2 DDI Combined Lifecycle Model

The Data Documentation Initiative (DDI) Combined Lifecycle Model (DDI, 2012) is a more linear curation model for research data, particularly social science data. First versions of DDI were developed by an informal network of individuals from the social science community. Historically, DDI was focused on data archiving. However, the DDI version 3 represented a major change from preceding versions in a way that the full data lifecycle is supported.

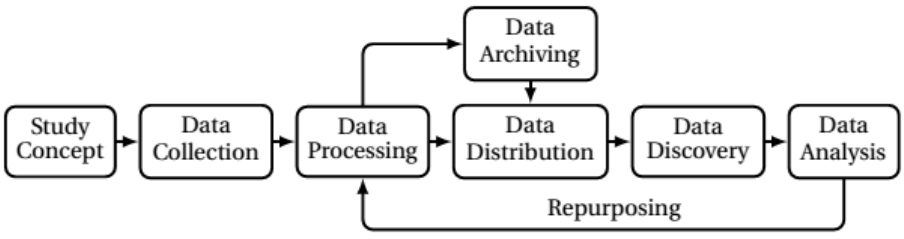

Figure 5 – DDI Combined Lifecycle Model. Source:(DDI, 2012)

As Figure 5 shows, this model has eight elements or steps which can be summarized as follows:

- Study concept: At this stage, it is defined the research question and the methodology to collect data;
- Data collection: At this stage, the data is collected from its sources. The DDI model proposes different methods to collect data, like surveys, health records, statistics or Web-based collections;
- Data processing. At this stage, the collected data are processed to answer the proposed research question;
- Data archiving. Both data and metadata should be archived to ensure long-term preservation and access;
- Data distribution. This stage involves the different ways in which data are distributed, as well as questions related to the terms of use of the used data or citation of the original sources;
- Data discovery. Data can be discovered by others;
- Data analysis. Data is used by others to achieve different goals;
- Repurposing. Data can be used outside of their original framework, restructuring or combining it to satisfy diverse purposes

DDI has its strength in the domain of social, economic, and behavioral research data (BOSCH et al., 2012). However, there is a common understanding is that metadata is part of a data lifecycle. DDI consider metadata must be maintained and versioned over time in order to support comparison between studies within a larger data collection (*e.g.*, Geographic Structures, Geographic Locations, Concepts, Universe hierarchies, and Organizations and Individuals). The metadata production should begin early in a research project and should be done when it happens. Moreover, the metadata could be then reused along the data lifecycle. In this sense, a series of metadata management activities are spread along data lifecyle.

2.2.3 DataONE Lifecycle Model

The DataONE is a foundation aimed at supporting science by engaging the relevant science, data, and policy communities as well as providing easy, secure, and persistent storage of data (ALLARD, 2012). The data preservation is accomplished by disseminating integrated and user-friendly tools for data discovery, analysis, visualization, and decision-making (ALLARD, 2012). Besides such tools, DataONE also provides provides a data lifecycle model as an aid to manage and preserve data.

It uses this lifecycle to place the tools from its toolkit into context. In particular, both tools and lifecycle were conceived through a comprehensive program of research, design, and development to create systems to preserve, disseminate, and protect research objects in a secure, reliable, and open approach (ALLARD, 2012).

Figure 6 – DataONE Lifecycle Model. Source:(ALLARD, 2012)

The model defines the following eight stages, as presented in Figure 6. At each stage, different users may interact with the data. An user may skip one or more stages and it is unlikely that one user will interact with the data at all stages. The stages are:

- Plan: Creating a data management plan that describe data that will be compiled, and how the data will be managed and made accessible throughout its lifetime;
- Collect: Acquiring data by hand or with sensors or other instruments and the data are placed a into digital form;
- Assure: Ensure the data quality through checks and inspections as well as standard formats, codes, units;
- Describe: Providing sufficient metadata to ensure the data are accurately, understandable, discoverable and reusable;
- Preserve: Depositing data in an appropriate long-term archive so they may be protected from bit-level degradation and format obsolescence;

- Discover: Exposing data through a searchable interface, along with the relevant information about the data (metadata);
- Integrate: Transforming and combining several data from disparate sources to form one homogeneous set of data that can be readily analyzed;
- Analyze: Analyzing data through statistical and analytical models to the data in order to extract meaningful answers to research questions;

Like DCC lifecycle model, Various steps of DataONE filecycle are on some level associated with the metadata curation. In particular, the Describe stage is responsible for creating sufficient metadata to understand the origin, organization and characteristics of a data.

2.2.4 OAIS Reference Model

The Open Archival Information System (OAIS) Reference Model (LEE, 2010) is a conceptual framework for building a complete data curation organization, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a target community. Although originally developed by the Consultative Committee for Space Data Systems, the OAIS had became the ISO 14721:2003 and recently updated to ISO 14721:2012 (ISO, 2012), which is a standard model for digital preservation systems at many institutions and organizations.

The OAIS Model describes functions, roles and responsibilities of data repositories. The OAIS data repository involves the interaction of four actors, which are: data producers, data consumers, management, and the repository itself. The OAIS Model was developed to attend the archiving of data to Space Science communities. However, it can be framed sufficiently generally to apply to maintain all kinds of digital and physical resources, such as data collections and metadata.

The OAIS Reference Model defines two models to describe the curation environment: information model and functional model. The former describes the types of resources, called Digital Objects, that an OAIS repository hold. The Data Objects can represent either physical objects or digital objects. Each data object has associated a set of metadata description elements, called Representation Information, which can be structural (*e.g.*, file formats), semantic (*e.g.*, a code book) or some other type (*e.g.*, software). Moreover, data object also has metadata about the Content Information, and combined with it, it forms an Information Object. The Reference Model is used to define which actors are responsible for interpreting and understand the information contained in a data object either because of their established knowledge base or with the assistance of supplementary representation information that is included with the data object.

The Figure 7 presents the Functional Model, which describes the functions, processes, roles and responsibilities of an OAIS environment. There are three packages: Submission

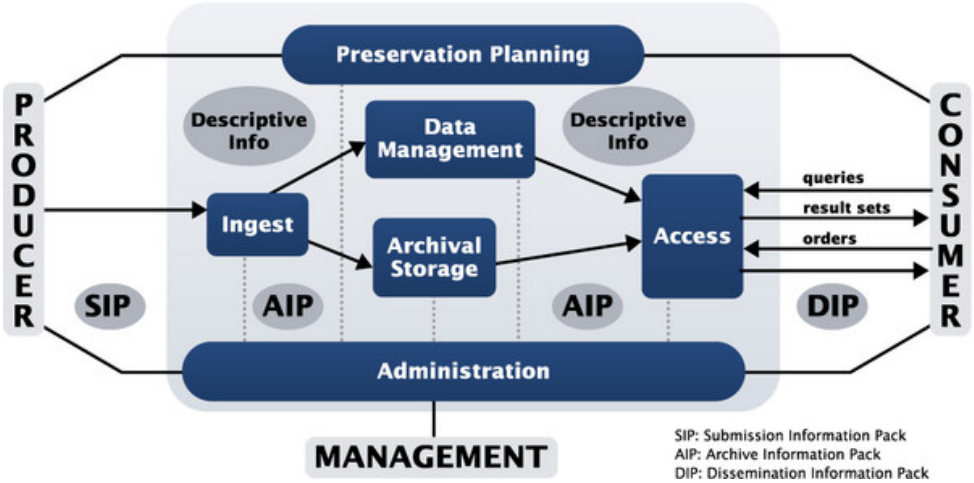

Figure 7 – OAIS Functional Mode. Source:(LEE, 2010)

Information Package (SIP), Dissemination Information Package (DIP) and Archival Information Package (AIP). A package is a set of relevant metadata necessary to describe a particular object and curation function. In the Functional Model, there are also five functions in an OAIS environment:

- Ingest function: receives information from producers and packages it for storage.
- Archival Storage function: stores, maintains, and retrieves AIPs.
- Data Management function: coordinates the system information and descriptive information about digital objects.
- Administration function: manages the daily operations of the archive. This function attains submission agreements from information producers, performs system engineering, audits SIPs to ensure compliance with submission agreements, develops policies and standards. It handles customer service and acts as the interface between Management and the Designated Community in the OAIS environment.
- Access function: This function includes the user interface that allows users to retrieve information from the archive.

Besides the related metadata management functions, the creation of all the three packages includes creation of myriad metadata. OAIS also define some protocol standards to search and retrieve metadata information about digital and physical data sources.

2.2.5 Summary of Data Curation Models

Data curation models are shaping the way data have been curated by individuals and research institutions. Four models analysed here represent a wide range of approaches and purposes in data curation: from abstract representations of the whole data lifecycle

through to much more detailed models of data management and reuse, and even just of data curation.

According to Carlson (2014), the data curation models aid to define and illustrate the data curation work visually, making it easier to identify the component parts or distinct stages of non-trivial processes. These models can vary across several aspects. For instance, curation models vary according to their form (*e.g.*, linear, circular, non-linear or other models) (FEICHENG; JUNCHENG, 2010; MÖLLER, 2013), according to the context (*i.e.*, individual-based, organisation-based and community-based models) (CARLSON, 2014) or according to their focus (whether the model seeks to define what would be good practice or is describing actual practice). Moreover, the models can be classified whether data and metadata are differentiated (MÖLLER, 2013).

Furthermore, the presented models also cover different concerns related to metadata curation. In general, all of them, describe activities for metadata creation. Others goes beyond by proposing functions to enable the archive and preservation of metadata. Finally, each model has its strengths relevant to its particular purpose. Some are specific to a particular academic field or institution, others seek to generalise to any form of data. Using (CARLSON, 2014; MÖLLER, 2013; COX; TAM, 2018) as starting points we classify the presented curation models with the aspects described above, as presented in Table 3. Each column corresponds to a work, cited before, and each row corresponds to an aspect.

The subject matter of the data curation model, which define the type of object to be curated. It could be anything from the whole research process to a model just concerned with digital objects and data (COX; TAM, 2018). While DCC and OAIS are focused on data, the scope of DDI and DataONE is whole research process. DDI is an individual-based model meant to represent the work performed by a particular team for a set duration time. In its turn, OAIS is an organization-based model, which provide a broader representation of stages and activities that are common across different projects. DCC and DataOne are community-based models meant for a particular academic community or discipline (including professionals that support them) to define existing or good practice. All the presented models are prescriptive instead of descriptive. Prescriptive models which try to establish a set of steps which are then suggested for use by others, while a descriptive model will look at a given system and find a lifecycle in it. DCC and DataONE models provide a high level of abstraction, while OAIS and DDI provide much details on how to perform the curation of data.

Regarding to the centralized nature, most of the works does not describe a concrete system at any one place. DataONE is is an exception, as such model recommends the use of a network of archive repositories. Open and Closed describe whether a system allows arbitrary data from the outside to enter its lifecycle, which wasn't initially meant for this particular system. Again, DataONE is the exception since it has a strong emphasis on dissemination and re-use of data. However, DCC includes an explicit import stage

Table 3 – Relation among data curation models

	DCC	**DDI**	**DataONE**	**OAIS**
Subject Matter	Data Curation Lifecycle	Research Data Lifecycle	Research Lifecycle	Data Curation
Curation Context	Community-based	Individual-based	Community-based	Organization-based
Prescriptive/ Descriptive	Prescriptive	Prescriptive	Descriptive	Prescriptive
Abstraction Level	High	High	High	Low
Centralized/ Distributed	Centralized	Centralized	Distributed	Centralized
Closed/Open	Partially Opened	Closed	Open	Closed
Collaborative Effort	No	No	No	No
Distinction data vs. metadata	Yes	No	No	Yes
Metadata Curation Activities	Planning; Creation; Preservation	Creation; Preservation; Dissemination	Creation; Preservation; Dissemination	Creation; Preservation; Dissemination

(*i.e.*, create and receive) for the inclusion of external data. None of the models propose a collaborative effort to the curation work.

And, finally, with regards to metadata curation work, most of the models recommend activities to create and preserve metadata. A few also recommend activities for metadata dissemination. And, only DCC recommends activities for metadata curation planning. An important point in looking at presented models is the question whether or not a distinction between data and metadata is made. DCC and OAIS explicitly distinguish between data and metadata about those. However, the rest of the models doesn't discuss metadata in detail, *i.e.*, the distinction is not so clear.

Although we not present an exhaustive picture of the range of data curation models, this analysis seems to support the value of a metadata curation model proposal.

2.3 DATA MANAGEMENT FRAMEWORKS

Data Management is technology-enabled discipline aimed at collecting, aggregating, consolidating, quality-assuring, persisting and distributing data throughout an organization. Data management focuses on ensuring a common understanding, consistency, accuracy and control of data (KAREL et al., 2006). For instance, it looks for ensuring that a organization does not use multiple and inconsistent versions of the same data in different parts of its systems.

Data Management is composed of mixture of processes and tools. Functions commonly covered by master data management include source identification, data collection, data transformation, data cleansing, data storage, data distribution, data classification, data

enrichment, data governance and others.

This section presents a general view about two important data management models, which inspired the conception of the framework proposed in this thesis.

2.3.1 DMBOK

Data Management International (DAMA) is an international non-profit organization, composed by professionals and technicians dedicated to promote concepts and best practices related information management and data governance. DAMA is responsible for the Data Management Body of Knowledge (DMBOK) that compiled a set of processes and practices to serve as a comprehensive guide to data management activities. The DMBOK was developed in 2009 with the collaboration of more than 120 professionals. It provides an overview of data management as well as gives definitions of data management processes, roles, and delivery results and their standard terminology.

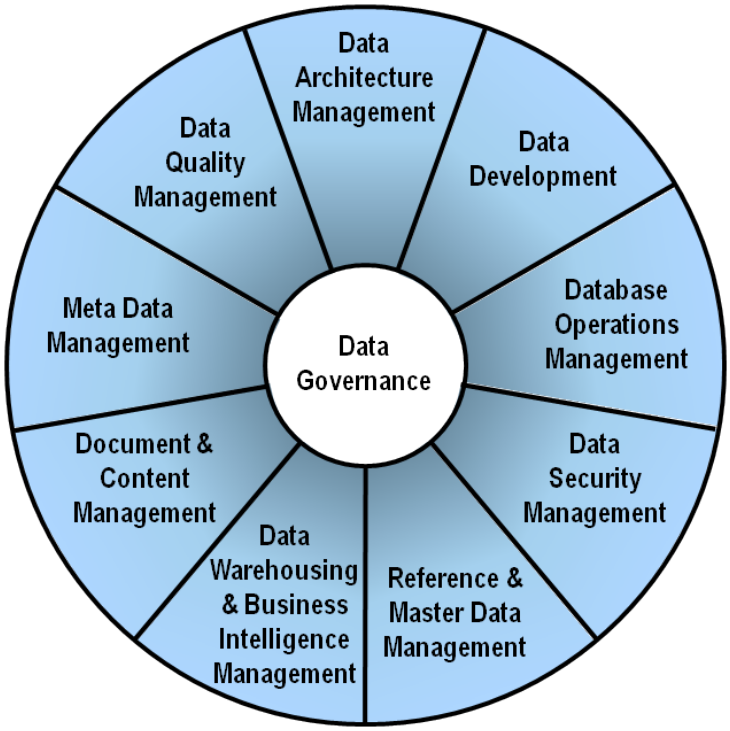

Figure 8 – DMBOK functions. Source:(MOSLEY et al., 2010)

DMBOK has 10 main data management functions, as presented in Figure 8. Taking data governance as the core, it narrates the scope of each function by clockwise rotation. Each one of functions are detailed as (MOSLEY et al., 2010):

- Data Governance is responsible for the high -level planning, supervision and control of data asset management;
- Data Architecture Management is responsible for defining the data management blueprint to meet the enterprise data needs. This function includes the development

and maintenance of enterprise data architecture, within the context of all enterprise architecture;

- Data Development is responsible for designing, implementing, and maintaining solutions to meet the data needs of the enterprise. It includes data demand analysis, implementation, testing, maintenance and other solutions;
- Data Operation Management is responsible for planning, controlling, and supporting of data lifecycle, ranging from data acquisition to archival and purge of data;
- Data Security Management is responsible for planing security policies and measures to ensure data confidentiality and hierarchical access rights.
- Reference and Master Data Management is responsible for planning, implementation, and control activities to ensure consistency of data values.
- Data Warehouse and Business Intelligence Management is responsible for planning, implementation, and control processes to provide decision support data and support for knowledge workers engaged in reporting, query and analysis;
- Document and Content Management is responsible for the management of electronic files and physical records (including text, graphics, images, audio, and video);
- Metadata Management is responsible for integrating, controlling, and providing high-quality metadata;
- Data quality management is responsible for planning, implementation, and control activities that apply quality management techniques to measure, assess, improve, and ensure the fitness of data for use.

Although DMBOK is useful for understanding data management issues, it was designed for organization context. Typically, the majority of organizational data is structured, internal to an organization and stored into organizations' internal (relational) databases. In addition, DMBOK identifies metadata as the categories of data. However, the definition for metadata is too narrow. Example DMBOK's metadata definition are "any data used to organize or categorize other data","[it is] used relating data to information both within and beyond the boundaries of the enterprise" and "usually it consists of codes and descriptions of definitions". These definitions exclude various applicability of metadata, including the description of other resources (*e.g.*, software) or even administrative and operational information. Another problem related to DMBOK is its institutional focus. The most important consequence of the institutional focus is that data management becomes largely dependent on centralized organizational structures. DMBOK does not consider sufficiently the role of collaborative work.

2.3.2 ISO 8000

The International Organization for Standardization (ISO) is an international federation of national standards organizations dedicated to create and provide common standards between nations. ISO is responsible for the family of standards ISO 8000, which provides a set of frameworks for improving data quality for specific kinds of data. These frameworks can be used independently or in conjunction with quality management systems (ISO, 2011). ISO 8000 is being applied in a variety of industry sectors and countries around the world [1].

ISO 8000 covers industrial data quality characteristics throughout the product lifecycle from conception to disposal (ISO, 2011). It also describes the vocabulary and features as well as defines the requirements for standard exchange and assurance of quality of data. The strength of ISO 8000 lies in the fact that their definitions about quality data is based on international agreement. It focus on a specific a kind of data including, but not limited to, master data. Master data term represents the subset of data in a organization that contains the most valuable information (KAREL et al., 2006).

The family of standards ISO 8000 was first proposed in 2002, and the first components were approved in 2009. Each standard is being published as a number of separate documents, which ISO calls "parts". Currently several standards of the 8000 series are released. The main standards are (ISO, 2011):

- Part 1 (Overview): Contains an introduction to ISO 8000. It contains a statement of the scope of ISO 8000 as a whole, principles of data quality, and a summary of the content of the other parts of ISO 8000;
- Part 2 (Vocabulary): Defines terms relating to data quality used in the ISO 8000 series of parts;
- Part 8 (Concepts and measuring): Describes fundamental concepts of information and data quality, as well as specifies prerequisites for measuring information and data quality;
- Part 61 (Data quality management: Process reference model): Specifies the processes required for data quality management;
- Part 100 (Exchange of characteristic data: Overview): Describes fundamentals of master data quality and specifies requirements on both data and organizations to enable master data quality;
- Part 110 (Master data syntax and semantics): Specifies general, syntax, semantic encoding and data specification requirements for master data messages between

[1] https://www.dataqualitypro.com/iso-8000-data-quality-certification-options/

organizations and systems. The focus of ISO/TS 8000-110:2008 is on requirements that can be checked by computer;

- Part 115 (Quality Identifier Prefixes): Specifies the requirements for the quality identifiers that form part of an exchange of master data;
- Part 120 (Master data provenance): Describes requirements for representation and exchange of information about provenance of master data and master data sets;
- Part 130 (Master data accuracy): Describes requirements for the capture and exchange of data accuracy information and a conceptual model for data accuracy information in the form of representations and warranties of data accuracy;
- Part 140 (Master data completeness): Specifies requirements for representation and exchange of information about assertions of completeness of master data;
- Part 150 (Quality management framework): Contains an informative framework that identifies processes for data quality management.

The ISO 8000 family is underpined by a few fundamental principles, which are:

- People: data quality management is a people based activity and not merely a technology implementation;
- Process: effective management is based upon a number of key processes;
- Continuous improvement: as well as striving to continuously improve the quality of data, the processes used to achieve this should also be continually improved.

Among the standards of the 8000 family, ISO 8000 part 150 is the one that most resembles the Louvre. It propose a model to manage quality of master data. This model is divided into three vertical 'processes' and three horizontal 'roles'. The three generic roles are: Data Manager, Data Administrator and Data Technician. The three key processes are as follows:

- Data Operations: Focus on the factors that affect data quality and the usage of data;
- Data Quality Monitoring: Defines a systematic approach to assess the levels of data quality;
- Data Quality Improvement: Corrects data errors detected and eliminates the root causes of data errors.

For many scenarios, process levels are probably too simplistic, however, they do provide an indication of whether the processes can be implemented. Actually, all ISO 8000 standards are very simple. For instance, standards core content covers just a few pages. These standards also have a clear and limited intended use, which is the exchange of Master Data between organizations. This design requirement turn difficult to adapt to other kinds of content, such as non-organizational data or metadata. Like DMBOK, ISO 8000 standards have also institutional focus.

2.4 CONSIDERATIONS OF THE CHAPTER

Approaches about digital curation, data management and metadata management exist way before the appearance of the modern Data Ecosystems. Thus, there are numerous of works addressing these issues. In this Chapter, we first introduced the basic concepts that are related to metadata curation. Then, we provide an overview of the main related work. In particular, we analyzed studies and solutions for digital curation, as well as data management standards. About the latter, the emphasis was placed on the family of standards ISO 8000 (ISO, 2011) and the DMBOK (MOSLEY et al., 2010).

During these studies, it was possible to observe that the curation of metadata was not explored in depth by the literature. All the presented models focus on data instead of metadata. Themes common across all the presented curation and management models include planning the curation project; gathering, processing, analyzing, describing and storing the data, and archiving the data for future use. In addition, their overarching concerns include standards, governance and policy, planning, risk management, evaluation, and metrics, sustainability, and outreach. The presented models would be adapted to curate metadata. However, in practice, these curation models have had limited impact even in the maintenance of data (MAUTHNER; PARRY, 2013).

The majority of curation works are based on the proposition of lifecycle. Lifecycles are useful approaches to frame the stages related to curation of metadata. Although useful, the idea of a lifecycle can be difficult to employ, since every data lifecycle is different depending on the needs, aims and approaches of their potential users. Moreover, these models recommend to follow a linear perspective for the curation work. However, the relationships between metadata curation activities are non-linear and bidirectional in many cases. Actually, every curation activity can benefit from experiences acquired throughout the metadata curation initiative.

Despite both ISO 8000 and DMBOK propose a metadata curation management process, these frameworks classify the metadata management as an atomic process, disregarding the complex nature of metadata curation. Moreover, their enterprise focus make difficult to adapt their processes to the distributed and autonomous environment of Data Ecosystems. These frameworks depend of organizational structures that are not common on several Data Ecosystems. Moreover, these frameworks demand some significant effort

and investment to start. A more gradual evolution and systematic procedures are more suited to Data Ecosystem context.

Based on the presented works, we may notice that a promising alternative is to propose a metadata curation framework aimed at describing a series of elements for guiding and supporting the creation of metadata curation initiatives, where each element describes a coherent set of engineering and management practices in metadata curation. This framework construction need to be underpined in key factors identified from previous works ((GOBLE et al., 2008; CURRY; FREITAS; O'RIÁIN, 2010; KOUPER et al., 2013; QIN; BALL; GREENBERG, 2012; FREITAS; CURRY, 2016)) and in this study. These factors are related to metadata curation and specific characteristics of Data Ecosystems, which are:

- Support to metadata lifecycle: The framework must ensure support to all stages of the metadata lifecycle, through creation and collection, storage, manipulation, sharing and collaboration, publishing, archiving and use.
- Involvement of actors: Actors must collaborate to the curation metadata work.
- Adaptability and flexibility: Curation of metadata should embrace adaptability and flexibility in order to adapt and adjust by meeting different uses and contexts.

Considering the key factors presented above, models proposed until now fail in enabling a metadata curation environment, in a distributed, collaborative, flexible and adaptive way. Thus, a metadata curation framework needs to give those who use and contribute to the metadata a sense of ownership and control in order to curate metadata in a practical manner. That is, Data Ecosystem actors should be involved with the capture and the collection of metadata, the preservation of metadata, and other tasks related to the metadata maintenance.

Given this context, this work considers the agile and open collaboration paradigms as sources of inspiration for the development of a framework for metadata curation that balances the needs for documentation and micro-management with a more collaborative, practical, flexible and adaptive effort of Data Ecosystem actors.

Next Chapter presents the state of the art on Data Ecosystems highlighting their origins, principles, issues, features, and other essential issues.

3 STATE OF THE ART ON DATA ECOSYSTEMS

The way that individuals and organizations are producing, sharing and consuming data has changed with the advent of new technologies. As consequence, data became tradable and value good. There are now Data Ecosystems, in which communities of actors interact with each other to exchange, produce and consume data. Data Ecosystem is an area that has been gaining popularity in the last five years.

An example of Data Ecosystem is a community of healthcare institutions. In the healthcare industry, a large volume of datasets has been produced (MURDOCH; DETSKY, 2013). For instance, the Electronic Health Record (EHR) systems alone collect a huge amount of data (MURDOCH; DETSKY, 2013). Every patient has his own digital record which includes demographics, medical history, allergies, laboratory test results etc. Records are shared via secure information systems and are available for providers from both public and private sector. Over the last decade, pharmaceutical companies and research institutions also have been aggregating years of research and development data into medical databases. Medical researchers can use large amounts of data on treatment plans and recovery rates of cancer patients in order to find trends and treatments that have the highest rates of success in the real world (LEBIED, 2018). Another example hospitals can use data from a variety of sources to come up with daily and hourly predictions of how many patients are expected in a specific time interval. In this example, the EHR systems, pharmaceutical companies and research institutions act as data producers. In addition, the medical researchers and hospitals act as data consumers. These data collections are provided by public and private organizations involved in health treatment, city planning, traffic monitoring and security enforcement.

The emergence of Data Ecosystems has been driven by several factors, including the emergence of digital technologies and political/institutional initiatives. For instance, most Data Ecosystems have been mainly driven by the Open Data movement and Open Government Data (OGD) programs, which call for the free use, reuse and redistribution of data by anyone (OPEN GOVERNMENT WORKING GROUP, 2007). Several governments have already launched Open Data Portals to stimulate and promote Open Data production and consumption (CHUN et al., 2010). Improvements in the technology (*e.g.*, mobile Internet or technology) and trends in the technology (*e.g.*, social media or mobile apps) also have been driving private and public organizations to publish data as well as to integrate their services with external data.

In this Chapter, we provide an overview of the current literature on Data Ecosystems by presenting a subset of the results found from a wide systematic mapping study published in (OLIVEIRA; LIMA; LÓSCIO, 2019). In particular, this Chapter is intended to function as a snapshot of the research in the Data Ecosystem area by (i) analyzing the

evolution of Data Ecosystem research, (ii) defining and analyzing how Data Ecosystems are structured and organized, (iii) analyzing how Data Ecosystems are created, (iv) what the available infrastructure is, (v) identifying the actors and their role as well as the Data Ecosystem organizational structure, (vi) presenting an understanding about models and frameworks that can help in understanding the process and activities related to Data Ecosystem, (vii) analyzing how the Data Ecosystems management and the generation of value are, and (viii) presenting Data Ecosystem theoretical foundations.

3.1 REVIEW PROTOCOL

The work developed in this literature review adopts as a methodological reference a combination of the following approaches: (PETERSEN et al., 2008; KITCHENHAM; CHARTERS, 2007). It sets out to provide an overview of this field by undertaking a Systematic Mapping Study, which is a protocol-driven review. Figure 9 illustrates the adopted protocol, whereas the individual steps are explained in Subsections 3.1.1-3.1.4. The applied review protocol is focused on a set of research questions.

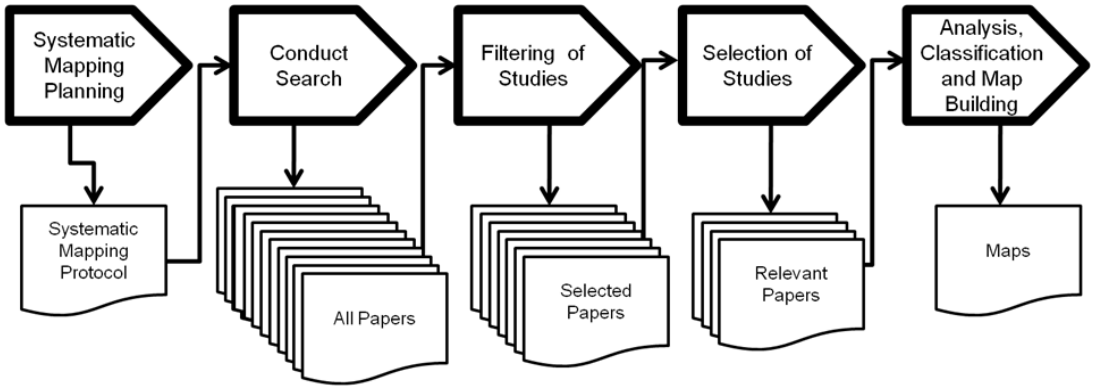

Figure 9 – Systematic literature mapping process. Adapted from: (LOPEZ-HERREJON; LINSBAUER; EGYED, 2015)

3.1.1 Research Questions

The purpose of this systematic literature review is to provide an overview of the research reported in the field of Data Ecosystems. For this, the following general research question was defined: *RQ: What is the current state of the art regarding Data Ecosystems research?*

We also defined 13 specific research questions to guide and structure the extraction and analysis of research data, and the synthesis of evidence. In particular, these specific research questions are used to categorize and quantify the key contributions and the evolution of research concerning Data Ecosystems, as well as to discover issues that still need to be explored and the limitations of existing studies.

- RQ1: What has been the evolution in the number of Data Ecosystems studies over the years?

- RQ2: What level of focus (full and partial) do the studies dedicate to Data Ecosystems research?
- RQ3: Who are the main publishers of Data Ecosystems studies?
- RQ4: Which individuals and organizations are the most active in Data Ecosystems research?
- RQ5: What types of contributions are reported by the studies?
- RQ6: Which topics or themes have been addressed by Data Ecosystems studies?
- RQ7: Which research methods have been used in Data Ecosystems studies?
- RQ8: How is the term "Data Ecosystem" defined?
- RQ9: What theoretical foundations are adopted by Data Ecosystems studies?
- RQ10: What are the main roles of actors in a Data Ecosystem?
- RQ11: How are Data Ecosystems structured and organized?
- RQ12: What are the main domains in which Data Ecosystems have been and are being developed?
- RQ13: What is currently known about the benefits and limitations of Data Ecosystems?

One of our initial aims was to identify various essential elements and features of Data Ecosystems. The definitions and concepts presented by the studies analyzed assisted characterizing Data Ecosystems. The identified concepts provided some clarity on what is known about Data Ecosystems. Furthermore, we also obtained an overview of real-world Data Ecosystems by mapping which of them were their drivers, how these ecosystems were created and what the available infrastructure is. This consolidated knowledge about Data Ecosystem can be a useful resource for researchers and practitioners looking for understanding Data Ecosystems. The full and detailed analysis of each question is available in (OLIVEIRA; LIMA; LÓSCIO, 2019).

3.1.2 Data Sources and Search Strategy

The strategy for collecting the relevant literature was to undertake a keyword search in five academic libraries (Table 4). The automated search process which took place in March 2017 retrieved 244 studies. The keyword query string (Figure 10) consists of synonyms of the search term "Data Ecosystem" defined according to our expertise. Also, we used wild-card characters to capture the plural and singular forms of the keywords. The query

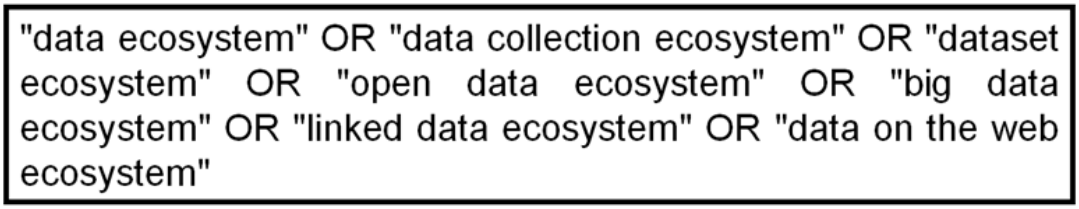

Figure 10 – Keyword-based query used to automate for search studies. Source: Author

string was intentionally kept simple, with no predicate, which enabled us to extract the maximum number of studies containing the terms.

From the initial set of 244 studies, we selected studies that mentioned concepts, theories, guidelines, discussions, lessons learned, and reports about the Data Ecosystem field (inclusion criteria). We excluded studies that fell into any of the six Exclusion Criterias (ECs), which were as follows:

- EC1: The study is not peer-reviewed (*e.g.*, presentation slides, extended abstracts, invited papers, keynote speech, workshop reports, books);
- EC2: The study contains less than one page;
- EC3: The study is not written in English;
- EC4: The study is not accessible on the Web;
- EC5: The study does not present any type of findings or discussion about Data Ecosystems;
- EC6: The study is duplicated.

3.1.3 Selection of Studies

The automatic search for and selection of studies was led by the first and second authors, who conducted the steps presented in Figure 9. Many of the retrieved studies were eliminated in the filtering step, during which the researchers evaluated the studies by looking at the title, abstract and venue information. After the exclusion criteria were applied, 48 potentially relevant studies remained.

After the filtering of studies step, the researchers worked on the selection step. Initially, the 48 potentially relevant studies were analyzed by both researchers, each of whom worked independently. The researchers applied the inclusion and exclusion criteria on the complete texts of potentially relevant studies. The differences were resolved in a meeting at which all authors reached a consensus. Duplicated studies were excluded in this step. However, for studies published in more than one academic library, all versions were reviewed for the purpose of data extraction.

Table 4 – Sources and number of studies

Academic Library and Snowballing	Studies Retrieved	Studies Excluded	Studies Included
IEEE	27	22	5
ACM	20	19	1
Science Direct	18	14	4
Scopus	105	93	12
Springer	74	73	1
Snowballing	6	0	6
Total	250	221	29

Moreover, we also applied the snowballing technique to find the hidden population of studies, *i.e.*, those studies that were not returned by our automatic search process but that could be interesting for the research. According to Jalali e Wohlin (2012), snowballing, when compared to automated searches in databases conducted in systematic literature searches, is less costly and brings less "noise" (articles that will not be selected). In particular, we used the backward snowballing approach that looks for new studies in the references of our list of studies. The whole selection process considered 29 studies as relevant ones for the data extraction and analysis (*cf.*, Table 4).

During the filtering and selection phases, we kept track of the rationale for each exclusion, as shown in Table 5. Although all studies mention the term "Data Ecosystem", a significant number of them do not investigate Data Ecosystems narrowly nor do they take a broad view. In fact, a large number of the studies used the term "Data Ecosystem" only once throughout the text or used the term as a keyword so that the study would be flagged during a search. Another set of excluded studies present a solution for data publication or data consumption, but they do not, for instance, focus on the relationships between data producers and other actors within an ecosystem. Moreover, 40 studies are indexed by two or more academic library. In most of the cases of duplication, we chose the study from the Scopus library, since this library retrieved the largest number of studies.

Table 5 – Rationale for excluding papers

Rationale	No. of Studies
Do not address Data Ecosystems	154
Duplicated	40
Paper is not an academic work or peer-reviewed work	21
Not in English language	4
Document body no longer than one page	2
Total	**221**

3.1.4 Data Extraction and Synthesis

The data were extracted using an electronic spreadsheet in Google Spreadsheets™. Each study was analyzed and the information collected was recorded on a form. Each researcher worked independently to extract data from the whole universe of studies. In the end, the most experienced researcher reviewed, and if necessary, revised all the data extracted by analyzing the studies again. This revision activity is meant to improve the accuracy of the extraction process and, therefore, the reliability of the results. Also, conflicts during the extraction phase were discussed and resolved by consensus at a meeting with the three authors of this study.

The results from the data extraction phase were integrated on spreadsheets, which were also used to generate mind-maps and tables. All descriptive information was calculated and organized using Google Spreadsheets™. The data extracted from each study were synthesized using thematic analysis and qualitative coding techniques. The synthetic data was organized into mind-maps. Each row of a spreadsheet as well as the synthetic mind-maps were reviewed more than once by at least two researchers.

3.2 FINDINGS

In this section we analyze the literature and the results of the systematic mapping study. From an initial sample of 250 studies, we identified 29 primary studies. Table 6 presents the complete list of references of these studies.

3.2.1 Data Ecosystems Research Evolution

We analyzed the evolution in the number of studies related to Data Ecosystems published over the recent years (*cf.* Figure 11). The first Data Ecosystem studies were published by Tim Davis [S03] and Ding *et al.* [S08] in 2011. The former study claims that data initiatives require an ecosystem that consists of Open Data infrastructures and standards in order to actively encourage the use of Open Data as well as the development of new technologies. Such a Data Ecosystem involves mobilizing a wide range of technical, social and political resources, and the need for interventions beyond data supply to support the coordination of activity around datasets. The latter study presents a platform called TWC LOGD Portal, which encompasses a model infrastructure that supports linked open government data production and consumption. By using this platform, a community of actors may interact to form an ecosystem. These studies neither create a theory nor define Data Ecosystems. However they provide the first insights into the field. It is important to remark that the most cited oldest references in the literature are the studies (POLLOCK, 2011) and (POIKOLA; KOLA; HINTIKKA, 2011).

We consider that a total of 29 articles on such an important topic to be a very small number. However, the number of studies might have increased because we did not consider

Table 6 – List of Selected Studies

Study ID	Reference	Study ID	Reference
S01	(SMITH; OFE; SANDBERG, 2016)	S16	(ZUIDERWIJK; JANSSEN; DAVIS, 2014)
S02	(MERCADO-LARA; GIL-GARCIA, 2014)	S17	(MAGALHAES; ROSEIRA; MANLEY, 2014)
S03	(ZELETI; OJO, 2016b)	S18	(LEE, 2014)
S04	(MOISO; MINERVA, 2012)	S19	(BOURNE; LORSCH; GREEN, 2015)
S05	(SHIN, 2016)	S20	(ATTARD; ORLANDI; AUER, 2016)
S06	(HA; LEE; LEE, 2014)	S21	(LINDMAN; KINNARI; ROSSI, 2016)
S07	(ZUBCOFF et al., 2016)	S22	(IMMONEN; PALVIAINEN; OVASKA, 2014)
S08	(DING et al., 2011)	S23	(ZELETI; OJO, 2016a)
S09	(ZUIDERWIJK et al., 2016)	S24	(ZUIDERWIJK et al., 2015)
S10	(DAWES; VIDIASOVA; PARKHIMOVICH, 2016)	S25	(HEIMSTÄDT; SAUNDERSON; HEATH, 2014)
S11	(DONKER; LOENEN, 2017)	S26	(HARRISON; PARDO; COOK, 2012)
S12	(KÖSTER; SUÁREZ, 2016)	S27	(DAVIES, 2011)
S13	(KOZNOV et al., 2016)	S28	(ZELETI; OJO, 2014)
S14	(SHIN; CHOI, 2015)	S29	(GAMA; LÓSCIO, 2014)
S15	(SCHALKWYK; WILLMERS; MCNAUGHTON, 2016)	-	-

studies that were published in 2017. Moreover, with the exception of 2013 and 2015, the average number of publications has been growing since the first study was published in 2011. The growth in the number of studies published has been strongly influenced by both the Open (Government) Data movement and the Big Data market.

3.2.2 Data Ecosystem Terminology

There is little agreement about nomenclature and the definition of Data Ecosystems. Although a discussion in greater depth on terminology is beyond the scope of this thesis, in order to be able to analyze the Data Ecosystem field as well as to guide the study process, we should first review some of the existing definitions. Hence, an important goal of this work is to provide an overview of how the research community defines the term Data Ecosystem.

Table 7 shows the studies split into two different groupings, namely those that include a definition for a Data Ecosystem and those that do not. Our first finding is that there is a large number of studies (13 studies, over 44%) that do not define the term Data Ecosystem. However, some of these studies make reference to previous studies that had done so. In most cases, this happens because these studies only partially focus on Data

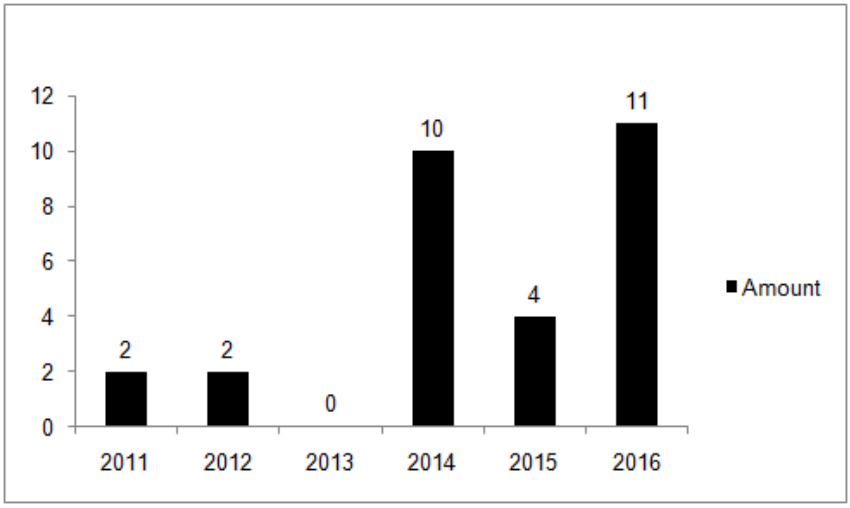

Figure 11 – Distribution of the number of studies published regarding Data Ecosystems between 2011 and 2016. Source: Author

Ecosystem research. In some cases, the authors consider a previous study (written by themselves or other researchers) as the basis for the background and definitions needed for what was then their most recent study.

Table 7 – Studies classified according to whether or not they offer an explicit definition for a Data Ecosystem.

Definition	**Studies**
Available	S01, S04, S05, S07, S08, S09, S10, S13, S14, S15, S16, S18, S20, S22, S26, S29
Not Available	S02, S03, S06, S11, S12, S17, S19, S21, S23, S24, S25, S27, S28

With regard to the studies that provide a definition, 7 of them provide a new definition (*i.e.*, with their own words) for Data Ecosystems. The other 10 studies define the field by using one or more definitions from the existing literature. Furthermore, most of the referenced works (*cf.* Table 8) do not provide a formal definition. They only conceptualize aspects about Data Ecosystems. For instance, instead of providing a definition for Data Ecosystems, S25 and Ubaldi (2013)[1] identify a set of structural Data Ecosystem properties (*i.e.*, circular flow of resources, sustainability, demand that encourages supply, and dependence developing between actors). These aspects are used to construct definitions by other works.

Pollock (2011) provides the oldest definition cited in the selected works. According to him, *"An ecosystem has data cycles, in which intermediate consumers of data such as builders of apps and data wranglers may share back their cleaned, integrated, and packaged data into the ecosystem in a reusable way. This cleaned and integrated data is often more valuable than the original source"*. This definition emphasizes the need for data cycles in

[1] Despite used as a source for Data Ecosystem definitios, Ubaldi (2013) and Pollock (2011) are not a scientific study

order to create Data Ecosystems. Besides the cycle, Pollock's vision requires that actors should play roles, such as infomediaries and consumers.

Harrison *et al.* [S02] envision the idea of a Government Ecosystem, which is a kind of Open Data Ecosystem. According to them, *"A Government Ecosystem envisions government organizations as central actors, taking the initiative within networked systems organized to achieve specific goals related to innovation and good government"*. They also complement this definition by stating that the *"ecosystem metaphor [is] often used by policy makers, scholars, and technology gurus to convey a sense of the interdependent social systems of actors, organizations, material infrastructures, and symbolic resources that must be created in technology-enabled, information-intensive social systems, among them, open government"*. Like Pollock (2011), Harrison *et al.* advocate that roles should be defined, but they also emphasize the idea of a keystone role that controls and coordinates the ecosystem. Moreover, they also recognize a set of contextual factors (*e.g.*, social aspect) as a key element of Data Ecosystem. Furthermore, Harrison *et al.* pair the ecosystem metaphor with the concept of multiple and varying interrelationships between producers, users, data, material infrastructure, and institutions.

Table 8 – Previous papers most cited by the studies selected

Most Referenced Papers	**Referencing Studies**
Pollock (2011)	S10, S16, S17, S18, S25, S26
Ubaldi (2013)	S01, S07, S10, S11, S13, S16, S23, S28
Heimstädt, Saunderson e Heath (2014) [S25]	S07, S10, S18, S16
Harrison, Pardo e Cook (2012) [S26]	S10, S15, S16, S18, S25

A notewothy example is study S25 which is referenced by 4 studies. Heimstadt *et al.* [S25] conceptualize Data Ecosystems by identifying a set of structural properties: circular flow of resources, sustainability, demand that encourages supply, and dependence developing between suppliers, intermediaries, and users.

Other definitions are provided by S07, S08 and S09. According to S07, a Data Ecosystem *"is made up of many actors and small organizational structures that should recognize data like the raw material that is in a cycle and is capable of feeding the ecosystem, providing benefits to all parties."*. This Data Ecosystem view also advocates for a cycle as well as pointing to there being multiple actors, each with their own expectations.

A different perspective is presented in S09, which defines Data Ecosystems as *"all activities for releasing and publishing data on the Internet, where data users can conduct activities such as searching, finding, evaluating and viewing data and their related licenses, cleansing, analyzing, enriching, combining, linking and visualizing data and interpreting and discussing data and providing feedback to the data producer and other stakeholders"*. Similarly, S08 defines Data Ecosystem as *"a data-based system where stakeholders of different sizes and roles find, manage, archive, publish, reuse, integrate, mashup, and*

consume data in connection with online tools, services, and societies". Both definitions present the idea of activities that would develop some value or benefit to actors who use data. These activities can be assigned to specific roles that will be performed by actors (*i.e.*, stakeholders).

As matter of fact, despite Data Ecosystem being a term that has been used by a number of studies, the meaning of Data Ecosystem is still very much under construction. It seems that to do so, a rich framework is required to support practitioners and researchers so that they may acquire a much fuller understanding of the current state and potential future of Data Ecosystems.

3.2.3 Data Ecosystem Characterization

Ecosystem metaphor is often used "to convey a sense of the interdependent social systems of individuals, organizations, material infrastructures, and resources that can be created in technology-enabled, information-intensive social systems" [S02 apud S16]. This view is similar to that of S14, who argue that Data Ecosystems are a cultural, technological, and social phenomenon based on the interplay of technology, actors, businesses, industries and governments. Similarly, S10 argues that Data Ecosystem understanding should place the emphasis on an evolving, self-organizing system of feedback and adjustment among actors and processes. S16 emphasizes the multiple and varying interrelationships between data, actors, infrastructure, service, tools and other different kinds of resources. These connected components are extremely interdependent [S16]. These ecosystem views suggest that an ecosystem is a socio-technical system, the design and analysis of which should be based on a contextual understanding of human interactions that are served by technology, but also on practices, culture and the political and economic context [S15]. Indeed, successful Data Ecosystem initiatives emphasize the importance of considering all components that constitute a socio-technical system [S12][S14][S18][S07].

With a more actor-centered perspective, Data Ecosystems can be defined as a community of actors of data sharing initiatives such as data producers, end-users and intermediaries S16. These actors dynamically interact by exchange of resources (*e.g.*, data, services, tools, infrastructure). These interactions forms a kind of relationships, some of them represent short-time or long-standing relations S10. Furthermore, S16 argues that Data Ecosystems could be seen as kind of exchange network, in a which there could be cooperation of various actors to facilitate the use and provision of data. Often keystone actors play a critical role in facilitating exchange in the ecosystem thereby ensuring dynamism and constant functioning S15. A keystone actor normally occupies or creates highly connected hubs of actors and promotes the health of the ecosystem by providing value to the surrounding actors (MANIKAS; HANSEN, 2013b).

Although actors are the central element of a Data Ecosystem, they are often the most neglected. Some studies have found that many of data users consider quality of

Data Ecosystem resources very low, especially data resources. Such barriers may hinder the generation of benefits. However, actors expect positive and beneficial impacts when exploit and explore Data Ecosystem resources. It is pursuit of benefits that keeps the ecosystem in motion.

In general, a Data Ecosystem addresses the multi-aspect nature of data sharing initiatives, encompassing several dimensions such as the social, legal, political, cultural, technological, operational and economic ones. Such a holistic nature emphasizes that all aspects of an ecosystem are interconnected, influenced by values [S10]. As a matter of fact, Data Ecosystem initiatives in different countries have faced many barriers because of differences in government organization, legislation and other factors [S13]. Therefore, the design and study of a Data Ecosystem should include seeking to understand the context.

According to S15, in Data Ecosystems, there are at least three contextual conditions that can motivate or constrain the functioning of Data Ecosystem elements. The first of these is the regulatory context, which includes laws, policies, standards and agreements. This context guides how to specify how the elements of the ecosystem are structured and how they interrelate [S15]. The second is the institutional context in which the actors operate. Each institutional context provides values, rules and norms that inevitably propel and restrain the behaviors of actors in the ecosystem [S15]. The third is the technological context, which encompass the IT resources, the IT operators and other enabling technologies that connect and interconnect the Data Ecosystem elements [S15]. The environmental context, such as cultural, social and economical elements, also exerts an influence on Data Ecosystem initiatives [S18].

With respect to other Data Ecosystem properties, the current literature also conceptualizes that Data Ecosystems are cyclical, sustainable as well as being demand-driven and there is dependence between suppliers, intermediaries, and users [S25]. By cyclical, the authors state that any resource should be processed cyclically. Sustainability is understood as the ability to continue a defined behavior indefinitely (THWINK ORGANIZATION, 2014). Analyzing the UK Data Ecosystem, S25 found "clear signs of a cycle, of sustainability, of demand encouraging supply, and of dependence developing between suppliers, intermediaries, and users." With regard to demand-supply, Data Ecosystems instead of presenting a mutual interdependence, often demonstrate more of a one-sided dependency, *i.e.*, data users depend on data producers, or vice-versa [S25]. Despite these gaps and drawbacks, the structural properties presented are still desirable, and Data Ecosystem emergence should be fostered by the self-organization aspects rather than the explicit design goals of conventional IT [S14].

3.2.4 Data Ecosystem Elements

Regarding Data Ecosystem elements, in a literature review, S02 points to three important domains of elements, which are: (i) government policies and practices, (ii) innovators; a

combination of technology, business and government and (iii) users, civil society and business. S14 also identifies as key elements of a Data Ecosystem (i) infrastructure, (ii) software and technologies, (iii) service and applications, (iv) standards, (v) users, (vi) social and cultural factors, (vii) government and (viii) industry.

In order to create and capture value, S23 states that actors must employ emerging sets of capabilities, which are types of skill. Their function allow an activity to be performed or even to improve the productivity of some activities. These capacities can include a range of skills and expertise, such as general knowledge about Data Ecosystem resources (*e.g.*, systems and technologies), technical skills required to manage and to process data (*e.g.*, data integration and data mining techniques), and operational expertise to incorporate data related resources into existing institutional and business processes [S18].

Besides capabilities, actors also require proper resources in order to provide and consume data. Common examples of resources are datasets, services, tools, financial capital as well as human capital, equipment, materials, and proprietary technology. S24 distinguishes between three categories of resources: human resources, data resources and IT resources. Human resources refer to individuals who use their capabilities to explore and exploit data. Data resources refer to the static and dynamic data-based assets, such as databases, knowledge bases or simply datasets. IT resources refer to hardware (*e.g.*, infrastructures, networks and computers), platforms and applications (software). Actually, actors do not necessarily need to own, manage, or operate the underlying resources, but can consume or contract such resources through other actors, such as service providers or other kinds of intermediary actor. S24 also emphasized that these resources often need to be combined to be able to address an actor's expectation. Not only data is a primary resource, but also various kinds of resources are complementary and are needed so that Data Ecosystemsfunction properly.

S02 emphasized dependency relations in Data Ecosystems. According to these authors, all Data Ecosystem elements are interconnected in a way that when one element is changed, effects can be felt throughout the whole system. In fact, actors affect and are affected by the creation and delivery of resources performed by the other actors [S22]. Moreover, actors have their own interests and benefits which could lead to conflicts. In particular, data consumers are strongly influenced by the decision of a data producer to not publish or update a certain piece of data (anymore), to change the format in which the data is published, to compromise the quality of the data or to change how it can be used [S03].

3.2.5 Data Ecosystem Actors and Roles

In Data Ecosystems, actors can play several roles. Actors are autonomous entities such as enterprises, institutions or individuals, which act in the Data Ecosystem in one or more specific roles. The actors are motivated by a set of interests [S16]. Actors involved in the

ecosystem usually have a commitment to the ecosystem. They may have an incentive for being active in the ecosystem.

A role is a function played by an actor in the ecosystem. It is related to a position and a set of duties. Several roles can be identified in Data Ecosystems. Typically, at least the roles of a data user and a data producer are identified in contemporary Data Ecosystems. However, there are several additional roles, each of which undertakes different duties. Moreover, it is possible to find two different roles sharing the same duty. For instance, both data intermediaries and data producers provide data. Furthermore, Data Ecosystem characteristics strongly determine the need for a particular role. In fact, some roles exist only in very specific ecosystems. For example, in Data Ecosystems based on medical/health data, typically there are roles responsible for assessing ethical issues.

Like Data Ecosystem terminology, there is little agreement about the actors' roles and their respective duties and activities. According to Lundell et al. (2009), defining the roles in each ecosystem is essential for researchers and business analysts in order to understand and manage an ecosystem and estimate its success.

Table 9 lists the roles identified in the studies. In total, we identified 13 major roles and a further 22 minor roles (*i.e.*, a specialized role that is responsible for some of the duties and activities of a major role). However, in most studies, both how the roles are defined and how their activities and duties are set out are underspecified or not specified at all. Moreover, the large majority of the studies only list Data Ecosystem actors (*i.e.*, stakeholders) such as citizens, developers, governmental organizations and private companies, leaving it to the reader to identify and classify their roles.

Table 9 – Roles of Data Ecosystem actors

Role	Studies
Data Provider	S01, S03, S05, S06, S07, S09, S11, S12, S13, S14, S15, S17, S19, S20, S21, S22, S23, S24, S25, S27, S29
Storer	S22
Developer	S22
Aggregator	S22, S29
Harmonizer	S22
Publisher	S22
Register	S22
Data Maintainer	S19
Data Producer	S02, S04, S08, S16, S19, S26,
Data Owner	S21
Policies, Laws and Rules Parties	S03, S09, S11, S10, S15
Policy makers	S09, S11
Standardized and regulation parties	S03, S15
Keystone Actors	S10, S15, S19, S26
Founder	S19
Service Provider	S05, S06, S14, S21, S22, S29
Data-as-a-Service providers	S22
Analytics-as-a-Service providers	S22
Data Service Developer	S01, S29
Re-user	S01, S02, S08, S12, S18, S17, S21, S22, S23, S26, S29
open data-driven organization	S23
Application developer	S08, S22
Interpreter	S21, S22
Data Intermediaries	S01, S02, S04, S09, S15, S16, S17, S21, S23, S25, S26
Data Brokers	S01, S04, S22
Enablers	S17, S23
Integrators	S17
Data Extractor and Transformer	S21
Data Analyser	S21
Data Visualizer	S21
Data User/Data Consumer	S01, S02, S03, S04, S05, S06, S07, S08, S09, S10, S11, S12, S13, S14, S15, S16, S17, S19, S20, S21, S22, S23, S24, S25, S26, S27, S28, S29
Data Curator	S08, S19
Infrastructure Provider	S06, S21, S22
Data Sponsor	S19, S29
Data Consultant	S01, S22

The role most often identified is Data User, which is responsible for directly or indirectly consuming data. Almost all the studies present the concept of a Data User role. However, this role is presented with a myriad of names in the studies, such as end-users, data consumers, data beneficiaries, etc. Data Users do not necessarily have the ability to consume data directly from data producers. They usually rely on services provided by

Re-users (users who create value out of the available data sources), Data Intermediaries or Service Providers. Moreover, Data Users usually represent the end-users of an ecosystem.

The second most highlighted role in Data Ecosystems is data producer, which is responsible for data supply or provision. Almost all the studies (21 studies) present the concept of a data producer role in a very similar way to that of the Data User. There are also a few studies that present minor roles related to the provision of data. For instance, S22 presents different data producer minor roles related to supplying data, such as "Storer to collect and save raw material, a Developer to manage and process raw material, an Aggregator to combine and edit data from different sources, a Harmonizer to standardize and homogenize data from different sources, an Updater to update information, a Publisher to publish the data, and a Register to maintain the administration of data resources".

It is important to note that data producers are not necessarily responsible for data generation. This responsibility may be assigned to another role, called Data Producer, which is responsible for the capture or generation of data. This role may also compile, aggregate, and package data. In particular, 6 studies identify the Data Producer role.

Another identified role is a Re-user responsible for adding value to the data to be re-used. According to S12, Re-user is responsible for using data to develop applications or services aimed at Data Users. The Re-user role is identified in 11 studies. There are several studies presenting the Re-user's minor roles. For example, [S22] defines the Application Developer role (*i.e.*, they use the data as part of the service) and Interpreter (*i.e.*, they interpret data for end-users). Another specialization of Re-users is Data-driven Organization [S23], which represents the private companies that use, transform, or invest in Open Data.

The Keystone Actor role is responsible for driving forces behind the ecosystem as well as providing stability in unstable environments (IANSITI; LEVIEN, 2004). This role is very common in Software Ecosystems. In S10, the researchers claimed that in Data Ecosystems, Keystone Actors are responsible for providing most data as well as for promoting the ecosystem. S18 states that this role must be assigned to actors who lead the OGD programs. Under this scenario, Keystone Actors should foster a set of formal directives, rules, and practices to drive a Data Ecosystem. Taking a slightly different view, S15 states that Keystone Actors are enablers, not necessarily drivers in the ecosystem; They are useful but they are not essential to the continued functioning of an ecosystem.

The Service Provider is responsible for producing and supporting software resources such as tools, applications, services, libraries, or other software products [S05][S06][S14][S21] [S22][S29]. Actors that do not have the abilities and resources to perform the data processing themselves can contract Service Providers [S22]. S22 presents two minor roles (*i.e.*, Data-as-a-Service providers and Analytics-as-a-Service providers) related to service provision based on the cloud computing paradigm.

Introduced as one of the most important roles in S10, the Policies, Laws and Rules

Parties represent the role responsible for creating the rules and policies to encourage and to control the participation of actors [S03][S09][S11][S10][S15]. Typically, this role is performed by Governments, Data Ecosystem Founders or Standardization Institutions (*e.g.*, Open Knowledge Foundation and W3C).

Other major roles identified in the literature are the Infrastructure Provider, Data Consultant, Data Sponsor, and Data Curator. Infrastructure Provider is the role that supports the activities of other roles. Infrastructure includes the provision of Information and Communication Technologies (ICT) resources or services such as hosting or storage capacity [S06][S21][S22]. A Data Consultant assists other roles to analyze the possibilities of data and also identifies the actor's needs [S22]. A Data Sponsor is responsible for promoting the Open Data initiative through both public funding programs and private investments [S19][S29]. Finally, Data Curators are responsible for the quality and availability of data [S19][S08].

Even though we have identified several different roles, there is still a myriad of different minor roles in the literature. However, a more in-depth discussion on classifying roles is classification beyond the scope of this research.

3.2.6 Data Ecosystem Structure

In a Data Ecosystem, each actor is connected to other actors by a set of interests or business models. The whole network of relationships may follow an organizational structure, ranging from an *ad-hoc* diffuse approach to a keystone-centric approach (HANSSEN; DYBÅ, 2012). An ecosystem organizational structure taking account of the way actors are connected and the properties of their relationships, such as relationship dependency (MANIKAS; HANSEN, 2013b). Studying the organizational structure of Data Ecosystems is important to understand and to govern the interaction and organization of actors (CHRISTENSEN et al., 2014). Table 10 shows the different groupings of Data Ecosystem structure found in the selected studies. We identified 5 different approaches: Keystone-centric, Data Intermediary-Based, Platform-Centric, Marketplace-Based and Business Model-Oriented.

Table 10 – Studies that describe a form of Data Ecosystem organizational structure

Data Ecosystem Structure	**Studies**
Keystone-centric	S07, S10, S12, S13, S15, S16, S18, S25
Intermediary-Based	S02, S04, S15, S21, S22
Platform-centric	S08, S12, S27,S29
Marketplace-Based	S01, S04
Business Model-Oriented	S06, S19, S20, S21, S22, S23, S24, S28

In the keystone-centric structure, actors are organized around a keystone actor, which is directly or indirectly responsible for providing much of the data. However, the keystone actor does not have complete control over the other actors. However, the keystone actor

does not have complete control over the rest of actors. They can leave (or enter) the ecosystem at any time. Hence, the keystone actor should be responsible for monitoring, evaluating, making decisions, and taking actions [S25]. There is also a specific kind of keystone-centric ecosystem performed by government administrations or public institutions which has been emerging from Open Data movements [S10][S12][S13][S18][S25].

The intermediary-based structure depends on the presence of data intermediaries in order to generate value from data. As mentioned in Subsection 3.2.5, a data intermediary is a role that facilitates the use of data for other actors. Therefore, in a Data Intermediary-based structure, data producers and data users (*i.e.*, the two extremes of a supply chain) are organized around the intermediaries. Some studies recognize the critical role that intermediaries can play in ecosystems [S15][S22][S21]. They also state that data intermediaries undertake a wide range of functions.

In the platform-centric structure, a platform defines the organizational structure of a Data Ecosystem. According to S10, the platform provides infrastructure and services to support both the provision and consumption of data. S10 and S08 emphasize that the costs of providing data are reduced when the data is released via the platform. The platform can also mitigate interoperability and usability problems. Open Data catalog tools (*e.g.*, CKAN[2]) are common, but limited examples of platforms are used to create Data Ecosystems [S29]. According to S29 and S03, Data Ecosystems should show not only data but also services. They need a platform that provides core services and allows developers to create new services. This platform can be based on a cloud computing infrastructure, for example [S22].

In marketplace-based structures, marketplaces provide the required infrastructures, business models, rules and services for transactions of data and software between actors [S01]. In general, marketplaces encompass a technical platform with the capacity to link data producers and data users. They also enable the sale of data, services, and applications. In this sense, S22 states that there are several pricing models suitable for data-related businesses. According to S22, "services and applications can be priced commonly based on features or performance, or the customer is charged a predefined price for customer-tailored services and applications usage".

Despite not defining how the actors should be organized, some studies present business models, which describe the rationale of how an actor creates, delivers, and captures value. In particular, value refers to any benefit that an actor obtains from the Data Ecosystem, such as satisfaction, utility, problem-solving or revenue. Common business models are business-to-business, business-to-consumer, and consumer-to-consumer ecosystems [S06]. According to Christensen et al. (2014), the business model is important in reasoning the cost, revenue, and/or sustainability of the Data Ecosystem. While S06, S19 and S21 did not present a well-established business model strategy, S20, S22, S23 and S28 and S24 are

[2] https://ckan.org/

based on the Value Chain Theory and Resource Dependency Theory, respectively.

3.2.7 Data Ecosystems Creation

A Data Ecosystem is initially created by some actors pursuing potential social and economic benefits [S11]. According to S02, Data Ecosystems are also created in part by executives and government administrations that want to achieve potentially beneficial factors [S02]. Moreover, within government, industry or academy, data are created, registered, processed, kept, used, shared and published for certain purposes [S12]. However, these available data are capable of other uses. In the case of governments, Data Ecosystems are not only driven by financial incentives or rewards. Many government administrations have encouraged society and private institutions to consume government data to create economic and social value [S15].

Similar beliefs across the world have spurred a growing number of private companies seeking to release internal data in order to improve their image and even increase profits by enabling external individuals and other companies to analyze information and come up with valuable findings and service/product innovation [S21][S23]. According to S22, private companies can earn direct profits from data sales in addition to which creating a Data Ecosystem centered on their private data can provide other benefits, such as new partners, new interests in the company's main products, new kinds of business activities and new customers. For instance, in 2011, the European Commission estimated the economic impact of directly and indirectly using Open Data in the 27 countries then in the EU to be about $140 billion annually [S21][S24][S23].

In fact, the move towards Data Ecosystems is not only desired, but inevitable. The marriage of political goals for transparency and contemporary ICT tools reduced the cost of information capture, management, and use. These new possibilities contribute to the pressure on private and public institutions to make data and documents available as well as not to disclose the data [S02].

Data Ecosystems can be seeded, modeled, developed, managed, *i.e.*, intentionally cultivated for the purpose of achieving a managerial and policy vision [S02]. In fact, many Data Ecosystems were created by using data sharing programs, which typically comprise a set of formal directives, rules, and practices that apply to all or most actors within a community or institution. Moreover, these Data Ecosystems may be conceived from scratch or by extending the existing infrastructure platforms, such as CKAN[3], which allow the amount of work needed to establish an ecosystem to be reduced.

S10 presents five different Data Ecosystem development approaches, namely: (i) Data-oriented approach focusing on the characteristics, quality, and availability of open datasets; (ii) Program-oriented approach addressing the purposes and features of a data sharing program with emphasis on data release initiatives; (iii) Use and user-oriented approach

[3] https://ckan.org/

focusing on the factors that influence data use by actors; (iv) Scorecard and impact approach addressing a wider array of considerations that influence how and how well Data Ecosystems work; and (v) Network-based approach focusing not only on Data Ecosystem elements but their dynamic relationships and their influence on the ecosystem performance. All these approaches allow the emergence of a Data Ecosystem to be stimulated. However, an ecosystem does not develop solely through top down governance. It is crucial to provide a role and benefits for all the Data Ecosystem actors, since an ecosystem depends on fruitful interaction between cooperating and competing actors [S25].

In this sense, S22 states that an ecosystem cannot be created, but it should naturally emerge. Data Ecosystem demands that actors are equal and each one actor should identify its role in the ecosystem S22. Moreover, the joining of the ecosystem should be fast and easy through registration, after registration, the actor should has access to all data in the ecosystem and the actors in the ecosystem cooperate in value networks S22. This suggests that monopolies should not exist in Data Ecosystems. Indeed, the existence of monopolies may compromise the Data Ecosystem sustainability. For instance, many Data Ecosystems are government-funded (*e.g.*, UK Open Data Ecosystem S02 and Russian Open Data Ecosystems). This kind of dependency creates a bottleneck for the ecosystem, as it is governments are the majority data holder and finantial sponsor.

3.2.8 Data Ecosystems Infrastructure

A number of infrastructures have been developed in recent years to create Data Ecosystem platforms in order to explore the potential of data, such as Open Data portals, the European Open Data infrastructure [4], infrastructures of statistics (*e.g.* Barometer[5]), data publishing solutions (*e.g.* CKAN, Junar[6] and Socrata[7]). Another alternative is to use services as building blocks so as to construct platforms for Data Ecosystems.

In general, these infrastructures support or accelerate the access and exchange of resources, mainly data, between the supplying actors and consumer actors. This may impose tailoring or repacking resources to fit to technologies such as Web applications, databases and APIs so that both computing-skilled actors and domestic actors can navigate, explore, and subsequently gain insights from the Data Ecosystem resources. Moreover, such infrastructures must also promote the interactions between the actors of the Data Ecosystem and they should also enable innovative products and services to be created in large-scale by using standard formats and common tools.

In fact, the literature on Data Ecosystems frequently advocates for efforts to establish rich new data infrastructures and standards, and to actively encourage the development of Data Ecosystems [S04][S14][S03]. However, providing a simple and reliable infrastructure

[4] https://www.eudat.eu/
[5] https://opendatabarometer.org/
[6] http://www.junar.com/
[7] http://www.socrata.com/

involves more than the release and aggregation of those datasets. The lack of standardized technologies and well-accepted guidelines may hinder ease of use and reduce productivity. For instance, the myriad heterogeneous solutions imply that actors must deal with different APIs and data access methods, the structural, syntactic and semantic heterogeneity of data, and the variability of quality levels and the diversity in types of metadata.

Promising alternatives are intermediary roles to simplify and promote the access to resources. For instance, Poikola, Kola e Hintikka (2011) define the intermediary roles Aggregator (to combine and edit data from different sources), Harmonizer (to standardize and homogenize data from different sources), Application developer (to utilize the data as part of the service) and Interpreter (to interpret the data). Hence, those actors that do not have the capabilities and resources to discover, access and process data themselves can delegate these activities to intermediary roles. Intermediary roles also can provide the infrastructure to support other Data Ecosystem activities.

Another option is the creation of marketplaces to facilitate trading and sharing of data, services and other resources. Since data itself becomes an asset, they may be traded in marketplaces as a commodity along with goods and services. For example, Microsoft Windows Azure Data Marketplace[8], as the name implies, integrates data with its applications. Another example is Bloomberg Marketplace[9] that brings data from a variety of sources and providers, curates and makes them available to its customers who pay per-transaction and/or subscription fees. According to Ordanini e Pol (2001), digital marketplaces in general provide technological infrastructure, rules, business models and services for transactions between providers and consumers.

3.2.9 Data Ecosystems Value Generation

In Data Ecosystem, actors are required to employ a set of capabilities and resources to catalyze value. According to S17, often the onus is on the consumers to extract value from the available resources. This creates a problem since the average consumer lacks the necessary skills (ZUIDERWIJK et al., 2012). Due to these barriers, value is not created by a single actor but rather a value chain (*i.e.*, in a network of actors). A value chain is a set of independent value-adding activities that is used to exploit a set of resources. Moreover, a value chain consists of different actors conducting one or more activities (*e.g.*, data provision, data curation, data analysis), and each activity can consist of a number of actions or value creation techniques (*e.g.*, Gathering, Visualization, Service Creation). In Data Ecosystems, the minimal value chain consists of data producers, data intermediaries, and data consumers [S25]. As value-adding activities offer different complexity, it is possible that each action can consist of one or more value chains [S20].

[8] http://datamarket.azure.com/
[9] https://www.bloomberg.com/professional/product/market-data/

The introduction of incentives and rewards can stimulate the flow of resources in a Data Ecosystem [S04][S21]. In fact, the production, provision and exploitation of Data Ecosystem resources need investments [S14][S21]. S14 states that without money, it becomes very difficult to sustain Data Ecosystem initiatives. However, there is little incentive to invest in resources and capabilities [S22][S10]. The lack of knowledge on the benefits of sharing data and the lack of new operation models are the main impediments that explain why actors, mainly private companies, are not currently motivated to engage on Data Ecosystems [S22]. Therefore, it is important to develop sustainable business models that give them an incentive to keep data up-to-date and accessible and, in addition, to create sustained commercial applications and tools [S21][S22].

Business models support the value proposition for actors in an ecosystem. A number of business models applicable to data were described in the literature (TEECE, 2010; ZOTT; AMIT, 2010), such as support, subscription and professional services/consulting business models, which could be applicable both for data producers and application developers [S22].

Figuring out how to earn revenues (*i.e.*, earn financial profit) from providing data or data-based resources to consumers is a key element of business model design in Data Ecosystems. Services and applications can be priced commonly based on features, cost or pay-per-use go models, for instance [S22]. Data resources can be priced using the Subscription model, in which the consumer pays a fixed price for a certain time frame. Another suitable model is the Flickr multiple revenue stream model (TEECE, 2010) as it involves collecting subscription fees, charging advertisers for contextual advertising, and receiving sponsorship and revenue-sharing fees from partnerships. However, in general, data is often complex to price, and consumers have many ways to obtain certain types of data without paying (TEECE, 2010). Another alternative is to earn revenues by attempting to extract meaningful and actionable information from it in order to support other business, such as targeted advertising and improvement of services and products.

3.2.10 Data Ecosystem Management

Although Data Ecosystems have the potential to generate benefits, many ecosystem initiatives are struggling to establish effective management of their resources and actors (POLLOCK, 2011; UBALDI, 2013). In fact, while the potential of Data Ecosystems is real, the realization is unsuccessful in many cases [S25][S02] (ZUIDERWIJK et al., 2012). For instance, there are fundamental concerns related to Data Ecosystem development despite high expectations and financial investment. According to S10, as a consequence of numerous barriers and limitations, the performance of Data Ecosystem initiatives tends to be simplistic. They focus on releasing and promoting short-time contests, such as hackathons and code fests. Even the applications developed for government-sponsored application contests present unsatisfying outcomes. Most applications in such scenarios end up being

quickly abandoned. In a long-standing perspective, very few applications resulting from such contests are actually scalable and sustainable [S29].

Establishment of the right Data Ecosystem means the proper coordination of various categories of actors and the provision of business support and stimulation of resources development and usage [S13]. Other essential elements in a successful Data Ecosystem are actor collaboration, integration of scientific, social, and economic information, preservation of ecological processes, and adaptive management [S16]. In fact, Data Ecosystem management is important in order to actively facilitate the effective functioning and accomplishing the goals of ecosystems [S02]. Moreover, if a Data Ecosystem does not have a management structure, it becomes difficult to drive the ecosystem forward, and to build and learn from past experiences [S18].

Data Ecosystems will perform well only if they are designed with an appreciation of their full complexity. According to S02, the management of a Data Ecosystem requires sketching some basic topics, that focus on (i) identifying the most active actors that act as essential components of the ecosystem; (ii) analyzing the nature of the transactions that take place between those actors; (iii) recognizing what resources are needed by each actor and how they engage transactions; and (iv) studying the indicators that signal the status of ecosystem activity [S02]. Therefore, these considerations demand a more systemic approach to program planning and designing Data Ecosystems.

However, so far, Data Ecosystems management initiatives are simplistic. Governments around the world have developed programs and policies that aim to promote both the supply and use of public sector data [S02][S13]. Such policies very often focus on ensuring the availability and quality of data resources. These management policies have helped to expand Data Ecosystems and improve the provision of data. However, such policies fail to include other key actors such as data consumers and data intermediaries, who actually demand the supply. Hence, it is crucial to include, from the beginning, the point of view of all the ecosystem's actors. An integrated and collaborative management must ensure that goals included in a Data Ecosystem agenda would meet the needs, rights and interests of all actors who are part of the ecosystem [S12].

With a different management approach, S14 recommends the intervention of government to trigger or guide Data Ecosystems development and to link these with the government's sector objectives. Among a list of recommendations, S14 advocates governmental investment in a core data infrastructure that will become the foundation for advanced Data Ecosystems. Moreover, governments need to leverage social forces and integrate them into technological arrangements when implementing the ecosystem as a long-term, advanced development strategy [S14]. Similarly, S03 also argues that the role of government should be to provide a simple, reliable and publicly accessible infrastructure that exposes underlying data, with any use of such data left entirely to other actors, either nonprofit or commercial ones. However, S03 states that governments should provide an

infrastructure and then get out of the way in order to promote self-organization for the ecosystem.

Assessment and evaluation methods are useful tools to improve the effectiveness of Data Ecosystems. The literature refers to the concept of health of an ecosystem as a means to monitor and assess ecosystem functioning, identify and predict areas for improvement, and evaluate changes in the ecosystem (MANIKAS; HANSEN, 2013a). Until now, alternatives to measure Data Ecosystem health are still naive, as the focus on relatively simplistic metrics such as number of datasets published, number and percent of existing datasets downloaded, number of datasets scheduled for release, number of APIs, and basic site analytics (*e.g.*, number of page views, downloads, etc.) [S10]. While informative to some extent, these health indicators focus only on the data producer perspective and as consequence they not evaluate how useful are the resources provided.

3.2.11 Data Ecosystems Privacy

Despite positive functioning and beneficial impacts expected from Data Ecosystems, governments, private companies and individuals are becoming aware of these risks on their privacy, and are starting claiming greater protection on their data [S04,S05,S14]. Each day, a growing amount of personal data that is collected, analyzed, and used by private and public entities. This data include relationships and personal information of users from shopping choices to health-care information. According to S14, such data are already commercialized.

Unfortunately, not all personal data custodians offer full guarantee to their users on the control on data privacy. This is a well-known issue for social network platforms. Although all of them include privacy control features, mainly to avoid criticism, they neither promote them, nor make them easy to use [S04]. For instance, Korea's largest theft of personal data, which were stolen from KB Kookmin Bank, Lotte Card, and NH Nonghyup Card, involved over 40

Consequently, in realization of the seriousness of illicit acts and the misuse of data, governments and international organizations are creating a number if regulations, including tougher penalties for cyber libel. For instance, the Korea Communications Commission released data privacy guidelines which allow industries to collect and use the private information and history of individuals without their consent [S14]. Another example is being carried out by the European Commission that are requesting greater levels of transparency, correctness and control features. The priority is, therefore, on the protection of data privacy, to reduce the risks due to their uncontrolled use, instead of promoting their exploitation, under the control of individuals [S04].

Because a large amount of data is private (*e.g.*, data about individual persons), the management of data privacy is important [S22]. Data privacy is an element that should be carefully considered when handling any data. A myriad of practices should be in place

to ensure that data protection laws are adhered to. Moreover, there is a growing consent to avoid that data cannot be linked back to an individual. The only exception to this is information about individuals that is legally in the public domain for transparency purposes. In all cases of data privacy is a fundamental right that must be protected and the data publisher must comply with data protection laws.

Some mitigation measures can be employed in order to protect data privacy. For example, providing clear privacy guidelines for publishing data, liaise with the national statistics authority on best practice for data anonymisation/aggregation and with the data protection office for guidance on legislation, create an awareness campaign to clarify differences between different kinds of data publication. Before deciding what data to publish, data producers need to have an overview of what data they currently manage, and could therefore potentially be published. This can be a challenging task, as data in large organisations is typically dispersed over multiple websites, databases, shared-storage, and personal-computers. According to S18, a data audit serves to establish an inventory of what private data currently exist. For each data to be published, the following information should be collected: privacy information (does the data contain personal/sensitive information), legal information (what license/terms and conditions are currently associated with the data), operational information (how often the data is collected), technical information (what format the data is in), and other valuable information.

3.2.12 Data Ecosystems Theoretical Foundations

We looked for references to established theories in order to help either to define the term Data Ecosystem or to develop the background for the research. Similarly to Hanssen e Dybå (2012), we also looked at either explicit well-established theories, such as Socio-Technical theory, or wider concepts, such as value creation and co-innovation.

Table 11 lists the most common theories used by the authors of the studies selected. Social-Technical Systems Theory (CLEGG, 2000) provides a base for 6 of the studies. This theory refers to the systematic integration of social and technical aspects of an organization or society as a whole (FISCHER; HERRMANN, 2011). The basic idea is that the interaction between social and technical factors influences the outcomes (FISCHER; HERRMANN, 2011) of a process, a system or an organization. Value Chain Theory (LEE; YANG, 2000) is the second most common theory, identified in 4 studies. The theoretical studies about other categories of ecosystems were used to assist in the conceptualization of the field. In particular, Business Ecosystems studies were referenced by 4 of the studies selected. (MOORE, 1999) defines a Business Ecosystem as "an economic community supported by a foundation of interacting organizations and individuals", which includes customers, producers, competitors, and other stakeholders. The key elements to a Business Ecosystem are the keystone species, which are companies that display leadership and exert a strong influence on co-evolutionary processes (MOORE, 1999).

Table 11 – Theoretical foundations adopted by Data Ecosystems studies

Theory	**Studies**
Resource Dependency Theory(PFEFFER; SALANCIK, 2003)	S03, S24
Social-Technical Systems Theory(CLEGG, 2000)	S05, S10, S14, S15, S18, S27
Normalization Process Theory(SOOKLAL; PAPADOPOULOS; OJIAKO, 2011)	S05
Information Polity(HELBIG et al., 2012)	S15
Actor Network Theory(MUNRO, 2009)	S15
Value Chain Theory(LEE; YANG, 2000)	S19, S20, S22, S28
Dynamic Capability Theory(DANIEL; WILSON, 2003)	S03
Digital Innovation(YOO; HENFRIDSSON; LYYTINEN, 2010)	S01
Natural Ecosystems(WIKIPEDIA, 2001)	S25
Business Ecosystems(MOORE, 1999)	S18, S22, S25, S29
Software Ecosystems(JANSEN; BRINKKEMPER; FINKELSTEIN, 2009a)	S16, S29
Digital Ecosystems(NACHIRA; DINI; NICOLAI, 2007)	S25

Besides the term Business Ecosystem, it is important to note that the term Data Ecosystem spans ideas that are borrowed from other approaches and for these the terminology varies. For instance, Data Ecosystems are inspired by the notion of Biological Ecosystems, which, in particular, denote a natural unit consisting of all plants, animals, and microorganisms in an area functioning together with all of the non-living physical resources of the environment (WIKIPEDIA, 2001). Moreover, a Data Ecosystem among other elements includes systems, databases, workflows, people, the market, government and an infrastructure. These elements suggest that a Data Ecosystem needs to combine components from different ecosystems.

The creation of a Data Ecosystem is also prompted by there being a Digital Ecosystem and a Software Ecosystem. According to Nachira, Dini e Nicolai (2007), Digital Business Ecosystems "provide an open source distributed environment, where software components, services, applications and also business models are regarded as digital species that can interact with each other, reproduce and evolve according to laws of market selection". The term Software Ecosystems is also recent. This refers to networked organizations or individuals, who base their relations on developing, commercializing and using a central software technology (HANSSEN, 2012; JANSEN; CUSUMANO; BRINKKEMPER, 2013). All of these ecosystems break down the internal boundaries of organizations' production lines by allowing contributions from partners, vendors, and other external parties.

A Data Ecosystem can be viewed as another instance of a Business Ecosystem, a Digital Business Ecosystem or a Software Ecosystem. Despite sharing network and co-evolution characteristics, Data Ecosystems differ from previous ecosystems. Unlike other ecosystems, Data Ecosystems do not rely on an explicit common platform in which different actors can collaborate. The common platform is actually the wide collections of data

exchanged by the actors. In particular, the data do not necessarily need to be provided by a single actor. The lack of a common platform creates a more diffused supply-demand network. Another difference is related to how products traded between the actors are perceived. In Business Ecosystems, business operations and actors *per se* are the products (MANIKAS; HANSEN, 2013a). In Software Ecosystems, the products are software components or services. In Data Ecosystems, the product is data.

Actually, the boundaries between different kind of ecosystems are difficult to define [S16]. For instance, Data Ecosystems may entail Software Ecosystems in relation to the network of actors involved in developing and providing data-related software. Hence, understanding these related ecosystems allow us to use some concepts in the Data Ecosystem world. It can provide us with solutions to the challenge of building up information and knowledge about a Data Ecosystem.

In summary, a heterogeneous theoretical background has been used by the Data Ecosystem research studies. This situation is a consequence of two factors: the field is in its infancy and different research and industry communities have been investigating the area independently.

3.2.13 Data Ecosystems Benefits and Barriers

Table 12 shows the benefits expected in establishing/developing a Data Ecosystem according to the selected studies. The most cited benefits are related to enabling or improving aspects of political and social life, such as improvements in the quality of life and social trust, economic growth, the support of policy making processes and enhancing citizen services [S01] [S10]. The second most cited benefits are related to economic aspects, such as creating new business opportunities by using data and data services, and enabling innovation and value creation [S06].

Table 12 – Benefits expected from Data Ecosystems

Benefits	**Studies**
Improvements in political and social aspects	S01, S02, S07, S08, S09, S10, S11, S12, S13, S15, S16, S17, S18, S23, S24, S26, S27, S28, S29
Improvements in economic aspects	S02, S03, S06, S09, S10, S11, S13, S12, S14, S17, S18, S19, S20, S22, S23, S25, S26, S27, S29
Ease of data consumption and production	S02, S04, S05, S06, S08, S10, S11, S12, S22, S24, S25, S27
Communication and interaction between actors	S01, S02, S03, S13, S17, S18, S20, S21, S24, S27
Improvements in the quality of data and services.	S01, S03, S05, S08, S09, S11, S12, S15, S22, S24, S29

Another benefit expected is the ease of consumption and production of data by using Data Ecosystems. Actors that do not have the abilities and resources to consume or

provide data can contract Service Providers or Data Intermediaries. For instance, S22 presents several specialized roles related to data provision. Another example is S21 S21 which presents the roles of Data Analyzer and Data Visualizer to help data users in the data consumption process. S12 and S15 emphasize that Data Ecosystems also help to promote the interoperability of data and services, thereby aiding the reuse of data and data transparency.

Another great benefit expected is to prompt actors to interact and participate. According to S13, when actors communicate with each other and interact. This contributes to establishing a Data Ecosystem. A well-maintained network with internal and external actors working collectively increases competitive advantages [S24]. As stated in S22, the communication between actors facilitates the delivery of services and sharing of knowledge, and also enables several types of partnership and cooperation. Actor engagement and interaction can also be improved by holding events such as hackathons, seminars, conferences and competitions [S13].

And lastly, Data Ecosystems also contribute to the delivery of better data and services (due to feedback from the actors. According to [S29], data producers and data intermediaries benefit from ideas and feedback about their own processes received as a result of transactions with other actors. Based on relevant feedback, it is possible to improve the ecosystem as a whole by analyzing which applications and services should be continued, revised, or abandoned in favor of alternatives [S29]. In particular, attention must be paid to the quality of the data. For most actors, the data consumption problem is less a matter of data volume than the quality of data [S05]. Actors can give feedback to data producers and thus improve the correctness and quality of the data [S22]. In Data Ecosystems, it is common for there to be rules developed that seek to increase the quality of data [S09].

Table 13 shows barriers and challenges expected that Data Ecosystems identified in the selected studies will face. The most cited barriers and challenges are the lack of technical knowledge and resources to maintain an ecosystem. In S11, the authors express their concern about the long-term sustainablity of a Data Ecosystem. According to them, the actors need to generate revenue to cover part of their operating costs. The cost of providing data and the resources needed to make them useful have received relatively little attention [S19]. In particular, the cost of simply keeping the data is usually only a small fraction of the total cost of data management [S19]. From the data consumption perspective, the cost of consuming data is largely related to supporting finding, accessing, and reusing data [S19]. Data Ecosystem actors need to find ways to keep their activities financially and politically viable [S02].

Other barriers pertain to the complexity of activities needed to produce, identify, access, understand, and use data. Even when data has already been created, there are a number of aspects to consider in order to meet the data provision criteria. For instance, the creation of appropriate and sufficient metadata to assist data consumption [S07][S29].

Table 13 – Barriers and challenges for Data Ecosystems

Barriers and Challenges	Studies
Lack of technical knowledge and resources to maintain the ecosystem	S01, S02, S03, S04, S05, S07, S08, S09, S10, S11, S13, S15, S16, S17, S18, S19, S20, S21, S22, S23, S26, S27, S29
High complexity of tasks such as data discovery and data consumption	S01, S03, S07, S08, S09, S10, S12, S13, S16, S18, S20, S21, S23, S24, S25, S26, S27, S29
Lack of actor participation and interaction	S01, S09, S10, S12, S11, S13, S14, S15, S23, S24, S25, S26, S29
Lack of organizational structure	S01, S05, S06, S07, S10, S13, S14, S16, S18, S21, S22, S25, S26
Presence of aspects related to privacy, confidentiality and liability	S04, S06, S10, S14, S18, S26, S28

Furthermore, in most cases, data producers do not know what data other actors want and will use. The lack of feedback to effectively engage the actors makes it difficult to know what kind of data is valuable for release. Moreover, the way that data are published also influences possible barriers for using data (ZUIDERWIJK; JANSSEN, 2014). For instance, poor data quality [S10][S03], operational changes in the provision of data [S18] and usability problems [S08] are common barriers related to data consumption activities. By using the data consumption perspective, actors should possess several capabilities to find data, collect data, clean and prepare data for consumption, and, finally, consume the data [S16]. The lack of guidelines [S16] and the capabilities required [S03][S23] make it difficult for many actors to use data and to generate value from it.

The lack of participation and interaction between actors affects the ecosystem, causing actors to participate less [S13], and reduces the sharing of data, knowledge, and experiences [S01]. Actor participation and interaction barriers refer to the ease and attractiveness of joining and contributing to a Data Ecosystem [S01]. Underlying factors raising barriers in this area include lack of incentives, lack of ability or time to engage in an ecosystem, costs and competition from other ecosystems (ZUIDERWIJK; JANSSEN, 2014). According to S13, it is necessary to provide detailed business, economic and financial models for the Data Ecosystem to stimulate actors to participate. Furthermore, in order to obtain the best outcomes, it is crucial to raise awareness among the actors, i.e., the actors must understand that value generation calls for a data chain that enriches raw data and thus makes the content valuable.

The lack of organizational structure for Data Ecosystems hinders internal and external participants understanding and, subsequently, implementing and manipulating these ecosystems [S14]. Organizational structure consists of well-defined models for interaction between actors, their roles, specific theories that describe internal and external structures

in Data Ecosystems [S05], operation models, policies and principles [S06]. For instance, in S25 the authors state that the lack of business models that generate value for all the actors causes a one-sided dependency. The same problem is reported by S22.

Additional barriers are derived from concerns for privacy, confidentiality and liability [S10]. Besides operational activities and technical resources, the requirements for data provision also include privacy and confidentiality protection [S10] [S02] [S28]. Private data are often the data user's own data or information collected about users. In the Web, every day, individuals create personal data that is collected, analyzed, and used by private and public entities. The violation of privacy by opening data and the prospect of being legally liable when data are misused is an important barrier (ZUIDERWIJK; JANSSEN, 2014). Existing privacy concerns include the invasion of privacy, imperfections in security, the seriousness of illicit acts and the misuse of personal data [S14]. In S06 and S04, the authors call for technical standards and guidelines for privacy and security management to be put forward.

3.3 CONSIDERATIONS OF THE CHAPTER

The purpose of this Chapter was to find current relevant research on Data Ecosystems as well as to provide an overview of the field. Apart from providing an overview, we also identified several areas that are not covered in the literature body. Up until now, there are not many academic papers related to Data Ecosystems. In most cases, they are focused on some component technology or a solution that reflect only a small fragment of the whole research area. Three important factors should be considered in order to elaborate the development and evolution of Data Ecosystem field: theory, models, and engineering.

There is no well-established terminology in this area. So far, the terminology and definition for Data Ecosystem vary greatly. This diversity imposes a pressing problem for developing a clear understanding of the new opportunities and emergent capability challenges in exploiting data ecosystems. Accurate definitions are required in order to get a mutual understanding of what data ecosystems embody.

Moreover, the whole functioning of Data Ecosystem is still ambiguous in practice. It is not clear how to design, maintain or evolve Data Ecosystems. Moreover, we have not identified a model for describing the whole ecosystem and its essential elements as well as how to control and manage an ecosystem. A model will support the practitioners and researchers in having a proper idea about the current state of a data ecosystem. It also may help to define a strategic planning for achieving its goals, such as value creation and new businesses. Hence, a model tends to provide the means for developing a framework to control and manage an ecosystem.

Assessment model is another gap in the Data Ecosystem Literature. Like natural ecosystems that need to address resilience and sustainability properties, data ecosystems need the ability to survive and evolve. The health of an ecosystem depends in part on the

variety of ideas, actors, relationships, and technologies available. Hence, try to understand the ecosystem health and predict the feature conditions of the ecosystem by assessment models it is crucial for the field. Measuring the health of an ecosystem would provide large benefits. According to [S26], part of defining effectiveness and success for a Data Ecosystem lies in determining and utilizing the metrics that measure its health. However, very few studies elaborate, analyze or measure the health of a software ecosystem. For instance, we identified only one work addressing this theme (*i.e.*, [S11]).

Finally, a field that has not been covered in the literature, is the engineering of Data Ecosystems. Data ecosystem is also a socio-technical phenomenon. Hence, its health is very dependent on social matters like managing the actors and communication. It is essential having models for orchestrating social aspects of data ecosystem. Communication and interaction are two important benefits of data ecosystems. The management and orchestration of actors and their relationships require more research in academia. In particular, there is a lack of engineering and governance methods in order to organize a data ecosystem, to assist in the decision making and to materialize the benefits.

Another important gap is related to the metadata management. We had not found any study proposing or even mentioning how the metadata should be preserved or improved. This is still an immature area, thus there is a lack of detailed information on "How metadata curation in Data Ecosystems can be structured and systematized by means of practices and processes" and some research is needed in order to propose new solutions to this open problem.

Next Chapter presents the formalization of Data Ecosystems terminology. It also presents a meta-model proposal for describing essential elements related to Data Ecosystems.

4 A META-MODEL FOR DATA ECOSYSTEMS

Data Ecosystem covers a wide range of disciplines including design, orchestration, management and assessment. For each discipline, models would help in understanding the functioning and activities of Data Ecosystems. The term model is generally used to denote an abstract description of a study object (or part of it) (*e.g.*, concepts from the real world, behaviors or systems) related to a specific point of view (ATKINSON; KUHNE, 2003). In this sense, a model should be able to answer questions in place of the actual study object. Models must be precise enough to be subject to automation, and therefore it is important for them to be written in a well-defined language.

In general, models allow sharing a common vision and knowledge among technical and non-technical stakeholders, facilitating and promoting the communication among them. According to Silva (2015), models also allow more effective and efficient planning to be undertaken while providing a more suitable view of the system to be developed. They also allow system control to be achieved according to objective criteria.

In Data Ecosystems, models can be used as a kind of blueprint that can be used for running and managing Data Ecosystems. They also may help to define strategic plans for achieving ecosystem's goals, such as value creation and new businesses. Hence, a model tends to provide the means for developing a framework to control and manage an ecosystem (BOUCHARAS; JANSEN; BRINKKEMPER, 2009; PETTERSSON et al., 2010).

For Software Ecosystems, some studies had proposed models for describing the ecosystem and the environment in which software products and services operate. Boucharas, Jansen e Brinkkemper (2009) present and formalize a standards-setting approach to software product and software supply network modeling. Such approach enables software vendors to communicate about relationships in the software supply network, theorize about weak spots/links in their business model and anticipate upcoming changes in the software ecosystem. This Software Ecosystem modelling proposal was extended and analyzed in several other studies, such as (JANSEN; BRINKKEMPER; FINKELSTEIN, 2009b; JANSEN et al., 2012; GERMAN; ADAMS; HASSAN, 2013). Similarly, (PETTERSSON et al., 2010) propose a conceptual model for describing Software Ecosystems. However, (PETTERSSON et al., 2010) focus on Software Ecosystems for mobile learning, which bring up several aspects and insights for this particular domain. Despite these models for Software Ecosystems provide a holistic picture of the relationships between a vendor and its buyers and suppliers, they are based on value exchange related to a core platform.

Moreover, to the best of our knowledge, a model for describing a Data Ecosystem and its essential elements has not been proposed yet. In particular, we have found no evidence of a model that: (i) defines the constructs and the construction rules required to describe the core constructs of Data Ecosystems; (ii) can be used as a reference for

studies that aim to specialize (or reuse part of) for specific Data Ecosystem domains (*e.g.*, Open Data Ecosystem or bio-science data communities); and (iii) allows interchange or transformation of models between Data Ecosystem tools.

The above requirements suggest a kind of model to represent different Data Ecosystems as well as being extended to cover the particularities of these ecosystems. A suitable means for creating such kind of models is to develop a meta-model for describing the essential aspects of Data Ecosystems.

According to Seidewitz (2003), a meta-model is a specification model for which the objects under study are specified as models using a certain modeling language. Other authors define meta-model as an abstract model which represents other "instances" of models (BRAMBILLA; CABOT; WIMMER, 2012). In this sense, meta-models take place one or two levels of abstraction higher than the standard concrete models (BRAMBILLA; CABOT; WIMMER, 2012). Meta-modeling techniques play an important role in supporting the design and the development of complex systems. Meta-models can be used proficiently for defining new languages for exchanging and storing information as well as defining new properties or features to be associated with existing information (metadata) (OMG, 2016; BRAMBILLA; CABOT; WIMMER, 2012).

With meta-model constructs, it is possible to fully customize models to a certain domain, or class of stakeholders, or add new types, for example, (BRAMBILLA; CABOT; WIMMER, 2012). In this sense, the meta-modeling process can be defined as a modeling hierarchy with each level being characterized as an instance of the level above (BRAMBILLA; CABOT; WIMMER, 2012). For instance, going from an instance level (M0), consisting of objects or individuals that conform to the model level (M1) and further to a meta-model (M2) and meta-meta-model (M3) levels, which define a language for describing meta-models. An example is UML Modelling architecture (OMG, 2016; BRAMBILLA; CABOT; WIMMER, 2012). In the M3 level, MOF (Meta-Object Facility) provides a meta-meta-model to build metamodels. In the M2 level, UML metamodel provides a model that describes the UML itself. The M1 Level contains models written in UML, such a system's class diagram. The M0 level contains the actual data that is used to describe real-world objects.

In our context, (i) the meta-model level (M2) defines which modeling constructs will be used for creating a model regarding an abstract Data Ecosystem; (ii) the model level (M1) is an instance of a meta-model that specifies a representation of constructs for a specific Data Ecosystem domain (*e.g.*, smart cities or research data sharing community) and (iii) an instance level (M0) is composed of data or metadata items that conform to a specific application model.

In this Chapter, we aim to fill these gaps by presenting a formal definition for Data Ecosystems as well as presenting a meta-model for describing Data Ecosystems. Both contributions were published in (OLIVEIRA et al., 2018; OLIVEIRA; LÓSCIO, 2018). First,

to enable the development of a common knowledge base, we investigate the state of research on the Data Ecosystem field and extract constructs from relevant studies in order to build a common and cohesive definition for Data Ecosystems. Based on this formal definition, we propose a meta-model. Such meta-model defines the Data Ecosystem fundamental constructs and their inter-relationships for enabling analysis and description of ecosystems. In summary, we aim to contribute to the advance of the Data Ecosystem research by both identifying the core constructs used for defining Data Ecosystem and associating them in a meta-model. We expect that the results presented in our study will stimulate and provide support for a debate in the scientific community to the Data Ecosystem theory.

In the next sections of this Chapter are presented the research method, meta-model knowledge acquisition process, meta-model conceptualization, meta-model formalization, evaluation method and evaluation results.

4.1 META-MODEL DEVELOPMENT METHOD

Although the use of meta-models is very common in several domains, the construction of a meta-model is not a trivial task. A central activity in the development of a meta-model is to identify those constructs and aspects of a knowledge domain that are of interest to denote them and create terms to refer to them (ATKINSON; KUHNE, 2003). Hence, meta-models have to follow a defined structure as well as have to represent a knowledge domain. According to Cristani e Cuel (2005), the main idea is to develop an understandable, complete, and shareable set of classes, properties and relations that represent a domain. The use of a well-defined strategy that specifies explicitly or implicitly a series of steps should help modelers to produce a valid, well-formed and high-quality model (CRISTANI; CUEL, 2005).

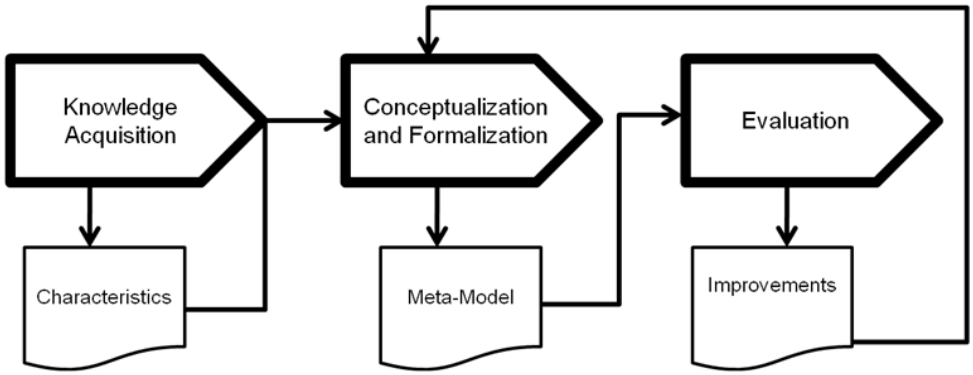

Figure 12 – Research design: construct-evaluate cycle. Adapted from:(SANTANA; ALVES, 2016)

In order to develop a meta-model for Data Ecosystems, we employ an interactive construct-evaluate cycle (*cf.* Figure 12) based on multiple qualitative methods, including

systematic literature review, qualitative data analysis, and qualitative evaluation. The whole research design of this work was based on the research method employed by Santana e Alves (2016), Santana (2015), which followed the interactive process composed of three phases: knowledge acquisition, conceptualization and formalization, and evaluation. The construct-evaluate cycle approach is suitable for building meta-models from scratch and also is aligned with Design Cycle and Rigor Cycle proposed by Hevner (2007).

4.2 RUNNING EXAMPLE

Like Zuiderwijk, Janssen e Davis (2014), we developed a motivational scenario to exemplify the constructs and concepts presented in following sections. According to Carroll (2000), scenarios allow creating knowledge to discuss problems, as well as can demonstrate an interpretation by the appearance and behavior of the system.

In this sense, an example of a Data Ecosystem is an open government data initiative promoted by fictional Government Administration, which aims to promote transparency and improve their services.

In this scenario, data must comply with the eight open government data principles [1]. Data shall be complete, primary, timely, accessible, machine processable, non-discriminatory, non-proprietary, and license-free. Creative Commons Attribution Share-Alike license constrains all of data access, use and re-use.

Such Data Ecosystem is centered on the Open Data Portal used to publish datasets and related documentation (*e.g.*, documentation, data dictionary and metadata). The Open Data Portal provide a seamless user experience for visitors by addressing usability and accessibility properties. Among its features, the portal provides mobile and cross-browser compatibility, and is integrated with social media platforms.

The portal was developed using CKAN, which is an open source software that provides support for the creation of data portals. Moreover, this portal is hosted at OpenShift[2], which is a company that provides a Web site hosting services. Such web hosting service is provided s a Freemium Model, which is a supply-demand business model by which a service is provided free of charge, but additional features for such service have an associated charge.

Two datasets provided by the Portal, which are: (i) Budget Expenses and (ii) Budget Financial Statements. These datasets are structured using a tabular data model and are distributed by CSV[3] files. These datasets conforms to the INDA (Normative Instruction of the National Open Data Infrastructure)[4], which is a Brazilian standard.

[1] https://opengovdata.org/
[2] https://www.openshift.com/
[3] https://tools.ietf.org/html/rfc4180
[4] <https://www.governodigital.gov.br/transformacao/cidadania/dados-abertos/inda-infraestrutura-nacional-de-dados-abertos>

Furthermore, this Data Ecosystem is made up of many actors and small organizational structures, which mostly belonged to one of following sectors: administration technicians, application developers, Small and Medium Enterprisess (SMEs) and citizens.

Alceu is a government technician responsible for publishing the Budget Expenses and Budget Financial Statements datasets. Developers develop applications, apps and games, or create visualizations and publish analysis, which are used by other actors. A popular application is a site My Government Budget (MGB) developed by a startup called OpenAccount. The MGB presents the government budget at a high level by integrating all Government's datasets. SMEs also perform an intermediary role. They usually supply infrastructure and services to government and other actors. Elba is financial journalist who plans to utilize the Budget Expenses and Budget Financial Statements datasets. She also use MGB to analyze the Government finantial status.

4.3 META-MODEL KNOWLEDGE ACQUISITION

The *Meta-Model Knowledge Acquisition* phase aimed to identify the relevant concepts, characteristics and aspects that describe Data Ecosystems. We used as input the papers identified in our recent systematic literature review on Data Ecosystems (OLIVEIRA; LIMA; LÓSCIO, 2019) and presented in Chapter 3. The selected studies (*cf.* Table 6) were used to identify the constructs of a Data Ecosystem. In particular, these studies have a wide variability on the level of detail they provide on Data Ecosystems ranging from mere reference to the ecosystem term (*e.g.*, Shin (2016), Shin e Choi (2015)) to papers in which Data Ecosystems conceptualization is the main focus (*e.g.*, Zuiderwijk, Janssen e Davis (2014), Heimstädt, Saunderson e Heath (2014), Harrison, Pardo e Cook (2012)).

Our investigation of the literature provided various characteristics related to Data Ecosystem, as can be seen at Table 14. These characteristics were used to extract the essential constructs related to Data Ecosystems. Most of the characteristics are informal and with a low degree of maturity of their understanding. Another problem is the lack of knowledge about their relationship. Analyzing the connection between the constructs is difficult. On the other hand, the lack of conceptual knowledge about the Data Ecosystem and the absence of a model for its description appear as important barriers to Data Ecosystems initiatives. These conclusions reinforced the idea that the meta-model proposed in this work can be a relevant contribution to this area.

[5] *cf.* Table 6

Table 14 – Data Ecosystem Characteristics

Data Ecosystem Characteristics	**Works ID**[5]
Data Ecosystems enable sharing of data, services, experience and knowledge	S01, S02, S05
Data Ecosystems enable communication between actors	S29
Data Ecosystems enable value creation	S07, S09, S13, S23, S24, S25, S26, S28, S29
Data Ecosystems are composed by multiple and different actors	S01-S29(All Selected Papers)
Data Ecosystems are composed by multiple networks of actors	S01, S05, S16, S14, S13, S18, S26
Data Ecosystems rely directly or indirectly on infrastructure services and resources	S01, S05, S16, S22, S26, S08, S29
There are large ecology of actors, such as individuals, markets or (private and public) organizations	S01, S06, S07, S05, S07, S09, S13, S14
An actor produces, acts on or is responsible for one or more Data Ecosystem resource, such as data, software or infrastructure	S01-S29(All Selected Papers)
An actor has dynamic behavior	S03, S05, S09, S15
An actor is autonomous	S16, S29
An actor has goals and requirements	S02, S03, S09, S26
An actor has different capabilities, experience and knowledge	S02, S07, S03, S23, S24, S28, S29
An actor may be composed by other actors	S07, S22
An actor performs one or more role	S01, S02, S04, S06, S08, S12, S16, S19, S20, S22, S25, S26, S29
A role has duties and is responsible for a set of activities	S10, S26
An actor interacts with others through multiple and varying relationships	S16, S14, S13, S16, S22, S01, S05, S13, S25, S26
A relationship relies on the use of services or other ICT technologies	S08, S22, S10, S16, S25
A relationship follows a dependency relation	S02, S04, S07, S14, S19, S26
A relationship enables value generation	S07, S09, S13, S23, S24, S25, S26, S28, S29
A relationship may be influenced by business models	S01, S06, S13, S19, S22
A relationship involves transactions of Data Ecosystem resources, such as data, services, tools or infrastructure;	S01, S02, S09, S16, S21, S22
A transaction has one flow that can carry out a Data Ecosystem resource	S02, S04, S07, S14, S19, S26
A transaction may involve costs, such as time and resources	S01, S02, S22
A Data Ecosystem resource has different features and quality attributes	S13, S07, S17, S25, S20, S21, S29, S26, S22
A Data Ecosystem resource may follow standards and regulatory policies, such as open data principles	S08, S12, S15, S25
A Data Ecosystem resource may be constrained by a license	S08, S12, S15, S25
Actors, roles, relationships, transactions and resources are subjected by regulatory conditions, cultural, political and economic aspects or current information and communications technologies	S01, S05

4.4 META-MODEL CONCEPTUALIZATION

We seek to identify the constructs from characteristics presented in Table 14 and connect them to propose a common and cohesive definition for Data Ecosystems. Despite of abstractness of identified characteristics, four main constructs stand out (*cf.* Table 15): (1) actors, (2) roles, (3) relationships and (4) resources. Typically, Data Ecosystems rely on a vast and heterogeneous set of actors, each one with different properties, capabilities and expectations. Similarly, Data Ecosystem resources are heterogeneous. For instance,

datasets are heterogeneous regarding structural (schema), syntactic (format) and semantic (meaning) issues. Actors may produce and consume data using different activities and under different conditions.

Table 15 – Data Ecosystem Main Constructs

Construct	Construct Description	References
Actor	A Data Ecosystem is composed by multiple actors	(MERCADO-LARA; GIL-GARCIA, 2014; ZELETI; OJO, 2016b; ZUIDERWIJK et al., 2016; SCHALKWYK; WILLMERS; MCNAUGHTON, 2016; HARRISON; PARDO; COOK, 2012)
Capability	Actors have different capabilities	(MERCADO-LARA; GIL-GARCIA, 2014; ZELETI; OJO, 2016b; ZUBCOFF et al., 2016; IMMONEN; PALVIAINEN; OVASKA, 2014; ZELETI; OJO, 2016a; ZELETI; OJO, 2014; GAMA; LÓSCIO, 2014)
Expectation	Actors have different goals and requirements	(MERCADO-LARA; GIL-GARCIA, 2014; ZELETI; OJO, 2016b; ZUIDERWIJK et al., 2016; HARRISON; PARDO; COOK, 2012)
Role	A Role is a function played by an actor in a Data Ecosystem	(SMITH; OFE; SANDBERG, 2016; MERCADO-LARA; GIL-GARCIA, 2014; MOISO; MINERVA, 2012; HA; LEE; LEE, 2014; DING et al., 2011; KÖSTER; SUÁREZ, 2016; ZUIDERWIJK; JANSSEN; DAVIS, 2014; BOURNE; LORSCH; GREEN, 2015; ATTARD; ORLANDI; AUER, 2016; IMMONEN; PALVIAINEN; OVASKA, 2014; HEIMSTÄDT; SAUNDERSON; HEATH, 2014; HARRISON; PARDO; COOK, 2012; GAMA; LÓSCIO, 2014)
Duty	A Role has duties and requires some capabilities	(DAWES; VIDIASOVA; PARKHIMOVICH, 2016; LINDMAN; KINNARI; ROSSI, 2016; HARRISON; PARDO; COOK, 2012)
Activity	A Role may be associated to a set of activities related to the production and consumption of data	(DAWES; VIDIASOVA; PARKHIMOVICH, 2016; LINDMAN; KINNARI; ROSSI, 2016; HARRISON; PARDO; COOK, 2012)
Relationship	Actors interact with others through relationships	(ZUIDERWIJK; JANSSEN; DAVIS, 2014; KOZNOV et al., 2016; SHIN; CHOI, 2015; IMMONEN; PALVIAINEN; OVASKA, 2014; SMITH; OFE; SANDBERG, 2016; SHIN, 2016; KOZNOV et al., 2016; HEIMSTÄDT; SAUNDERSON; HEATH, 2014; HARRISON; PARDO; COOK, 2012)
Transaction	Relationships may involve transactions of ecosystem resources	(DING et al., 2011; KÖSTER; SUÁREZ, 2016; GAMA; LÓSCIO, 2014)
Business Model	Relationships may follow a business model	(SMITH; OFE; SANDBERG, 2016; HA; LEE; LEE, 2014; KOZNOV et al., 2016; BOURNE; LORSCH; GREEN, 2015; IMMONEN; PALVIAINEN; OVASKA, 2014)
Resource	Resources are a useful or valuable product or possession produced, provided or consumed by actors	(SMITH; OFE; SANDBERG, 2016; MERCADO-LARA; GIL-GARCIA, 2014; ZUIDERWIJK et al., 2016; ZUIDERWIJK; JANSSEN; DAVIS, 2014; LINDMAN; KINNARI; ROSSI, 2016; IMMONEN; PALVIAINEN; OVASKA, 2014)
Standard	Resources may conform to standards	(IMMONEN; PALVIAINEN; OVASKA, 2014; HARRISON; PARDO; COOK, 2012; SHIN, 2016; GAMA; LÓSCIO, 2014; DAVIES, 2011)
License	Resources may be constrained by licenses	(DING et al., 2011; KÖSTER; SUÁREZ, 2016; SCHALKWYK; WILLMERS; MCNAUGHTON, 2016; HEIMSTÄDT; SAUNDERSON; HEATH, 2014)
Quality Metric	Resources may be evaluated regarding quality metrics	(IMMONEN; PALVIAINEN; OVASKA, 2014; ZUIDERWIJK; JANSSEN; DAVIS, 2014; HARRISON; PARDO; COOK, 2012; SHIN, 2016; GAMA; LÓSCIO, 2014; DONKER; LOENEN, 2017)

Based on these constructs, we define a Data Ecosystem as: *a set of networks composed by autonomous actors that directly or indirectly consume, produce or provide data and other related resources (e.g., software, services and infrastructure). Each actor performs one or more roles and is connected to other actors through relationships, in such a way that actors collaboration and competition promotes Data Ecosystem self-regulation.*

Thus, we define a Data Ecosystem in terms of a quadruple, denoted by $E = (Re, Ro, A, Rl)$, where:

- E is the Data Ecosystem name;
- Re is a set of data-related resources exchanged, produced or consumed by ecosystem actors.
- Ro is a set of roles that can be performed by the ecosystem actors;
- A is a set of actors who participate or participated in the ecosystem;
- Rl is a set of relationships engaged by the ecosystem actors;

Each of these constructs is detailed in the following subsections.

4.4.1 Resource

Resources are a useful or valuable product, possession or capability produced, provided, curated or consumed by actors. In Data Ecosystems, resources range from datasets and data-based software to infrastructure. In particular, data-based software includes reusable assets (*e.g.*, data visualization services and data storage services) or software assets (*e.g.*, apps and tools that ease the analysis and visualization of some specific datasets) that are used to consume, produce or provide data. Resources may be exchanged individually or in combination through relationships' transactions. Furthermore, resources usually conform to standards or are constrained by licenses. Finally, all resources may be evaluated regarding quality metrics.

A Resource may be defined in terms of a quadruple denoted by: $Re_i = (\{Fe_1, ..., Fe_n\}, \{St_1, ..., St_m\}, \{Ql_1, ..., Ql_k\}, \{Li_1, ..., Li_l\})$, where:

- Re_i is the resource name;
- $\{Fe_1, ..., Fe_n\}$ is a set of features. Each feature Fe_i represents a distinctive attribute or aspect that characterizes the resource Re_i;
- $\{St_1, ..., St_m\}$ is a set of standards. Each standard St_i represents a specification, procedure and guideline to which the described resource Re_i is conformed;
- $\{Ql_1, ..., Ql_k\}$ is a set of quality properties. Each quality property Ql_i represents a distinct metric through which a quality property can be evaluated regarding the resource Re_i.

- $\{Li_1, ..., Li_l\}$ is a set of licenses. Each license Li_i represents a license giving official permission to perform some activity related the resource Re_i.

Regarding our motivational scenario, the list of resources are defined as:

- *Budget Expenses dataset = (Fe={resource type=dataset, data structure=tabular, format=CSV}, St ={INDA}, Ql = {complete, primary, timely, accessible, machine processable, non-discriminatory, non-proprietary, and license-free}, Li= {Creative Commons Attribution Share-Alike})*;

- *Budget Financial Statements dataset = (Fe = {resource type=dataset, data structure=tabular, format=CSV}, St = {INDA}, Ql = {complete, primary, timely, accessible, machine processable, non-discriminatory, non-proprietary, and license-free}, Li = {Creative Commons Attribution Share-Alike})*;

- *My Government Budget site = (Fe = {resource type=solution, mobile compatibility, cross-browser compatibility, social media integration}, St = {}, Ql = {usability, accessibility}, Li = {})*;

- *OpenShift web hosting service = (Fe = {resource type=solution}, St = {}, Ql = {availability, reliability}, Li ={proprietary}).*

4.4.2 Role

Role is a function played by an actor in a Data Ecosystem. It is related to a set of duties and activities. Several roles can be identified in a *Data Ecosystem*. Typically, at least data consumers and data producers are identified in contemporary Data Ecosystems. However, there are several additional roles which are responsible for different duties and activities. Moreover, roles may or may not overlap their responsibilities.

A Role may be defined in terms of a triple denoted by $Ro_i = (\{Fe_1, ..., Fe_n\}, \{Du_1, ..., Du_m\}, \{Ac_1, ..., Ac_k\})$, where:

- Ro_i is the role name;

- $\{Fe_1, ..., Fe_n\}$ is a set of features. Each feature Fe_i represents a distinctive attribute or aspect that characterizes the role Ro_i;

- $\{Du_1, ..., Du_m\}$ is a set of duties. Each duty Du_i represents a moral or legal commitment that arise from the role Ro_i;

- $\{Ac_1, ..., Ac_m\}$ is a set of activities. Each activity Ac_i represents a piece of work done as part of role Ro_i's duties.

Regarding our motivational scenario, the set of identified roles are:

- *Producer = (Fe = {}, Du = {produce datasets, publish datasets}, Ac = {{select data, extract data, prepare dataset, publish datasets, maintain datasets})*;
- *Consumer = (Fe = {}, Du = {consume data, provide feedback}, Ac = {access dataset, process dataset, analysis data, provide feedback})*;
- *Intermediary = (Fe = {}, Du = {provide goods or services}, Ac = {develop solutions, provide solutions, maintain solutions, ensure quality metrics}).*

4.4.3 Actor

An Actor is an autonomous entity such as an enterprise, institution or individual, which plays one or more specific roles in a Data Ecosystem. An actor is considered as a basic construct of a Data Ecosystem with identity and that has its distinct existence. Actors bound to a role must possess the capability of discharging the commitments a role imposes on them. A set of interests motivates actors and each one has different expectations. Actors usually commit to the ecosystem, which may have an incentive for being active in the ecosystem.

An Actor may be defined in terms of a quintuple denoted by $A_i = (\{Fe_1, ..., Fe_n\}, \{Ex_1, ..., Ex_m\}, \{Ca_1, ..., Ca_k\}, \{Ro_1, ..., Ro_l\}, \{Re_1, ..., Re_j\}, \{Rl_1, ..., Rl_h\})$, where:

- A_i is the actor name;
- $\{Fe_1, ..., Fe_n\}$ is a set of features. Each feature Fe_i represents a distinctive attribute or aspect that characterizes the actor A_i;
- $\{Ex_1, ..., Ex_m\}$ is a set of expectations. Each expectation Ex_i represents a purpose or objective related to the Data Ecosystem expected by the actor A_i;
- $\{Ca_1, ..., Ca_k\}$ is a set of capabilities. Each capability Ca_i represents a named piece of functionality that is declared as supported by the actor A_i.
- $\{Ro_1, ..., Ro_l\}$ is a set of roles. Each role Ro_i represents a role performed by the actor A_i.
- $\{Re_1, ..., Re_j\}$ is a set of resources. Each resource Re_i represents a resource produced, curated or provided by the actor A_i.
- $\{Rl_1, ..., Rl_h\}$ is a set of relationships. Each relationship Rl_i represents a relationship engaged by the actor A_i with other actor.

Regarding our motivational scenario, the list of actors includes:

- *Government = (Fe = {}, Ex = {promotion of transparency, improvement of government services}, Ca = {funding open data program}, Ro = {producer}, Re = {Budget*

Expenses dataset, Budget Financial Statements dataset, Open Data portal}, Rl = {Government-OpenShift, Government-OpenAccount})

- *Alceu = (Fe = {}, Ex = {}, Ca = {programming skills, knowledge about data domain}, Ro = {producer}, Re = {Budget Expenses dataset, Budget Financial Statements dataset})*
- *OpenAccount = (Fe = {organization type=startup}, Ex = {promote transparency}, Ca = {application development}, Ro = {intermediary}, Re = {My Government Budget}, Rl = {Government-OpenAccount})*
- *OpenShift = (Fe = {organization type=private company}, Ex = {make proffit}, Ca = {Web site hosting}, Ro = {intermediary}, Re = {Web hosting service}, Rl = {Government-OpenShift})*
- *Elba = (Fe = {}, Ex = {analyze government data}, Ca = {finantial data analysis}, Ro = {consumer}, Re = {}, Rl = {OpenAccount-Elba})*

4.4.4 Relationship

Relationships are the interactions among Data Ecosystem actors. Relationships are often based on a common interest or are also related to the role each actor serves in the ecosystem. Also, they may vary according to several aspects, such as economic, political, cultural and technological context. Actors exchange data or other types of resource through transactions. Finally, the relationships may follow business models.

A Relationship may be defined in terms of a quadruple denoted by $Rl_i = (A_i, A_j, \{Fe_1, ..., Fe_n\}, \{Bm_1, ..., Bm_m\}, \{Tr_1, ..., Tr_k\})$, where:

- Rl_i is the relationship name;
- A_i, A_j is a a pair of actors, which represents the actors that participates in the relationship Rl_i;
- $\{Fe_1, ..., Fe_n\}$ is a set of features. Each feature Fe_i represents a distinctive attribute or aspect that characterizes the relationship Rl_i;
- $\{Bm_1, ..., Bm_m\}$ is a set of business models. Each business models Bm_i describes a business strategy of how the relationship Rl_i creates, delivers, and captures value;
- $\{Tr_1, ..., Tr_k\}$ is a set of transactions. Each transaction Tr_i describes a set of exchange or transference of resources between the actors engaged on the relationship Rl_i.

Regarding our motivational scenario, the list of relationships includes:

- *Government-OpenShift = ({Government, Open Shift}, Fe = {type=commercial relationship}, Bm = {Freemium Model}, Tr = {open data portal hosting})*
- *Government-OpenAccount = ({Government, Open Account}, Fe = {type=commercial relationship}, Bm = {}, Tr = {provisioning of datasets related to financial data})*
- *OpenAccount-Elba = ({Open Account, Citizen}, Fe = {}, Bm = {}, Tr = {data-based software provision/consumption})*

4.4.5 Data Ecosystem Properties

In addition to the constructs described in the previous sections, in a Data Ecosystem, there are also two essential properties: networked character and self-organization.

Like Software Ecosystems, Data ecosystems have a networked character. In general, Data Ecosystems are composed of loosely coupled networks of actors in such a way that value is not created in a chain but more in a network of actors (IMMONEN; PALVIAINEN; OVASKA, 2014). Data Ecosystems also have multiple levels and dimensions (ZUIDERWIJK; JANSSEN; DAVIS, 2014). An actor can be composed of several other actors, such as a government. Moreover, an actor can participate in multiple Data Ecosystems. In addition, Data Ecosystems involve both Software Ecosystems, which provide the data-related software, and data-based Business Ecosystem, which is formed by organizations having their own parts and know-how in the data-based business (ZUIDERWIJK; JANSSEN; DAVIS, 2014; HARRISON; PARDO; COOK, 2012). According to Zuiderwijk, Janssen e Davis (2014), in Data Ecosystems, there are pre-existing networks of interaction with different levels of training and capacity.

Data Ecosystems are a self-organizing environment. Relationships emerge through interaction and feedback between the actors (MAGALHAES; ROSEIRA; MANLEY, 2014; POLLOCK, 2011). For instance, while a data user depends on the data that are published by a data producer, the latter depends on the feedback on their data provided by data users. According to Pollock (2011), Data Ecosystems contain data cycles with feedback loops and sharing of data back to publishers. Furthermore, various aspects of the ecosystem affect its health (HEIMSTÄDT; SAUNDERSON; HEATH, 2014). For instance, actors affect and are affected by the creation and delivery of the offerings of the other actors (DAWES; VIDIASOVA; PARKHIMOVICH, 2016). Since there are dependent relations between actors, each can influence participation and relationships between one another. In some cases, keystone actors can provide rules and policies for coordinating the relationships.

Furthermore, Data Ecosystems is subjected by a set of contextual elements, which are physical and non-physical factors that influence the Data Ecosystem's and its elements. Moreover, the contextual elements may contain information that helps actors understand the background and purpose of a Data Ecosystem element. A Data Ecosystem encompasses a variety of actors working in a diverse commercial and cultural environment.

Concerns related to Data Ecosystem can be manifold and might include policy, licenses, technology, financing, organization, culture, and legal frameworks and are influenced by ICT infrastructures. Each of these concerns are contextual elements and they influence directly and indirectly how actors act and participate in the Data Ecosystem.

Resources, roles and relationships are also subjected to contextual elements. For instance, each actor exists in a wider context. The data producers and data consumers are involved in networks of interaction. If a keystone actor lacks of funding to sustain its activities, the Data Ecosystem can have its sustainability compromised. Hence, it is important to take a deeper lock on context in which all Data Ecosystem are subjected in order to better manage and sustain the ecosystem.

4.5 META-MODEL FORMALIZATION

Based on this set of constructs, we formalized a preliminary version of Data Ecosystem meta-model. It was not considered a specific instance of Data Ecosystem. The intent is to represent the most abstract constructs related to any Data Ecosystem. This stage consisted of numerous cycles of formalization modeling and verification of the constructs of the literature. It is worth mentioning that certainly the construction of the meta-model was strongly influenced by our personal experience.

Before presentation of the meta-model, the modelling practices and conventions used for designing and illustrating the meta-model are described in Section 4.5.1. The meta-model is founded on a few core constructs that are defined and formalized in Section 4.5.2.

4.5.1 Modelling practices and conventions

The modelling approach selected for this thesis is based on the Meta Object Facility (MOF). MOF is an industry-standard developed by the Object Management Group (OMG) (SILVA, 2015). It is is especially applicable for definition, development and management of modelling languages and models created by those languages. In essence, the MOF provides a definition for a so-called meta-meta-model, that is, a meta-model for declaration of meta-models. The foundational elements of MOF include for example definitions for such concepts as *Class* or *Property* applicable for description of classes and their attributes. In particular, the proposed Data Ecosystem meta-model formalization relied on the Eclipse Modeling Framework (EMF)[6] and the meta-modeling language ECore provided by EMF. EMF is a core modeling framework and code generation facility for building tools and applications based on models defined in the Ecore (SILVA, 2015). There are several tools and frameworks developed on top of EMF such as Graphical Modeling

[6] http://www.eclipse.org/modeling/emf/

Framework (GMF)[7], Eugenia[8] and Epsilon[9]. In its turn, ECore is the native meta-meta-model language of the EMF and it complies with Essential MOF (EMOF) as part of OMG MOF 2.0 specification (SILVA, 2015).

Furthermore, the meta-model has been designed with a UML-modelling tool. The UML abstract syntax is defined using MOF as meta-meta-language. The notation elements of the UML are enough to understand the knowledge as well as suitable to express the basic constructs of a meta-model: as constructs of the domain under study, its attributes and relationships (including relationships "super-type/sub-type" and "concept-part"). In addition, there are several open source tools have been found that support UML diagrams. UML class diagram notation (SILVA, 2015) was utilized for designing and illustrating the meta-model elements. The UML-model was then imported to the Eclipse framework. During the import the UML-model was converted to Ecore meta-modeling language.

All occurrences of classes with the same name in the diagrams represent the same element of meta-model; correspondingly, the meta-model classes are uniquely named. A same class can occur more than once in the class diagram to ease the model reading. All associations are named in the meta-model and they are considered as public properties of their owning classes. Moreover, all association are attached with a multiplicity conforming the UML-notation.

Two kinds of associations are used in the meta-model: regular associations representing named references between classes and composite associations representing whole-part relationships. A regular association is a relationship between two classes. An association is a "using" relationship between two or more classes in which the classes have their own lifetime and there is no owner (SILVA, 2015). All the associations used in the meta-model are navigable only to the direction of the arrow-head. For example, the *Feature* class owns the association named *hasValue* as illustrated in Figure 14. The corresponding association can be read as "Feature hasValue FeatureValue". Composite associations are identified by a filled diamond shape. The diamond shape is located at the composite-end (the owning end) of the association. In Figure 14, a *DataEcosystemElement* has a composite association which includes zero or more Feature. This composition is used to denote conceptual "ownership" or inclusion of classes.

4.5.2 Data Ecosystem Meta-model

The envisioned meta-model is presented in Figures 13-17. As discussed above, the UML class diagram notation (SILVA, 2015) is used for illustrating the meta-model classes. The proposed Data Ecosystem meta-model is a single, unified model where every meta-model class is connected to at least one other class of the meta-model by specialization or

[7] http://www.eclipse.org/modeling/gmp/
[8] https://www.eclipse.org/epsilon/doc/eugenia/
[9] https://www.eclipse.org/epsilon/

association. In particular, the meta-model provides modelling constructs for specifying capabilities, structure, behaviour, artifacts, actors and relationships of Data Ecosystems. All the presented classes are defined at Appendix A. A summary of the main classes are explained as follows.

In the proposed meta-model, all classes are derived from *NamedModelElement* class. A NamedModelElement is a named abstract element of Data Ecosystem. This element may have a name and a description. The *name* property provides a label or identifier for the class, commonly a descriptive name. The name may or may not be unique, as determined by the rules of the actors responsible for the Data Ecosystem. The *description* property represents a text description of the element. It may contain a simple text string content, or carry a reference to an external description.

A DataEcosystem instance comprises a collection of *DataEcosystemElement*, as illustrated in Figure 13. The *domain* property provides the main category of the Data Ecosystem. A Data Ecosystem can have multiple domains. The *subDataEcosystem* association refers to a grouping of DataEcosystem instances that composes a major Data Ecosystem. The *composedOf* association specifies DataEcosystemElement instances that compose the Data Ecosystem. In particular, *DataEcosystemElement* is an abstract class that represents a basic unit/construct of Data Ecosystem. This includes actor, resources, relationships and role as presented in Section 4.4.

A DataEcosystemElement can have multiple Feature. A *Feature* class represents a distinctive attribute or aspect that characterizes a DataEcosystemElement. For instance, a dataset resource includes as features, type of data content, data model and other aspects that defines how data can be processed. A feature may or may not directly affect the DataEcosystemElement behavior. In addition, the feature may or may not be considered as information suitable for the discovery process. The *hasValue* association refers to a group of *FeatureValue* instances.

Both DataEcosystem and DataEcosystemElement instances may be subjected to a context. Context can refer to any information that can be used to characterize the situation of an entity, where an entity can be a person, place, or physical or computational object (ABOWD et al., 1999). The related context class were based on those proposed by (VIEIRA; TEDESCO; SALGADO, 2011), In particular, the *Context* class identifies the relevant contextual elements related to a DataEcosystem or DataEcosystemElement, represented by *ContextualElement* class. A contextual element includes legal, cultural, economic and other factors that play their role and influence the impact of a DataEcosystem or DataEcosystemElement. The *hasValue* association refers to a group of *ContextualElementValue*. A ContextualElementValue has *value* and *timestamp* properties. The timestamp represents a sequence of characters or encoded information identifying when a certain ContextualElement value was recorded, usually giving date and time of day, sometimes accurate to a small fraction of a second. The *ContextSource* class represents

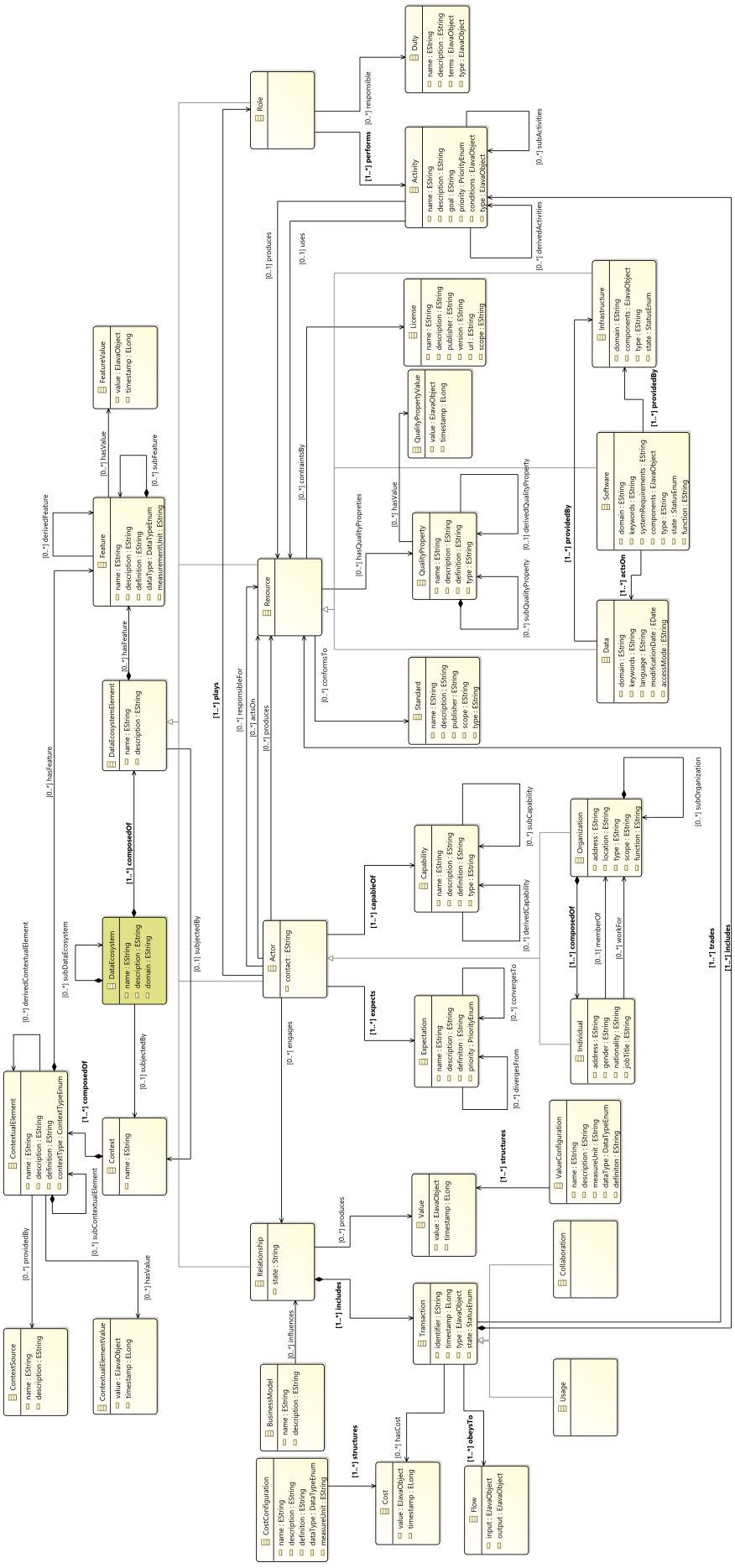

Figure 13 – Meta-model for Data Ecosystems. Source: Author

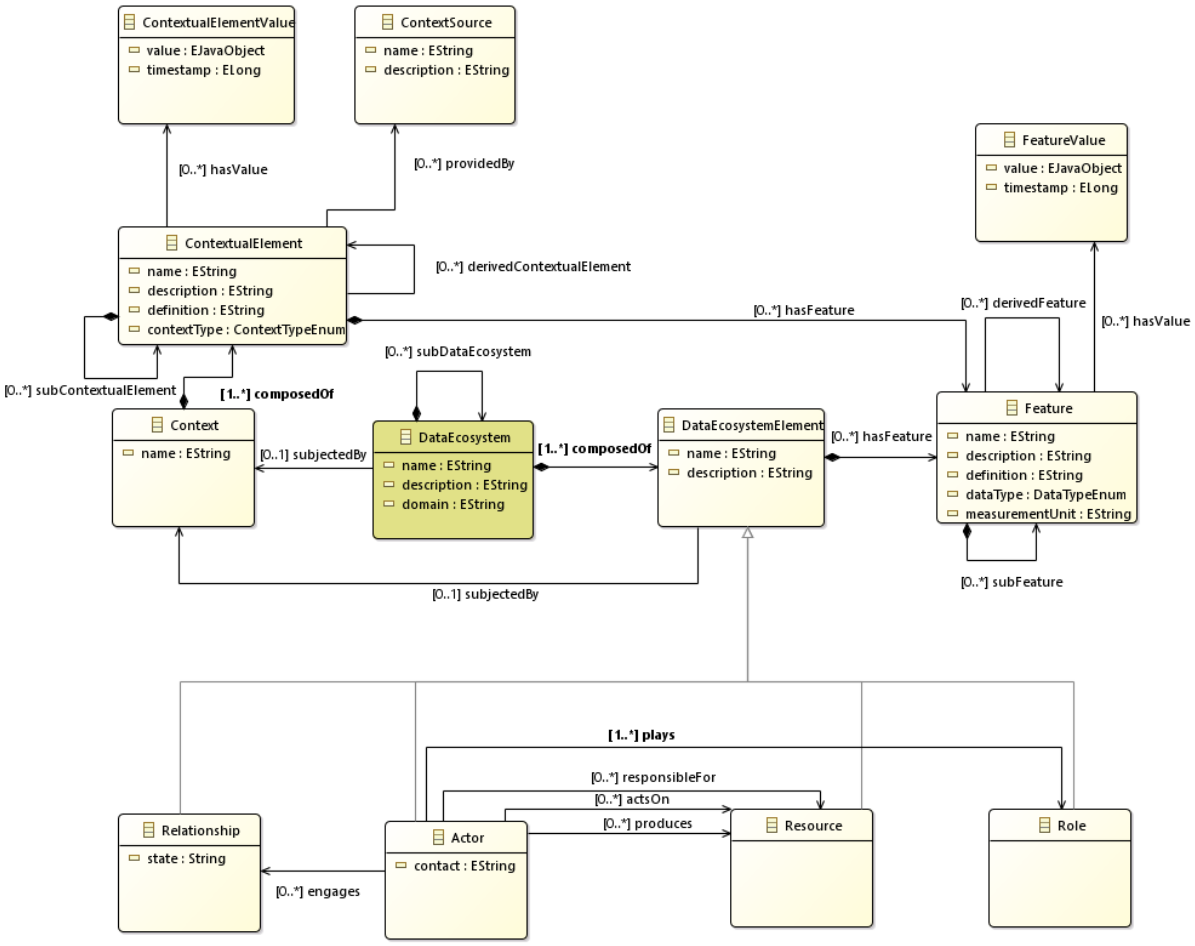

Figure 14 – Meta-model main classes. Source: Author

an entity responsible for certain ContextualElement.

Actor class is a sub-class of DataEcosystemElement, as illustrated in Figure 14. An Actor represents an entity that interacts consuming or providing a resource in a Data Ecosystem (*cf.* Figure 15). It is possible to name and to describe actors. The *Individual* class is a specialization of Actor, which represents a single person that exists as a distinct entity. The *Organization* class is a specialization of Actor, which represents an organized body of individuals with a particular purpose, especially a business, society, association, for example. There are a variety of legal types of organizations, such as corporations, governments and non-governmental organizations. An Individual can be linked to an Organization by using the *memberOf* and *workFor* associations. The *subOrganization* association refers to a group of Organization instances that composes a major Organization.

By using the *expects* association it is possible to define one or more expectations for an Actor instance. An expectation represents a feeling of expecting something to happen. It includes some goals, benefits and/or requirements to be accomplished. Actor's expectations are represented in the meta-model by the *Expectation* class. The Expectation class has two associations: *convergesTo* and *divergesFrom*. These associations allows to

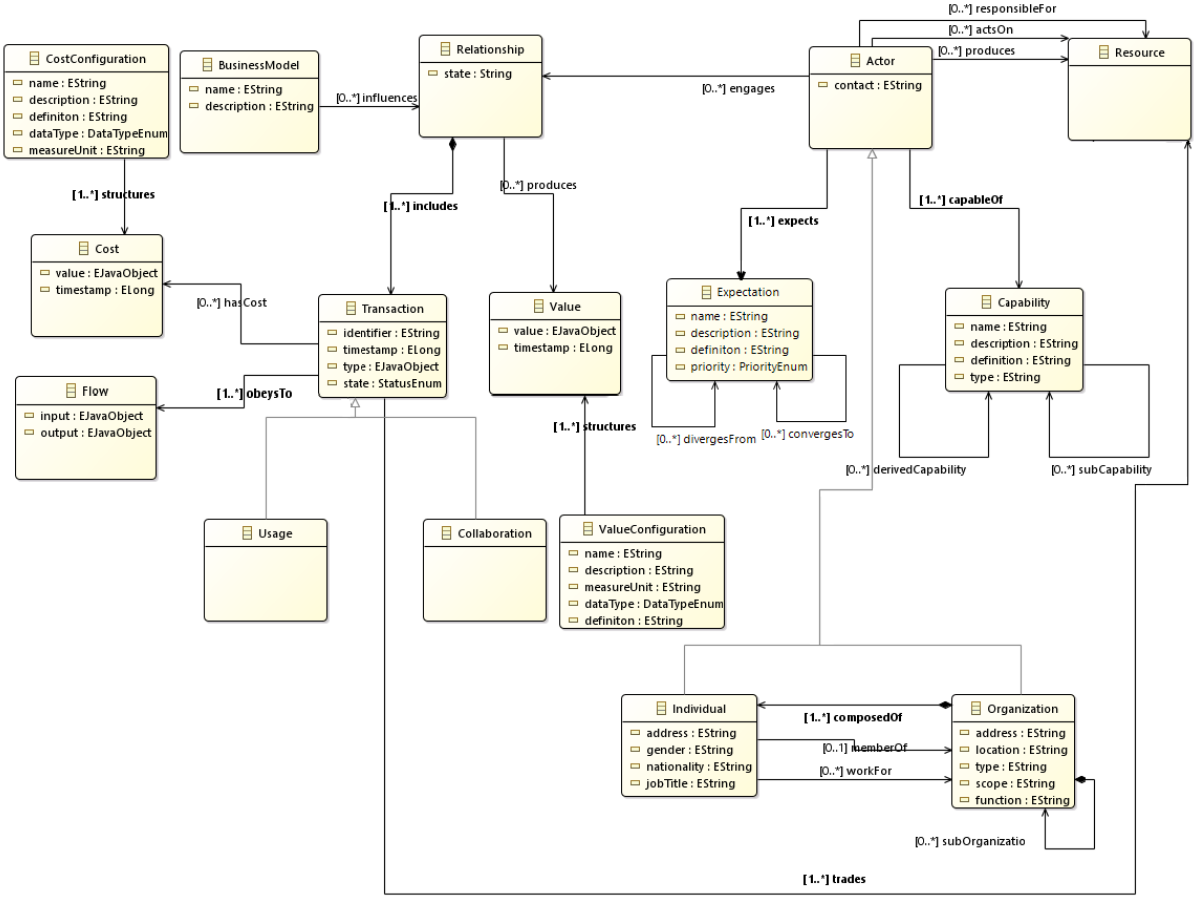

Figure 15 – Meta-model: Actor, Relationship and related classes. Source: Author

related different expectations with regards to their convergence or divergence.

By using the *capableOf* association it is possible to define one or more capabilities associated with an Actor. A Capability denotes an ability to perform actions. Actor capabilities are represented in the meta-model by the *Capability* class. The *actsOn* association links an Actor to to his/her consumed Resources instances. The *produces* association links an Actor to his/her produced Resources instances. Finally, the *plays* association links an Actor to his/her Roles played in the Data Ecosystem.

Role class is a sub-class of DataEcosystemElement, as illustrated in Figure 14. A role represents a function played by an actor in a Data Ecosystem (*cf.* Figure 16). The *performs* association links a Role to a group of duties. Each duty represents a moral or legal commitment that arise from a role. And, the *responsibleFor* association assigns to a Role a group of activities. Each activity represent a piece of work done as part of one's duties. The duty and activity constructs are represented by the *Duty* and *Activity* classes as presented in Figure 16.

Relationship class is a sub-class of DataEcosystemElement, as illustrated in Figure 14. A Relationship represents the way in which two actors are connected (*cf.* Figure 15). A Relationship have a state property, which specifies a condition or position that

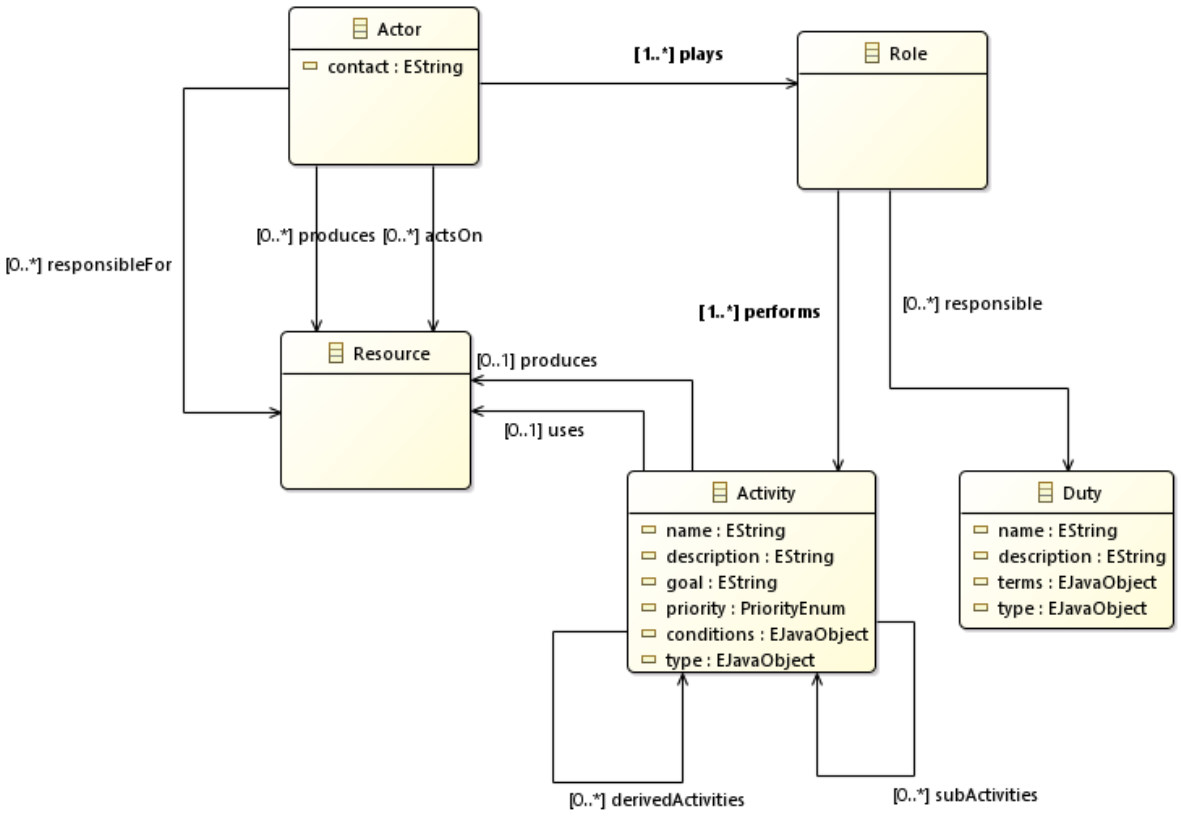

Figure 16 – Meta-model: Role and related classes. Source: Author

identifies whether a relationship is active. The business models construct represented by *BusinessModel* class describes the rationale of how a relationship creates, delivers, and captures value. The process of business model construction is part of business strategy. By using *influences* association it is possible to links business models to a set of relationships.

The *includes* association links a Relationship to a set of carried out transactions, which denote an exchange or transfer of Data Ecosystem resources. A transaction may have an identifier, type, state and timestamp. The *type* property describes a particular group of transaction that share similar characteristics and form a smaller division of a larger set. The *state* property describes a condition or position that identifies whether a transaction is active. Moreover, a transaction may have an associated cost as well as may obey to a *Flow* that structures how a transaction is performed. Furthermore, the *trades* association refers to a group of Resources exchanged or transferred between Actors. *Collaboration* and *Usage* classes specialize the transaction construct. A collaboration represents a special kind of transaction in which two or more actors are working together to complete a task or achieve a goal. In its turn, a usage represents the one-way transactions in which there is no feedback to the providing Actor.

Resource class is a sub-class of DataEcosystemElement, as illustrated in Figure 14. A Resource represents a useful or valuable product, possession or capability produced, pro-

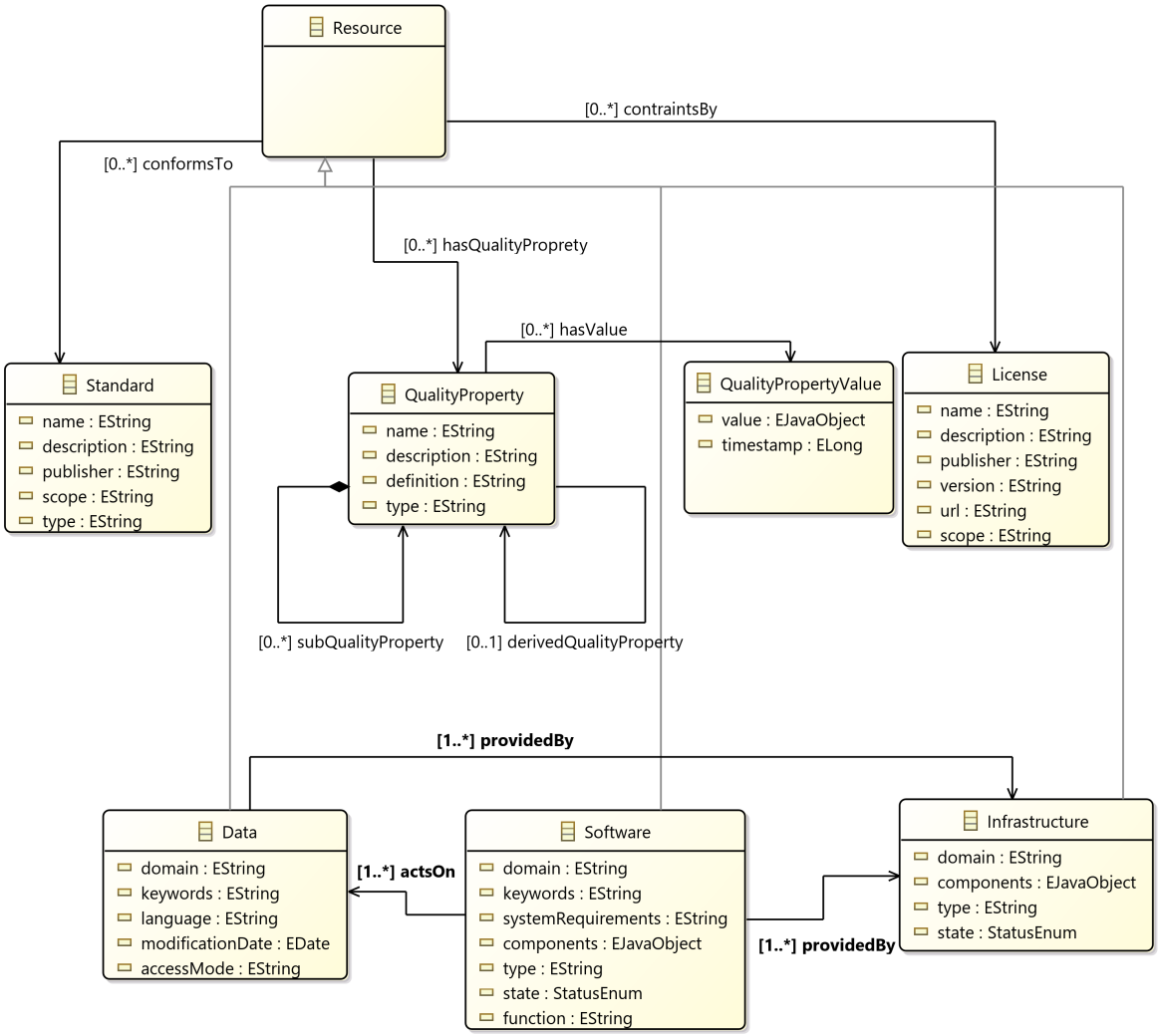

Figure 17 – Meta-model: Resource and related classes. Source: Author

vided, curated or consumed by Actors (*cf.* Figure 17). In Data Ecosystems, resources range from datasets and data-based software to infrastructure. The *Data* class is a specialization of Resource that represents data resources produced, provided, curated or consumed by Actors. The *Software* class specializes a Resource as a data-based software that are used to consume, produce or providing data. The *Infrastructure* class is a specialization of Resource that represents the underlying foundation or basic infrastructure that supports ecosystem activities. The *conformsTo* association assigns an established standard to which the described resource conforms. The *constrainedBy* association assigns an established license to which the described resource is constrained. And, the *hasQualityProperty* association refers to a group of quality properties, which are characteristics relevant to Actors (*e.g.*, the availability of a data). It is important to remark that there are several models formalizing quality management, such as Data Quality Vocabulary (DQV)[10] and

[10] <https://www.w3.org/2013/dwbp/wiki/Data_Quality_Vocabulary_(DQV)>

Quality of Service Ontology[11]. These models can be used to conceptualize and formalize specific aspects related to quality management.

Furthermore, the importance of temporal attributes in Decision Support Systems is broadly known. Some of the constructs formalized in this meta-model have a strong temporal nature. For example, the ContextualElementValue, FeatureValue, Cost, Transaction, and Value classes have a timestamp attribute that allows to track a time series. Other constructs could be extended to incorporate temporal attributes, such as QualityProperty. Relationships between classes could also generate tertiary classes in which temporal attributes could be recorded. However, these extensions would depend on the context in which the meta-model is being implemented. Finally, the timestamp attributes can be remodeled to allow more efficient and precise tracking. For example, using specific date manipulation formats allows you to analyze the time dimension through multiple levels of granularity such as day, month, and year.

4.6 META-MODEL EVALUATION METHOD

The *Evaluation* phase aimed at evaluating the developed meta-model. Usually, a meta-model evaluation implies making a technical judgment about its correctness. It is also important evaluating assessment of the practicality and utility of the proposed meta-model as well as other quality aspects, such as completeness and understandability. Hence, the conduction of the evaluation phase had the following objectives: (i) Validate the alignment of the meta-model classes with the different notions of Data Ecosystem; (ii) Assess quality and feasibility of the proposed meta-model; and (iii) Refine the proposed meta-model.

For this purpose, a case study was conducted in a real state-of-the-art Data Ecosystem. The case study was conducted in two phases: implementation and evaluation. The first phase consists of the implementation (*i.e.*, use) of the meta-model in a real case. In this phase, 10 participants (study subjects) were responsible for the implementation. More details about the case study context is presented in Section 4.7.2.

The second phase aims at evaluating the effectiveness of the application of the solution in the investigated context. To carry out this phase, the meta-model and artifacts obtained in the implementation phase were evaluated.

From the evaluation perspective, a promising alternative is to assess the quality of the meta-model from the evaluation of the quality of information derived from the meta-model. (LEE et al., 2002) presents a methodology, widely referenced by several authors, for information quality assessment called Assessment Information Methodology Quality (AIMQ), which covers an information quality model, a questionnaire for measuring the information quality dimensions and techniques for metrics interpretation.

[11] <ftp://ftp-sop.inria.fr/acacia/W3CAtelierWS/papers/QoSDescription.htm>

Another alternative is to evaluate the quality of the meta-model considering cognitive aspects. Cognition is "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses" (Oxford Dictionary, 2019a). A suitable approach to this type of assessment is the Bloom Taxonomy (FOREHAND, 2010), which establishes a hierarchy, composed of six categories, addresses the learning process and knowledge about a subject through a spectrum, ranging from the simplest to the most complex behavior.

Thus, both the AIMQ methodology and the Bloom Taxonomy can be considered appropriate alternatives for meta-model quality assessment. However, adaptations are necessary, while retaining the basic foundations of the two assessment methodologies.

Considering the AIMQ information quality dimensions, the meta-model was analyzed according to:

- Completeness: indicates if the stakeholders' perception of the meta-model contains all the concepts about Data Ecosystems (STRONG; LEE; WANG, 1997). In particular, a 'complete' meta-model should promote the structuring and understanding of the Data Ecosystem;
- Consistent Representation: refers to the extent to which the meta-model is presented in the same format (STRONG; LEE; WANG, 1997);
- Ease of operation: means the ease with which the meta-model can be derived to a model within time, budget and technology constraints (STRONG; LEE; WANG, 1997);
- Interpretability: refers to the extent to which the meta-model is represented in appropriate languages, symbols, and units, and the definitions are clear (STRONG; LEE; WANG, 1997);
- Objectivity: refers to the extent to which the meta-model is unbiased, unprejudiced, and impartial (STRONG; LEE; WANG, 1997);
- Understandability: refers as the ease with which the classes and other structures of the meta-model can be understood by stakeholders involved in Data Ecosystem
- Relevancy: refers to the ease with which the meta-model can cope with the purpose of modeling activity, *i.e.*, refers to the stakeholders' perception of the utility degree of the classes and other structures of the meta-model for the understanding about the Data Ecosystem under study (STRONG; LEE; WANG, 1997).

Considering the cognitive dimensions defined by Forehand (2010), the meta-model was analyzed according to:

- Knowledge: defined as remembering of previously learned concepts, *i.e.*, remembering appropriate information. It involves remembering the concepts, considering from specific facts to complete theories;
- Comprehension: defined as the ability to understand the meaning of knowledge;
- Application: refers to the ability to use learned knowledge in new and concrete situations;
- Analysis: refers to the ability to break down knowledge into its parts so that its organizational structure may be understood;
- Synthesis: refers to the ability to put parts together to form a new whole;
- Evaluation: is concerned with the ability to judge the value of knowledge for a given purpose.

The gathering of evaluation data was performed through questionnaires (electronic forms), which were composed of a series of questions used for collecting information from case study participants. A sketch of the questionnaire is available at <https://bit.ly/2sFocfP>. All the case study participants answered the questionnaire. It is important to remark that this questionnary does not represent a survey with the community. Moreover, the questions were answered without the presence of the interviewer. We used questionnaires because of their practicality and flexibility (MARCONI; LAKATOS, 2010). The questionnaire was composed basically by closed-ended questions adapted from AIMQ and Bloom Taxonomy, which aimed at evaluating the quality of the information derived from the proposed meta-model. Semi-structured interviews with open-ended questions were also applied, which is characterized by allowing the exploration of the participants' answers in depth without prior classification of closed-ended questions (MARCONI; LAKATOS, 2010). This procedure was used to clarify open-ended questions used in the questionnaire.

4.7 META-MODEL EVALUATION

The main idea underlying our work is that the proposed meta-model may be used to represent in an abstract way the relationships among the most essential constructs related to Data Ecosystems as well as to allow the development of Data Ecosystem specific models. After the proposal of the first version of the meta-model, it was performed a case study in order to evaluate it. It is important to emphasize that this case study was carried out as part of the Design Science Research approach presented in the Section 1.3, and not as a separate study.

4.7.1 Case Study Design

Our case study regards an instantiation of the meta-model considering a specific Data Ecosystem, in particular, an Open Data Ecosystem coordinated by a Brazilian public university. This case study, as mentioned in the 4.6 section, was divided into two steps: implementation and evaluation.

The implementation phase consisted of a meta-model instantiation, *i.e.*, the meta-model was derived into a model to represent the particularities of the ecosystem under study. This derivation involved: (i) the preliminary study of the meta-model, (ii) selection of a set of meta-model classes, and (iii) implementation of a concrete model derived from selected classes.

The meta-model instantiation was conducted with three partipants responsible for the creation of derived model. Other two participants helped to review and validate the derived model. The rest of participants contributed to select and analyze the documentation regarding the Data Ecosystem under study.

This selection was made mainly by looking for evidence of the use of the constructs formalized in the meta-model in the collection of documents related to the ecosystem, for example, Information Technology Master Plan, Data Governance and Open Data Plan documents. Such documentary research method was selected for this step as it provides concrete and detailed observation of information related to the ecosystem. In addition, a series of interviews were conducted with members of the ecosystem to complement the understanding of the ecosystem under study. Both strategies allowed to analyze the requirements and needs of the Data Ecosystem stakeholders, both internal and external, the available resources as well as the potential benefits and risks associated with the Data Ecosystem.

Based on the results of the documentary research and subsequent interviews, the following classes were selected: Actor, Resource, Role, Context, Activity, Duty, Standard, License, Expectation, Capability, Individual and Organization, Data, Software, and Infrastructure. The classes related to Relationship were not selected *a priori*, because the ecosystem is still in the creation stage and, as a consequence, there is not much information yet about relationships.

After selecting classes, a concrete model was implemented aiming at creating a catalog for documenting and promoting the ecosystem. This new concrete model included both the classes previously selected and specializations to represent the particularities of the ecosystem. For example, it was created a hierarchy of roles and also a taxonomy of organizations that are part of the ecosystem. This model was also developed using the EMF as a modeling platform. In addition, the catalog was used to gather information derived from the model and the meta-model as a consequence. The derived model is presented in Appendix B.

The evaluation step was made through questionnaires, which requests all the case

study participants to analyze a set of statements considering artifacts and information derived from the meta-model. The present section describes the case study context and also presents the results obtained with the questionnaires.

4.7.2 Case Study Context

As mentioned before, the research unit was a Brazilian public university. This university is developing an effort to create an Open Data Ecosystem. The main goals of the initiative are: (i) implementing a sustainable open data initiative, (ii) promoting the participation of members (*e.g.*, workers and students) of the university as well as society in general, (iii) disseminating the open data practices to the society and academic community, and (iv) promoting and enhancing processes for enabling transparency and free access to public information and data.

Since then, information about the Data Ecosystem efforts have been maintained in the form of various and dispersed electronic documents and several software applications. The open data publishing activities are also performed in *ad-hoc* way and carried out by multiple individuals. There isn't a consolidated method for publishing open data as well as there is no real knowledge of the existing open data demand. As a consequence, there are also no efficient means for evaluating results and planning the next steps of Data Ecosystem related activities.

The research subjects (aka participants) are the ones who provide the information to the case study. They are employees of the university who are conducting the open data initiatives, *i.e.*, they are directly involved in the consolidation of the university's Data Ecosystem. Moreover, we also included two members of the academic community, which acts mainly as data consumers. According to Runeson e Höst (2009), these research subjects can be considered adequate since they play different roles in the Data Ecosystem as well as they can provide significant information about the Data Ecosystem under study. They also have knowledge and experience with modeling techniques. In particular, the research subjects group consists of: two IT managers; two information science technicians; four developers with experience on the publication of open data; and two developers with experience on consumption of open data;

These participants are involved with open data practices in the university, with five grand areas: (i) structuring and planning the open data release process, (ii) creating and implementing data governance practices, (iii) developing open data pilot projects and (iv) selecting and applying information technology to support the open data practices.

Among the 10 participants, 50% were male and 50% were female. About 70% were aged 20−30 and 30% were aged 31−40. Of the participants, 80% were graduated under some Information System graduation or other computing graduation course and 20% were graduated under Information Science. Considering knowledge about Data Ecosystems, 20% had classified their knowledge as poor, 50% had classified as good and 30% had

classified as excellent. All of the participants perform some data publishing activity and 80% also act on some data consumption activity.

4.7.3 Evaluation Results

The case study participants evaluated the implemented model and its derived artifacts (*e.g.*, Data Ecosystem documentation) according to a set of information quality and cognitive dimensions.

These dimensions represent the identification of attributes considered fundamental to the meta-model understanding and its utility. Each dimension was evaluated using a set of statement. The participants if they agree with the statements using a scale with values ranging from 1 to 10, where the number 1 corresponds to "Strongly Disagree" and the number 10 corresponds to "Strongly Agree". Items labels with "(R)" are reverse coded [12].

Table 16 – Overall average of evaluation results of information quality assessment questionnaire

Scale	Total Responses
1 (Strongly Disagree)	1
2	1
3	3
4	2
5	3
6	4
7	36
8	62
9	69
10 (Strongly Agree)	89
Average	**8.58**

Table 16 presents the results obtained from the information quality evaluation. Column *Scale* lists each score (ranging from 1 to 10) used to evaluate the meta-model. Column *Total Responses* presents shows the total responses received by a score. For instance, the third line of Table 16 shows that only one participant choose the score 2 for a statement. Similarly, only one statement was evaluated with score 2.

Individually, the set of averages also indicates positive results in the quality of the information. The total average obtained (8.58) indicates that the information quality results were positive. In addition, over 81% of responses correspond to higher values of the scale (8, 9 and 10).

[12] In order to calculate the average, negative statements were considered with the inverse weight, since a negative answer for such affirmations is considered a positive result. For example, an inverse assertion evaluated as value 1 was considered as 10.

Table 17 – Detailed average of evaluation results of information quality assessment questionnaire

Dimension	Quality Statement	Average
Completeness	The meta-model derived information includes all necessary concepts.	9.00
	The meta-model derived information is incomplete. (R)	8.90
	The meta-model derived information is complete.	7.90
	The meta-model derived information is sufficiently complete for our needs.	8.90
	The meta-model derived information covers the needs of our tasks.	9.30
	Quality dimension average	**8.80**
Consistent Representation	The meta-model derived information is consistently presented in the same format.	9.10
	The meta-model derived information is not presented consistently. (R)	8.80
	The meta-model derived information is presented consistently.	9.10
	The meta-model derived information is represented in a consistent format.	9.20
	Quality dimension average	**8.90**
Ease of Operation	The meta-model derived information is easy to manipulate to meet our needs.	8.30
	The meta-model derived information is difficult to manipulate to meet our needs. (R)	8.70
	The meta-model derived information is easy to combine with other information	8.30
	Quality dimension average	**8.43**
Interpretability	It is easy to interpret what the meta-model derived information means.	8.00
	The meta-model derived information is difficult to interpret. (R)	8.70
	The meta-model derived information is easily interpretable.	7.70
	The meta-model concepts are clear.	7.90
	Quality dimension average	**8.08**
Objectivity	The meta-model derived information is based on facts.	8.50
	The meta-model derived information is objective.	8.80
	The meta-model derived information presents an impartial view.	7.30
	Quality dimension average	**8.20**
Relevancy	The meta-model derived information is useful to your work.	8.90
	The meta-model derived information is not relevant to your work. (R)	9.30
	The meta-model derived information is appropriate for your work.	8.80
	The meta-model derived information is applicable to your work.	8.70
	Quality dimension average	**8.93**
Understandability	The meta-model derived information is easy to understand.	8.20
	The meta-model derived meaning of this information is difficult to understand. (R)	8.90
	The meta-model derived information is easy to comprehend.	8.20
	The meaning of the meta-model derived information is easy to understand.	8.40
	Quality dimension average	**8.43**
Total Average		**8.58**

According to the Table 17, Interpretability and Objectivity scored the lowest values. This seems to reflect the constraints related to lack of both modeling skills and Data

Ecosystem background for some of the participants. However, in general, these results indicate that the meta-model was able to derive relevant knowledge for the participants and present it adequately.

Table 18 presents the total of responses attributed by case study participants evaluating aspects related to cognitive aspects. Like the previous tables, , the column *Scale* lists each score (ranging from 1 to 10) used to evaluate the meta-model. Also, column *Total Responses* presents shows the total responses received by a score.

The value obtained (8.63) indicates that the result was positive from the point of view of the learning objectives. In addition, over 81% of responses correspond to higher values of the scale (8, 9 and 10).

Table 18 – Overall average of evaluation results of cognitive quality assessment questionnaire

Scale	Total Responses
1 (Strongly Disagree)	0
2	0
3	0
4	0
5	5
6	4
7	31
8	56
9	55
10 (Strongly Agree)	69
Average	**8.63**

Table 19 presents a detailed view of responses related to each cognitive aspects. Individually, the set of averages also indicates positive results from the point of view of learning and educational objectives. The averages are: 8.7 for Knowledge; 8.7 for Comprehension; 9.0 for Application; 8.3 for Analysis; 8.6 for Synthesis and 8.6 for Evaluation. The final value (8.63) indicates that the result was positive from the point of view of the learning objectives. In addition, over 81% of responses correspond to higher values of the scale (8, 9 and 10). These results indicate that the meta-model was able to capture knowledge in the working environment of the participants.

After answering the questionnaires, the participants were interviewed about their evaluations of the meta-model. All of the participants agreed that the meta-model describe all the essential constructs of a Data Ecosystem. There seems to be no construct or notion missing. However, the majority of them stated that some constructs (*e.g.*, constructs for defining and ease the management of quality properties and more constructs related data resources) which probably make it more complete could be added in further extensions. The participants also confirmed that the information reflects the reality of Data Ecosystems in a general abstraction.

Table 19 – Detailed average of evaluation results of cognitive quality assessment questionnaire

Dimension	Quality Statement	Average
Knowledge	The main terms and concepts related to Data Ecosystems are present.	8.90
	Terms and concepts used in your work on Data Ecosystems are present.	8.30
	Terms and concepts used have little relation to your work. (R)	9.00
	Cognitive dimension average	**8.73**
Comprehension	The terms and concepts are correctly defined.	8.90
	It is possible to explain the theme/subject verbally.	8.30
	The meaning of the terms can be interpreted.	8.60
	It is not possible to justify facts from the terms. (R)	9.10
	Cognitive dimension average	**8.73**
Application	The information derived from the meta-model can be applied in my work.	8.90
	The information derived from the meta-model can be applied in new situations of my work.	8.80
	The information derived from the meta-model is not applicable to my work.	9.30
	Cognitive dimension average	**9.00**
Analysis	The information derived from the meta-model allows to identify reasoning failures.	6.70
	The information derived from the meta-model allows to identify the whole and its parts.	8.50
	The structure of information derived from the meta-model is adequate.	8.70
	The hierarchy of terms and relations is coherent.	8.80
	The structure of information derived from the meta-model is not adequate. (R)	9.00
	Cognitive dimension average	**8.34**
Synthesis	The information derived from the meta-model makes it possible to write about the theme/subject.	8.60
	The information derived from the meta-model makes it possible to elaborate solutions to problems.	8.50
	The information derived from the meta-model allows for new ways of classifying ideas within the subject.	8.80
	The information derived from the meta-model allows to produce a unique language on the subject.	8.30
	Cognitive dimension average	**8.55**
Evaluation	The information derived from the meta-model allows judging the adequacy of conclusions related to Data Ecosystems	8.20
	The information derived from the meta-model allows judging a fact related to Data Ecosystems	8.70
	The information derived from the meta-model does not allow adequate conclusions related to Data Ecosystems (R)	9.00
	Cognitive dimension average	**8.63**
Total Average		**8.63**

They also considered the meta-model relevant to the Data Ecosystem description as well as the reasoning behind it. In addition, the meta-model implementation was essentially relevant for all of them. Moreover, they argued that the meta-model identifies the

relevant constructs and relationships between them and additionally contains a set of logical assertions that ease the modeling of a specific Data Ecosystem. They also agreed that the meta-model and the meta-model implementation helps to understand their Data Ecosystem as well as contribute positively to their decisions about the ecosystem in question.

Regarding ease of operation, most of the participants judged the meta-model to be easy to use. One of the positive aspects pointed out in the ease of operation is a synthetic set of classes. The majority of them argued that an excessive list of classes could compromise both the understandability and operationalization. Two participants considered that the meta-modeling instantiation is straightforward. However, they recommended to look first into the literature as well as to look into examples to fully understand the classes defined by the meta-model.

Meanwhile, two participants who have little modeling skills had some difficult to interpret the meta-model. They suggested including a guide to explain the notation used and also including more examples to better illustrate the classes. In fact, the lack of knowledge of the modeling language was pointed out by other participants as a problem that can affect Interpretability. Moreover, a participant also recommended redesign the meta-model using different levels of abstraction. Each level involves a unique set of classes and compositions. While higher levels build on a relatively general and simplified view of Data Ecosystems, lower levels provide an increasingly detailed representation of a Data Ecosystem.

As for Consistent Representation, the participants judged that the meta-model appears to be consistent regarding syntax and format. However, they argued that is difficult to assess the correctness by looking at the models. According to them, examples would also contribute to demonstrate correctness and also for completeness to see if some classes is missing.

Some of the participants identified several of the classes are related to their activities, but they did not recognize them as part to Data Ecosystems. In addition, they argued that the relationships between these classes allowed them to identify 'blindspots' in the management of their ecosystem activities. According to one participant, the meta-model brought guidance on how the organization can manage their data resources, for example. Another finding reported was related to the management of data resources. According to two participants, their ecosystem management is fragmented and *ad-hoc*. By using the meta-model, it was possible to perceive the need for holistic management of the ecosystem.

4.8 DISCUSSION

According to participants' reports, the implemented model was potentially applicable and useful to support the understanding of their Data Ecosystem. It was said that it could also provide support for management activities of a Data Ecosystem. On the other hand,

it was also verified that the case study contributes to improving the meta-model itself as expected from Design Science Research approach.

The case study outcome was successful as: (i) it contributed to the solution of the lack of consensus about Data Ecosystem, as it consolidated the constructs necessary for a meta-model for describing Data Ecosystems; (ii) it supported the improvement of the clarity of definition of the classes of the meta-model; (iii) the feasibility of the meta-model was consolidated through its implementation and quality evaluation; and (iv) it provided the improvement of artifacts related to the management of the Data Ecosystems, serving as a potential information repository tool to support the coordination of the Ecosystem under study.

It is important to note that the proposed meta-model does not aim to model an exhaustive set of classes related to Data Ecosystems. The Data Ecosystem field has further representational requirements than the ones presented in the meta-model. Data Ecosystems are distributed, which requires being aware of the different types of available resources; are heterogeneous, which requires knowing how to access and consume these resources. Furthermore, in addition to the classes just mentioned, a Data Ecosystem also needs an appropriate way to model domain-specific management perspectives, such as business process and quality enforcement, so that practitioners can link the Data Ecosystem with their operational and business activities. A promising alternative is a multi-layer meta-model. Each layer formalizes a valid aspect of functioning and management of the ecosystem.

It is also important to note that the way of using the meta-model may differ according to the context or domain of the Data Ecosystem. For example, the meta-model does not necessarily need to generate a new concrete model. It can be used to guide the implementation of a computer system that includes data entry, processing and output. It can also be used as a conceptual basis for an integration strategy between existing applications in which elements related to the Data Ecosystem are already spread in multiples systems. At the other extreme, it can be argued that the meta-model can be used to catalog information contained in electronic documents (usually spread over several computers and under the custody of several individuals) by conceptual criteria. In any case, it is assumed that the model will support the creation and coordination of ecosystems by explaining what elements need to be monitored and how they are related.

4.9 CONSIDERATIONS OF THE CHAPTER

The meta-model described in this Chapter was designed to support the description of Data Ecosystems. In general, the meta-model formalize fundamental constructs of Data Ecosystems extracted from theoretical constructs obtained from the literature as well as dependencies between the constructs. These constructs are represented using an appropriate notation and, according to our case study, are capable of representing the main

constructs of a Data Ecosystem. To the best of our knowledge, this is the first meta-model for Data Ecosystems evaluated in the context of a real Data Ecosystem and by practitioners with experience in the field.

Moreover, as presented in Chapter 3, research on Data Ecosystems still lacks a greater understanding of how these ecosystems are characterized and how they work. Thus, due to lack of theory in Data Ecosystems, the meta-model presented in this Chapter allows to understand what are the main elements related to a Data Ecosystem and how they relate.

This understanding served as the basis for building the Louvre framework. First, we perceived the importance of relationships for maintenance and evolution of a Data Ecosystem. These relationships should be exploited to carry out different activities from the resource exchange. In our case, the relationships should be explored to perform metadata curation through a collective effort. In addition, since each participant has his own set of expectations, a metadata curation framework should address not only to disseminate the importance of metadata and metadata curation, but also to propose ways for engaging new participants. In the same way, since actors do not necessarily belong to the same organization, mechanisms and actions must be proposed that facilitate, stimulate and manage communication among the actors. Such communication coordination is an essential requirement for a metadata curation framework for Data Ecosystems.

Hence, the main idea underlying the proposed meta-model is to provide means to the constructs related to Data Ecosystem in such a way that it can be used as a conceptual framework to support the building of specific Data Ecosystem model or to design tools derived from the meta-model. Moreover, depending on the particular requirements of the ecosystem at hand, it allows the use of only part of the defined constructs.

Furthermore, the implementation of the meta-model in specific Data Ecosystems can generate new propositions that evolve the classes of the meta-model. Among many possibilities of applying the meta-model is the creation of an interrelated set of documents (*e.g.*, catalogs and planning documentation) or even a computerized system that implements the meta-model classes. Another practical implementation is to use the classes of the meta-model for the derivation of metadata related to Data Ecosystems.

In this sense, the meta-model proposed can also be viewed as a metadata specification intended to establish a common understanding of the meaning or semantics of the metadata related to Data Ecosystems. Thus, we can identify the meta-model proposed as a core metadata specification that is useful for several kinds of Data Ecosystems. Moreover, the proposed meta-model can be used to support (i) several modeling levels, (ii) homogeneous manipulation of these different levels and (iii) extensibility of the corresponding models and meta-models. Each modelling level formalizes a valid aspect of functioning and management of ecosystem. Such extensible approach allows to adapt the meta-model to specific Data Ecosystem domains as well as allows integration and interoperation of

metadata coming from different sources.

Next Chapter presents the Louvre framework describing the approach for creating the framework, its values and principles and its current structure.

5 THE LOUVRE FRAMEWORK

The heterogeneous, distributed and dynamic nature of Data Ecosystems requires active metadata management in order to ensure trustworthiness, integrity, access, fitness for use and reusability of metadata (CHESSELL, 2016). This active management is part of an emerging and growing discipline referred as metadata curation dedicated to the present and future maximization of metadata potential (CHESSELL, 2016). If the metadata curation activities are undertaken or neglected, it can compromise long-term access to metadata, and as a consequence it may hinder the growth of Data Ecosystems (STOCK; WINTER, 2011; DINTER; SCHIEDER; GLUCHOWSKI, 2015; XIAO et al., 2015). Hence, unless metadata can be preserved over time, current investment in Data Ecosystems will only secure short-term rather than lasting benefits.

Metadata curation provides the methodological and technological metadata management support to address quality issues maximizing the usability of the metadata as well as ensuring metadata preservation. Metadata curation methodologies covers different activities such as metadata acquisition, selection, classification, transformation, validation, and preservation. According to Higgins (2008), curation methodologies should define the basic functional components of a project dedicated to the curation. These descriptions would be expressed in terms of a well-defined set of concepts.

Furthermore, the selection and implementation of a metadata curation process is a multidimensional problem (FREITAS; CURRY, 2016), depending on the interaction between economical and cultural aspects, resources availability, standards, and technological dimensions. Moreover, in Data Ecosystems, making metadata curation attractive involves getting actors to a point where they understand the concepts, perceive potential benefits, and also envision a path to get there.

In general, the main idea behind this thesis is to propose a common framework for describing a metadata curation process in Data Ecosystems. It integrates several metadata management disciplines. It focus on the collaboration effort to perform metadata curation. It also adopts and adapts some Software Engineering values and practices. The framework was developed after an extensive literature review, during which other similar models and frameworks were analyzed in search for metadata curation activities. In addition, the experience of the Aladin research group[1] with a number of data production and data consumption projects served to give a practical foundation to Louvre framework.

[1] http://aladin.rf.gd/aladin

5.1 MOTIVATIONAL SCENARIO

We extended the running example presented in Section 4.2 to exemplify the use and importance of metadata curation in a Data Ecosystem. For this, we will describe some metadata curation issues related to four fictional actors, called Alceu, Elba, Fagner e Capiba. Alceu and Elba were already introduced in Section 4.2.

As mentioned before, Alceu is a government technician responsible for publishing the Budget Expenses and Budget Financial Statements datasets. The motivation of sharing these datasets is not only to promote transparency in relation to public spending, but also to enable the generation of value and innovation through the development of services and applications based on these data. So, different actors of the Data Ecosystem must be able to consume these datasets. Such motivation increases the need for descriptions of datasets. From Alceu's view-point he only needs to provide descriptive metadata, such as keywords and description notes. In order to provide these metadata items, Alceu publishes them as a HTML table in the Web page that hosts each dataset. However, Alceu's view does not care about the fact that other essential information like structural metadata, license and provenance are also needed by other consumers. Moreover, he also neglects other important issues like the preservation of metadata.

As mentioned before, Elba is a financial journalist who plans to utilize the Budget Expenses and Budget Financial Statements datasets to perform a longitudinal study on the government administration finances. She wants to gather and analyze the financial transactions of the government administration in order to pursue overpaid contracts. For Elba, it is crucial that these datasets have some quality properties. In particular, these quality properties include completeness (are values missing?), timeliness (does the metadata represent reality from the required point in time?) and accuracy (the degree to which the metadata represents reality?). Elba herself can verify these quality properties. However, she needs both procedures and tools to support the verification process. Any problems found by Elba should also be shared back with Alceu or other data consumers. But, she does not have access to any mechanism to share her findings.

Fagner is a system developer responsible for the My Government Budget site. This site uses the Budget Expenses and Budget Financial Statements datasets to provide some charts and easy-to-use interfaces to analyze government financial data. In order to develop this site, he needs to produce a set of metadata, which was not provided by Alceu. For instance, he created structural metadata to automatically process data. He also created operational metadata to automate the access of datasets. He wants to share these metadata with other data consumers. However, the open data portal that host Budget Expenses and Budget Financial Statements datasets does not provide any mechanism to allow external contribution.

Capiba is a researcher. His main research area is Web Semantic. Currently, he is developing a taxonomy and a vocabulary to describe government financial data. As a

case study, Capiba semantically enriched the Budget Expenses and Budget Financial Statements datasets. In order to share this refinement with other actors, he publishes the enriched datasets and their new semantic metadata in his own web site. If the datasets are updated by Alceu or another technician, the refinements made by Capiba may become obsolete.

All the mentioned actors are acting on the same Data Ecosystem's resources. They are creating, storing, consuming, evaluating or enhancing metadata related to these resources. However, as they are working separately, their efforts could be lost. Moreover, other Data Ecosystem resources are neglected in terms of metadata management. For instance, the My Government Budget site is an important resource, however no metadata is created or preserved to it.

Such scenario would benefit from a metadata curation framework to guide actors that are not physically co-located to collaborative creation, sharing and management of metadata. A principal commitment is to understand the metadata curation requirements of Data Ecosystem community, with the dual aim of promulgating good practice or proven solutions among that community and informing the orientation of metadata curation activities and tools in order to address their needs and requirements. This framework should enable the creation of an environment where greater collaboration can be explored.

5.2 FRAMEWORK DEVELOPMENT METHOD

The construction of the framework took place through an iterative and incremental process. In particular, we attempted to continually answer the following question: "How should a metadata curation framework be structured to guide metadata curation in Data Ecosystems in a collaborative, adaptive, and flexible manner while addressing particularities related to Data Ecosystems?"

As a result of the literature review, a consistent theoretical basis was consolidated, which guided the construction of the proposed framework. Initially, the alpha version of the Louvre framework received influences mainly from ISO 12207 (ISO, 2008) and some works identified during the execution of literature review. Among these works, it is possible to highlight the DMBOK (MOSLEY et al., 2010), the family of standards ISO 8000 (ISO, 2011), and DCC Curation Lifecycle (HIGGINS, 2008). The ISO 12207 was used as the standard for process specification. The DMBOK and the ISO 8000 family influenced the process definition of the alpha version of the proposed framework. The DMBOK still inspired the design of a structure based on dimensions, and the DCC Curation Lifecycle influenced the definition of Louvre's dimensions based on the stages of its lifecycle.

Given this context, we aim to bring together in the Louvre framework a set of best practices for curating metadata in Data Ecosystems. It is important to emphasize that the Louvre both adapts activities related to management of data and metadata initially gathered from DMBOK and ISO 8000 and extends these compendiums by proposing

new activities for curating metadata based on the theoretical framework raised from the literature review.

The Louvre was also strongly inspired by the theoretical basis created both from the systematic mapping and through the formalization presented in Chapter 4. The landscape presented in Chapter 3 allowed the identification of a knowledge base that could be applied in the construction of the framework proposed. In particular, the decentralized configuration of Data Ecosystems together with the absence in many cases of dedicated actors to perform ecosystem management activities require the proposition of methods that are more collaborative and less process-centric. In addition, the proposed framework should also be focused on promoting, at least indirectly, the network character of Data Ecosystems. The meta-model presented in Chapter 4 allowed to identify a set of essential elements that calls for metadata. This meta-model should then be used as a basis for identifying and modeling more specific aspects, for example the description of app and services.

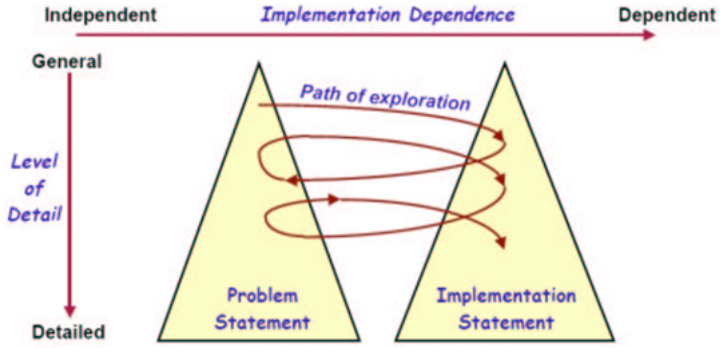

Figure 18 – The Twin Peaks model. Adapted from: (NUSEIBEH, 2001)

From the definition of the initial set of practices, the design of Louvre was continually verified and refined. Such construction process in which the refinement and verification phases were performed cyclically is slightly inspired by Bashar's Twin Peaks Model (NUSEIBEH, 2001), which is an iterative process that produces progressively more detailed requirements and design specifications, as Figure 18 suggests.

In this sense, the construction started through the elaboration of an initial version of the framework based from a sketch conceived through an extensive review in studies (theoretical and practical) of curation and management of information, data, and metadata. This initial version was improved through several cycles of verification and refinement. In particular, the verification was performed by the analysis and observation of the framework about the adherence of the key factors and how it would work in practice. All the elements that compose the Louvre were also verified on how each of them could contribute to the definition of the Louvre and how they could contribute to metadata

curation in Data Ecosystems. After that, the results of the verification process were used in the refinement of the proposed framework.

5.3 LOUVRE: A METADATA CURATION FRAMEWORK

The Louvre is structured to guide in a practical way the curation of metadata in Data Ecosystems. It is based on a set of dimensions, processes, purposes, activities, practices, and outcomes. The Louvre considers the body of knowledge related to metadata curation, digital curation and data management, which was adopted as a reference for the construction of the framework.

The Louvre framework follows a more descriptive than prescriptive approach. A prescriptive approach involves specifying, or even imposing, to individuals how they should do, rather than giving suggestions or describing what should be done. In its turn, the Louvre focuses on presenting which activities and practices should be considered by actors interested in curating metadata related to their Data Ecosystems.

Such descriptive approach allows adapting to the reality of a Data Ecosystem. Moreover, as far as the metadata curation work is being performed, a set of curation activities and practices are chosen to be implemented. This approach is desirable because the knowledge about curation increases as processes, procedures, services, and tools are being implemented.

Finally, the Louvre framework does not intend to replace existing models of data and metadata management. The Louvre was designed to be a complement to other reference models related to metadata management and to become an alternative focused on curation of metadata.

5.3.1 Principles, Values, and Mission

As mentioned before, the Louvre framework is based on the premise that metadata curation should be be practically performed and needs to rely on a collaborative effort and should also embrace adaptability and flexibility. Given this scope, some principles stand out, being:

- Practical Focus: The Louvre was designed to complement the lack of other existing curation and management models with regards to metadata curation by defining a set of elements that aid actors to curate metadata in Data Ecosystems. Moreover, the Louvre focuses on providing solutions that either curate metadata or support curation tasks. That is, each process is oriented to produce a set of outcomes. Being practical-focused means taking a holistic view of the overall problem, which can lead to suggested updates in infrastructure, processes, and overall organizational structures.

- Adaptability: Different contexts require different strategies (AMBLER; LINES, 2011). The Louvre allows and encourages actors committed to the curation work to freely identify which of the dimensions, processes, practices, and activities enable them to achieve their curatorial goals and to decide what should be employed in their current context.

- Collaborative effort: The Louvre looks for encouraging actors into a Data Ecosystem to perform the curation of their corresponding metadata. We believe that the engagement of actors is crucial in order to develop successful and effective curation of metadata.

- Flexibility: The Louvre does not specify technologies, systems or standards, which gives actors the flexibility to adapt the curation of metadata of diverse nature and volume.

The increasing requirements for planning and execution of systematic metadata management and curation have resulted in limited impact in practice. Moreover, in Data Ecosystems it is not possible to assume the existence of dedicated entities as well as corresponding funding to support of metadata curation activities or development of adequate infrastructure tools to facilitate metadata curation.

The implicit or explicit adoption of agile software development practices has allowed other domains (*e.g.*, project management (HIGHSMITH, 2009; FERNANDEZ; FERNANDEZ, 2008) and governance (LUNA, 2009; ALMEIDA, 2015)) to achieve success to reduce efforts and to overcome distributed environment issues. Given this context, in order to reduce the efforts demanded for curation work, the Louvre framework also adheres to and adapts the values of the Agile Manifesto (BECK et al., 2001) to metadata curation context:

- Individuals and interactions over processes and tools;

- Discoverable, understandable and usable metadata over comprehensive documentation;

- User collaboration over contract negotiation;

- Responding to change over following a plan.

The principles and values described above complement the mission, which is providing an understanding and increased awareness for curation metadata and agile and collaborative practices, as well as providing a basis for understanding the concepts, activities, and practices needed by actors for curating metadata in Data Ecosystems.

Table 20 – Main Elements, Influences and Inspirations

Louvre Elements	Influences	Inspirations
Dimension-based Structure	The Louvre organizes its processes in dimensions.	M3 (MAnGve Maturity Model)
Dimensions' Scope	The Louvre framework defines its dimensions following the stages proposed by data curation lifecycles.	DCC; DDI; OAIS; DataONE
Process Structure	The Louvre framework defines a standard structure for specifying its processes.	ISO/IEC 12207; RiSE-RM; Mangve; M3 (MAnGve Maturity Model)
Processes Specifications	The Louvre framework specifies its processes based on a set of initiatives involving data management and data curation.	DMBOK; ISO 8000; DCC; DDI; OAIS; DataONE; PMBOK; COBIT
Agile Practices	The Louvre framework tailors agile practices based on a set of initiatives involving agile management.	Mangve
Actors and Roles	The Louvre framework specifies its roles and how actors can organized based on a set agile software development methods	SCRUM; Disciplined agile delivery
Deployment Method	The Louvre recommends a deployment method inspired in the PDCA cycle.	Mangve
Graphical Representation	The Louvre framework presents its graphic representation in order to relate its elements.	Mangve

5.3.2 Structure

Louvre's structure was inspired by standards, models, and frameworks investigated in the course of this thesis. Table 20 presents some of the main influences of these works for the construction of the framework. In particular, the structure is similar to those used by RiSE-RM (GARCIA, 2010) and Mangve (LUNA, 2009). All of these works were based on the ISO (2008).

As presented in Figure 19, the Louvre framework is composed of the following structural elements: actors and roles, agile practices, dimensions, processes, activities, purposes, input, outcomes, work products and a deployment method. Some of these elements are used to group other (*i.e.*, dimensions and processes). Moreover, all these elements can be seen as enablers for metadata curation, through which the curation is guided and goals can be achieved. These elements are detailed as:

- Actors and Roles: recommend how actors may be organized to curate metadata as well as defines a set of roles involved in performing and supervising the curation work.

- Agile Practices: adopt and tailor strategies from mainstream agile methods to support curation of metadata.

- Dimension: groups a set of correlated processes related to one or more stages of the metadata lifecycle.
- Process: groups a set of correlated activities that are executed in order to generate expected outcomes (ISO, 2008). It indicates key curation areas where actors should focus on enabling curation of metadata. The set of activities does not imply any prescriptive order in their use. The order of activities is influenced by multiple factors, including organizational and technical considerations, each of which can vary according to the Data Ecosystem context.
- Activity: represents a set of actions defined to achieve a specific result (ISO, 2008).
- Purpose: represents a high-level goal related to a process (ISO, 2008).
- Input: represents tangible artifacts that each process needs to initiate its activities.
- Outcome: describes an observable result of the achievement of the process purpose (ISO, 2008), including the production of an artifact, a significant change in the Data Ecosystem, project or environment, meeting of specified requirements or goals, for example (GARCIA, 2010).
- Work Product: represents an artifact produced by a process. The set of work products is not an exhaustive list of required results, but rather just examples that can help in the implementation of these processes (ALMEIDA et al., 2015). Work products can include documents, specifications, metadata items, and software, including reusable assets (components and services) or software assets (applications).
- Deployment Method: provides procedures for adopting and improving metadata curation process in the Data Ecosystem context.

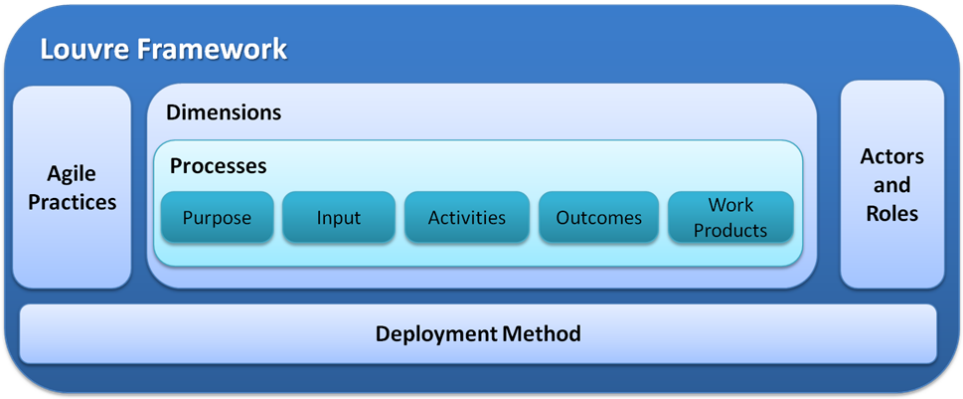

Figure 19 – Louvre Structure. Source: Author

The metadata curation work is captured in dimensions, processes, and activities. These elements describe or present a coherent body of actions and functions that systematize

metadata curation work. In particular, a process shows the activities that needed to produce a particular set of outcomes. In its turn, agile practices describes strategies and techniques that can be used to support the systematic elements (*i.e.*, dimensions, processes and activities). Hence, such agile practices are another kind of structural elements that are not directly related to a systematic description of some curation task to be performed.

The dimensions, processes, activities and agile practices enable the separation of concerns regarding to the metadata curation work, *i.e.*, these elements subdividing metadata curation work into independent, interchangeable modules. Moreover, all of the elements inside each of such modules are semantically related. In addition, such modularization allows the Louvre to be adapted according to the context in which the framework will be applied, in which the needs and preferences of the ecosystem that will be adopting the framework should be considered.

In addition to the components described above, the Louvre framework also recommends that the implementation of the metadata curation initiative should be performed through an iterative and incremental method. This approach is inspired by both agile software development models and the Plan, Do, Check and Act (PDCA) method (FONSECA; MIYAKE, 2006). In which, each iteration is chosen a set of processes, activities or practices to be implemented or improved. This method is iterative because the work is always performed and improved in subsequent iterations. At the same time, it is incremental because it performs staged planning in which various curation-related concerns can be developed in parts. The use of iterations allows greater adaptability, since the processes and practices are adapted to the current Data Ecosystem context. This approach allows continuously controlling and improving processes and services, increases the chances of delivering better results to metadata curation initiatives in Data Ecosystems. Finally, this iterative approach also allows taking advantage of what has been learned in previous iterations.

5.3.3 Louvre Framework Versions

As mentioned in Section 5.2, the construction of the Louvre framework started initially from the knowledge acquired during the execution of the literature review and other activities carried out in the exploratory research phase (cf. Figure 1). Table 21 presents the main elements of this first developed version (alpha version). In this version, the framework was composed of 6 roles, 8 agile practices, 6 dimensions and 19 processes.

The alpha version was evaluated through a survey (cf. Section 6.1). During this evaluation, the proposed framework was validated and some improvements were pointed out. From these results, the framework was then refined into its beta version. Table 22 presents the agile practices, roles, dimensions, and processes specified for beta version.

The beta version of Louvre framework was composed of 19 processes, which continued to be organized in 6 dimensions. However, the Metadata Curation Planning dimension

has been refined and renamed to more clearly represent the analysis process. In addition, some refinements were made in several processes based on the results of the previous version evaluation. In particular, the Metadata Curation Platform Maintenance process was refined to increase more platform management activities.

The beta version of the Louvre framework was evaluated through a focus group (cf. Section 6.2). During this evaluation the improved version of Louvre framework was verified and some new improvements were pointed out. From these results, the framework was evolved to its version 1.0. During this evolution, some processes were refined. Version 1.0

Table 21 – Louvre Framework - Alpha Version

Elements		
Actors and Roles	CurationMaster Stakeholder MetadataCurator TechnicalExpert PlatformOwner TeamLeader Stakeholder	
Agile Practices	User Story Persona Backlog Continuous Integration Continuous Refactoring Automated Tests Collective Ownership Burndown Chart	
Dimensions	Metadata Curation Planning	Metadata Curation Requirements Engineering Metadata Curation Planning
	Metadata Acquisition	Metadata Creation Management Metadata Harvesting Management Metadata Model Management Metadata Appraisal and Selection Management
	Metadata Quality Management	Metadata Quality Control Metadata Quality Improvement
	Metadata Preservation and Dissemination	Metadata Ingest Management Metadata Versioning Management Metadata Integration Management Metadata Access Management
	Metadata Curation Monitoring and Controlling	Metadata Curation Monitoring Recruiting and Engagement Management Communication and Feedback Management Metadata Curation Coordination
	Metadata Curation Platform Administration	Metadata Curation Platform Design Metadata Curation Platform Implementation Metadata Curation Platform Maintenance

Table 22 – Louvre Framework - Beta Version

Elements		
Actors and Roles	CurationMaster Stakeholder MetadataCurator TechnicalExpert PlatformOwner TeamLeader Stakeholder	
Agile Practices	User Story Persona Backlog Continuous Integration Continuous Refactoring Automated Tests Collective Ownership Burndown Chart	
Dimensions	Metadata Curation Analysis and Planning	Metadata Curation Requirements Engineering Metadata Curation Planning
	Metadata Acquisition	Metadata Creation Management Metadata Harvesting Management Metadata Model Management Metadata Appraisal and Selection Management
	Metadata Quality Management	Metadata Quality Control Metadata Quality Improvement
	Metadata Preservation and Dissemination	Metadata Ingest Management Metadata Versioning Management Metadata Integration Management Metadata Access Management
	Metadata Curation Monitoring and Controlling	Metadata Curation Monitoring Recruiting and Engagement Management Communication and Feedback Management Metadata Curation Coordination
	Metadata Curation Platform Administration	Metadata Curation Platform Design Metadata Curation Platform Implementation Metadata Curation Platform Maintenance

is the version presented in this chapter.

5.4 OVERVIEW OF THE LOUVRE FRAMEWORK ELEMENTS

Figure 20 lists all actors and roles, agile practices, dimensions, processes, and deployment method. For each process, Figure 20 also presents the total amount of input, activities, outcomes and work products. In the following sections, we present a summary of the main elements (actors and roles, agile practices, dimensions and processes) of the Louvre. The whole Louvre framework is presented in details in Appendix C.

5.4.1 Actors and Roles

Outlines how certain actors may be organized in order to achieve the metadata curation goals. It also defines roles and responsibilities. In the following, the concepts of contributors and team and the list of roles related to metadata curation work will be described.

5.4.1.1 Contributors and Teams

The people part of metadata curation involves actors who curate metadata, which they will make available to contributors and non-contributors alike (from here, all the actors who contribute to the curation of metadata will be called contributors). However, many contributors do not have sufficient skills to perform the whole metadata curation work (ZUIDERWIJK et al., 2012). Meanwhile, the contributors have varying areas of expertise. They also share similar goals related to metadata curation. With their diverse set of skills, the contributors should be able to curate metadata as a group. Hence, teams and the characteristics of teams are central to metadata curation in Data Ecosystems.

In this sense, the Louvre framework supports team-building and teamwork. Contributors should work in collaboration with each other to curate metadata in Data Ecosystems. Several agile approaches emphasize that collaboration is a key to success for agile project delivery. In fact, two of the four values in the Agile Manifesto highlight the emphasis on strong collaboration. One of the most expected benefits from team work is that contributors may learn from one another and mentor one another. Meanwhile, shared disparate sets of skill may unlock the power of a team to tackle heterogeneous work.

Due to distributed nature of Data Ecosystems, a metadata curation team should possess the ability to self-organize. The Agile Manifesto recognizes self-organizing teams as an essential principle, stating that “the best architectures, requirements, and designs emerge from self-organizing teams”. The best people for planning work are the ones who are going to do it. In metadata curation context, a curator should be respected as intelligent people who can determine their strategies for working together (COCKBURN, 2002). When given a bit of guidance, they can plan their work within established parameters, such as goals and policies.

As part of deciding the best way to curate metadata, some teams will decide that one person on the team will make all key technical decisions. Other teams may choose to

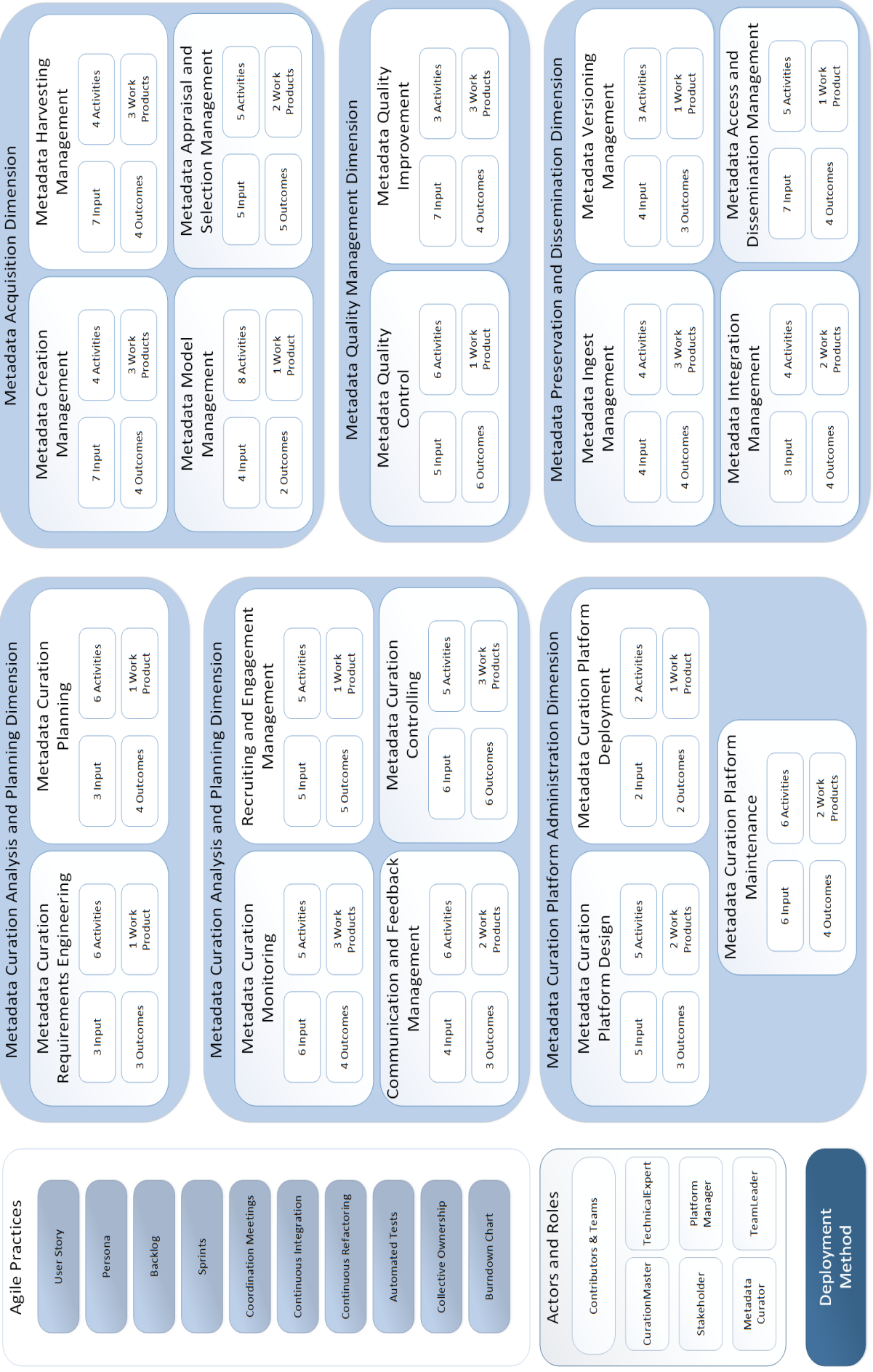

Figure 20 – Louvre Framework Elements. Source: Author

share the responsibility with multiple contributors. There isn't some optimal organization. Rather, the self-organization aims to encourage the team to own the problem of performing its work fully.

Moreover, for a metadata curation initiative to stay agile, teams must be formed and disbanded quickly. Hence, it is important to encourage teams and contributors to meet each other, share information transparently, and move from team to team depending on the goal to be addressed (AMBLER; LINES, 2011).

5.4.1.2 Roles

According to Welser et al. (2011), roles clearly defined makes coordination easier (though non-trivial). Inspired by Scrum and Disciplined Agile Delivery frameworks, the Louvre deemphasizes roles based strictly on skill sets in favor of primary roles that can include a variety of skills. Accordingly, the three primary roles are:

- CurationMaster: It represents an authority to exist beyond the potential crowd of curators who organizes whole metadata curation work providing a way for any actor to contribute to the common effort (BUDHATHOKI; HAYTHORNTHWAITE, 2013). The CurationMaster identifies and explains the requirements, prioritizes requirements, provides feedback on the curatorial strategy.

- Stakeholder: It represents someone who is materially impacted by the outcome of the metadata curation work. A stakeholder could be any Data Ecosystem actor who directly or indirectly consumes curated metadata. Stakeholder represents the needs and desires of the Data Ecosystem actors regarding metadata and curation needs. As such, a stakeholder clarifies any details regarding the metadata and is also responsible for maintaining a prioritized list of metadata curation tasks that curators will deliver. However, different from software development projects, often it is complex to identify one stakeholder to speak on behalf of a wide range of stakeholders. Moreover, it isn't reasonable to expect a stakeholder to be experts in every nuance in Data Ecosystem domain. Therefore, the Louvre framework explicitly recommends defining a set of stakeholder to represent the whole community.

- MetadataCurator: It represents contributors who focus on performing the actual curation of metadata. MetadataCurators will carry out all tasks required to curate the metadata, including acquisition, appraisal and selection, quality assurance, preservation, dissemination, etc. Note that not every MetadataCurator will have every single one of these skills. But, they will have a subset of required skills and may strive to gain more skills over time.

The Louvre framework also recognizes three secondary roles, which are typically introduced, often on a temporary basis, to address scaling issues. Accordingly, the three secondary roles are:

- TechnicalExpert: Since MetadataCurators should perform tasks in one or more disciplines, this doesn't imply that every MetadataCurator needs to be an expert at everything. Sometimes a team needs the help of a TechnicalExpert to overcome a difficult problem and to transfer their skills to one or more contributors (AMBLER; LINES, 2011).

- PlatformManager: It represents the authority that provides the services and functions for the overall creation and operation of the metadata curation platform. The platform used to support metadata curation tasks is a key source for successful curation and someone needs to be responsible for mitigates this risk. PlatformManager is responsible for planning, designing and maintaining the metadata curation platform and its related infrastructure. It is also responsible for establishing and maintaining preservation standards and policies, and providing curators support.

- TeamLeader: Often, a team needs an informal leader, called a TeamLeader, who emerges from the team. A TeamLeader is a kind of metadata curation coach, helping to keep the team focused on curation activities work items and fulfilling their iteration goals and commitments that they have made to the Stakeholders (AMBLER; LINES, 2011). A TeamLeader has some additional responsibilities such as being (i) the teacher of how metadata curation should be performed, (ii) monitoring the long-term health of curated metadata, (iii) organizing a team, (iv) organizing design workshops, and (v) reviewing curation work done.

Service providers, application developers, regulatory authorities and other roles may be involved in metadata curation. However, it is beyond the scope of Louvre to identify and defines all possible type of contributors, and all the roles that would apply.

5.4.2 Agile Practices

The Agile Practices provide a set of best practices that enable, encourage and guide contributors in the iterative, incremental and self-organizing curation of metadata. Since the publication of Agile Manifesto, a myriad of practices aligned with its values and principles has emerged. Agile practices help accomplish agile principles in a method (LEE; YONG, 2013). In fact, the agile methods are built on a set of predefined agile practices. Common examples of agile practices are daily stand-ups meetings, feature-driven development or pair programming. Diebold (2015) identified 155 agile practices.

Even though the benefits of agile practices are worthwhile (MONIRUZZAMAN; HOSSAIN, 2013; GLAIEL; MOULTON; MADNICK, 2014), often such practices require adjustments and

tailoring for context adequacy in both organization and project levels (LEE; YONG, 2013). According to Conboy e Fitzgerald (2010), tailoring can be supported and stimulated by the adopted software method or executed based on the enterprise needs. Hence, in the light of values and principles presented in Section 5.3.1, the Louvre adapts and tailors agile best practices from a variety of agile methods and software process frameworks to suit the metadata curation context. In the previous sections, some agile practices adapted to the context of metadata curation have already been presented. Among them, we highlight the iterative and incremental process, teamwork and self-organizing teams.

Besides that, the Louvre also adapts additional ten practices and provides advice for when and how to apply these practices. In summary, these ten practices focus on simplicity, coordination, flexibility, early solution delivery, and empowerment of contributors. The ten practices are presented in details in the following.

5.4.2.1 User Story

A User Story is a technique widely used by several agile methods to capture system requirements (COCKBURN, 2002). Typically, a user story is an informal, natural language description of one or more features of a software system. User stories are more commonly used as a checklist to define whether a project has delivered its objective (COCKBURN, 2002). User stories are often written from the perspective of an end user or a user of a system (COCKBURN, 2002). Each user story is expected to yield, once implemented, a contribution to the value of the overall product.

Metadata curation is not a software project, but the simplicity of user stories lends itself to the articulation of metadata curation requirements. The informal language may contribute to create a shared understanding with those contributors not familiar with metadata curation and metadata management concepts. The "content" of user stories may cover at least the following:

- Metadata policies and guidelines;
- Metadata requirements;
- Metadata standards;
- Metadata curation processes and controls;
- Metadata repositories and metadata curation systems.

The role-feature-reason template is another agile practice commonly recommended to help curation teams and product owners starting to write user stories. This template advocates paying attention not just to "what" the desired product is to do, but also "for whom" it does it and in pursuit of "what objective" (AGILE ALLIANCE, 2018).

Therefore, the role-feature-reason template recommends users stories to be written as statements in the form of "*As a [role], I want [goal/desire/feature] so that [benefit/reason]*" can help to explain a metadata curation procedure, a platform improvement, or even a quality expectation for the metadata (AGILE ALLIANCE, 2018). Some examples are:

- As a consumer, I need to know that the metadata provided to me is accurate so my interactions with the Data Ecosystem resource are as efficient as possible.
- As a consumer, I need to know that the metadata I receive is correct, so I have the confidence I am making the right decisions.
- As a MetadataCurator, I need to be able to trust metadata repository, as I rely on the repository to preserve metadata created.

In metadata curation context, user stories should also be written from the viewpoint of "useful deliverables" that make sense to contributors. Thus, it is not necessary to elucidate the "how" a requirement should be implemented. Moreover, user stories should have a single goal, desire or feature. User stories with multiple goals, desires or features can make it difficult to understand, estimate and prioritize the original requirements.

5.4.2.2 Persona

A Persona is a detailed, synthetic biography of a fictitious user. In other words, a persona defines an archetype user of a system, *i.e.*, an example of the kind of user who would interact with the system (COOPER; REIMANN; CRONIN, 2003). According to Pichler (2013), persona is a powerful technique to describe the users and customers of a product in order to make the better design decisions.

If end users are unknown, personas can help to write user stories. A persona must have the same characteristics of the user and should be based on knowledge gained from direct interaction with the target customers and users (PICHLER, 2013). Personas should not be confused with roles, which are primarily defined regarding tasks or job descriptions. The persona practice emphasizes goals and behaviors.

In the Louvre, personas are useful when both MetadataCurators and CurationMaster don't have easy access to Stakeholders, helping to guide your decisions about metadata needs and curation requirements. Questions like "How would Marcelo use this feature" or "Would Bernadette even be interested in this?" can start great conversations within a curation team, getting to think the way that users actually would. In short, personas are one of a range of modeling techniques to acquire and analyze requirements.

5.4.2.3 Backlog

Scrum's central point, product backlog is a list of all the work necessary for the product. A Product Backlog lists all requirements, features, improvements, and fixes that constitute

the changes to be made to the product in future releases (SCRUM ALLIANCE, 2016). Common formats for structuring backlog items include user stories, use cases, or any other requirements format that the development team finds useful. Product backlog does not need to be complete at the beginning of a project. Over time, product backlog grows and changes as development team learn more about the product and its users (SCRUM ALLIANCE, 2016).

The product backlog is what will be delivered, ordered into the sequence in which it should be delivered (SCRUM ALLIANCE, 2016). It is visible to everyone. The ProductOwner prioritizes product backlog items based on considerations such as risk, business value, dependencies, size and date required (SCRUM ALLIANCE, 2016). The ProductOwner gathers input and takes feedback from, and is lobbied by, many people, but ultimately makes the call on what gets built.

In metadata curation context, the backlog can be used to coordinate contributors and teams to perform tasks. In this sense, a curation backlog lists technical work related to metadata curation. Common examples are:

- create metadata related to a specific resource;
- evaluate the quality of a metadata item;
- implement or improve a system, service or tool;
- ensure regulatory compliance.

The backlog items reflect tasks related to Louvre's dimensions and processes. Moreover, curation backlog items could be prioritized based on how desirable a task is to achieve metadata curation goals. MetadataCurators can then browse curation backlog to select the suitable tasks based on their skills and availability. In its turn, a curation team can also select a set of tasks determines which they can complete. In this sense, a curation backlog enables flexible coordination of metadata curation work.

The curation backlog can be combined with a task board (AGILE ALLIANCE, 2018) to ensure the efficient diffusion of information relevant to the whole curation team. Typically, a task board is divided into three columns labeled "To Do", "In Progress" and "Done".

5.4.2.4 Sprints

In Scrum methodology, the development process is divided into regular cycles over time. Each one of these cycles are called Sprint. Sprints are used to achieve some defined goals. During each Sprint a set of requirements is implemented, resulting in an increment of the product being developed. Each Sprint has a goal of what is to be built, a design and flexible plan that will guide building it, the work, and the resultant product increment (SCHWABER; SUTHERLAND, 2013). Each Sprint have a time defined, meaning that the

schedule for an interaction must be considered fixed and the scope of the iteration content is actively controlled to respect the schedule.

In metadata curation context, a sprint starts from the creation of the curation backlog. Thus, each sprint would cover all the curatorial tasks related to a set of metadata. Thus, the metadata curation will be developed and delivered in small interactions, with a constant evolution of the processes or services of metadata curation, so that the curation initiative can obtain a quick return of curation benefits and minimize risks.

Base on Scrum method, the Louvre recommends that the first iterations should not be longer than two weeks. Although some metadata curatorial activities demands more effort, such as developing or deploying the metadata curation platform, strict metadata curative tasks can be accomplished within a shorter interval. For inexperienced teams, the interaction should be short enough for the team to learn how to perform metadata curation processes and procedures. In these cases, inexperienced teams should also select a smaller number of metadata items to be curated. With less curatorial tasks, they can learn and gain experience in relation to metadata curation strategy, policies and procedures in a faster manner.

Although the duration of a sprint is not a static and rigid rule, the literature have reported many experiences that, in too long iterations, the team tends to be relaxed in the initial phase and to become overloaded in the final phase of the interaction (BERCZUK, 2007). Therefore, it is important to pursuit short iterations.

5.4.2.5 Coordination Meetings

A coordination meeting is one of the most commonly practiced technique in software developments methods. These meetings present opportunity for a development team to get together on a regular basis to coordinate their activities. In several agile methods, it is recommended to schedule meetings in a short time basis. For instance, Scrum method recommends to perform daily meetings. Each day at the same time, the team meets so as to bring everyone up to date on the information that is vital for coordination. These coordination meeting is normally timeboxed to a maximum duration of 15 minutes, though this may need adjusting for larger teams.

In distributed environments like Data Ecosystems, it is a challenge to arrange daily meetings in which all curation team members are able to attend. Hence, the Louvre framework recommends to schedule coordination meetings at particular points in the sprint cycle. It's important that a coordination meeting focus more on status updates than problem solving. Any impediments that are raised in the meeting would become team's responsibility. And so, impediments can be resolved in the current or further sprints.

5.4.2.6 Continuous Integration

In software engineering, continuous integration is a practice that recommends developers to integrate their work frequently, and testing the modifications, as early and often as possible. Ideally, developers should integrate their work one or multiples times a day. According to Fowler (2006), the main benefit expected from continuous integration is to prevent integration problems. Hence, continuous integration seeks to minimize the duration and effort required by each integration episode as well as being able to deliver a product version suitable for release at any moment. In practice, continuous integration is achieved through version control tools, team policies and conventions, and tools specifically designed to help achieve continuous integration (AGILE ALLIANCE, 2018).

In Louvre, it is recommended that all metadata acquired should be as soon as possible transferred/ingested to a metadata repository. Once ingested in the repository, the metadata must be evaluated against compliance with quality policies, standards and other regulations. These verification procedures require a sort of automated test mechanisms to test the common problems related to quality and compliance assurance. The idea of continuous integration is to find issues quickly; giving each MetadataCurator feedback on their work and automated mechanisms evaluate that work quickly. The continuous integration practice may delivery some benefits. For instance, it would allow other MetadataCurators become aware of work done as well as to contribute for this work. Thus, continuous integrated curated metadata creates a channel for contribution. Moreover, more robust the test suite results in greater confidence that metadata will address the expected quality requirements.

In this context, the great advantage of continuous integration lies in the fast feedback. Each integration allows communication to contributors and possible inconsistencies or defects can be detected more quickly. In this way, curation teams or contributors responsible for a task would become aware of possible problems, thus verifying and correcting them faster. Continuous integration is yet another way to bring security to collaborative metadata curation.

However, unlike software projects that handle with code files that can be deposited in version control software, not all metadata curation tasks result in the creation or change of metadata to be ingested in a metadata repository, such as the development of metadata harvesting mechanisms. For these tasks, the intermediary results can be shared through agile communication practices (*e.g.*, forums and task boards), thus avoiding redundant tasks and allowing all contributors to have a complete view of the progress of the metadata curation work of Data Ecosystem.

5.4.2.7 Continuous Refactoring

Continuous Refactoring is a controlled technique for restructuring the internal structure of the existing program's source code without changing its external behavior (FOWLER et al., 1999). Refactoring improves nonfunctional attributes of the software. In other words, refactoring focuses exclusively on getting the code into a desirable state for the development team. Advantages include improving readability of code and reducing its complexity, which as a consequence can improve source-code maintainability (FOWLER et al., 1999). Another expected benefit of refactoring is encouraging each developer to think about and understand design decisions (AGILE ALLIANCE, 2018).

In Louvre, continuous refactoring practice is viewed as a refactoring metaphor proposed by (LUNA, 2009). In that sense, refactoring should be seen as a restructuring and possible optimization of metadata curation work. In this way, the refactoring refers to:

- people: how to work in a team and the integration and commitment of the contributors;
- processes: reviewing of procedures, techniques and methods to maximize the results of curatorial tasks;
- tools and solutions: identification, selection or development of tools and solutions to improve results and reduce manual efforts to perform curation tasks.

For example, planned metadata curation actions that are not being properly performed or resulting in a poor performance should be reviewed as soon as the problem is identified. Corrective or preventive actions should be taken in order to achieve the expected results. Another example of refactoring is improvemments on metadata curation platform. The platform architecture may degrade over time as new services and functionalities are incorporated. To prevent the platform from becoming complex to maintain, the PlatformManager can continually refactor the platform looking to modularize its structure and improve infrastructure. However, such changes do not change the behavior of the platform; only improve its maintenance and operation.

5.4.2.8 Automated Tests

In software development, an automated test is a specialized software to control the execution of tests and to verify an expected behavior of a software product (JANZEN; SAIEDIAN, 2005). Such practice automates some repetitive but necessary tasks in a formalized testing process already in place, or performs additional testing that would be difficult to do manually. Test automation is critical for continuous delivery and continuous testing.

Some different approaches have been proposed for automating tests, including acceptance tests and unit tests. Usually, automated tests have a binary result, pass or fail.

A failure suggests, though does not prove, the presence of a defect in the product. An expected benefit from automated unit tests is a decrease in defect rates.

In Louvre, an automated test suite enables curation teams to verify some quality of metadata safely. Moreover, such a practice can continuous assure quality. In other words, there are ongoing and periodic quality assessments performed by automated tests. In the presence of failure, a new metadata quality improvement task can be registered in curation backlog. It is particularly important given that multiples contributors curate metadata. Like automated testing for other things, it requires skill and tooling to implement. To effectively test metadata in an automated manner it is necessary to include those tests in a continuous integration approach.

However, there are some quality aspects which are able to be automatically assessed. For instance, the evaluation of descriptive metadata requires human checking. For these cases, the quality control would be performed manually by contributors.

5.4.2.9 Collective Ownership

In software development projects, development teams typically adopt conventions governing who is allowed to modify some source code that was originally developed or created by another, often referred to as kind of ownership (AGILE ALLIANCE, 2018). These conventions can be either formalized or entirely implicit among development team members. There are many different models of ownership. Very often only one developer owns each code file (AGILE ALLIANCE, 2018). Collective ownership is a explicit convention that any team member is allowed to make changes to any code file as necessary. These changes can be viewed is a positive duty, and are triggered either to complete a development task, to repair a defect, or even to improve the code's overall structure (AGILE ALLIANCE, 2018).

The Louvre framework recommends that metadata should be owned by all the contributors. Instead of creating an environment where everyone is responsible for their own piece of metadata, any contributor can be responsible for all metadata curated. Everyone has access and authorization to access, edit, and enhance any item in the repository at any time. Ownership is collective and everyone is equally responsible to all parties. With this, the overall contributions save time, since it does not have to wait for the authorization of another participant to edit an item. Collective ownership can also reduce the risk that the absence (or unavailability) of any contributor will stall or slow work. This practice also is a favorable factor in the diffusion of technical knowledge as well as encourages each developer to feel responsible for the quality of the whole (AGILE ALLIANCE, 2018).

However, this practice does not eliminate the need for boundaries between roles and responsibilities, or the need for coordination. In particular, access management is still essential to prevent unauthorized participants, as well as to allow accountability.

5.4.2.10 Burndown Chart

Burndown chart is a visual measurement tool to measure the development team progress. It shows the completed work per day against the projected rate of completion for the current project release. That is, it shows the total effort against the amount of work delivered in each iteration. A burndown chart is calculated from the sum of estimated backlog items. If a backlog item is set to "closed", it counts for the burndown.

Typically, project managers or team leaders create these charts and share them daily with their team. Among its purposes, a burndown chart allows verifying if a project is on track to deliver the expected solution within the desired schedule. Such charts are also useful for predicting when all the work will be completed. To some extent, burndown charts can force teams to constantly evaluate their performance on a daily basis and prioritize work as needed, which helps maintain the accuracy of development iterations and project backlogs. In addition, project managers save time by using software to create the charts, rather than creating them manually automatically.

Burndown charts can be applied to metadata curation to measure progress over time. Thus, the curation work can be represented regarding either time or curation backlog items. The burndown charts should be published in a portal or Web site. MetadataCurators or any contributors also save time by tracking progress visually, instead of sorting through email, tasks and documents for status updates. Besides, CurationMaster can plan more effectively, especially when teams have to add or drop members, by consulting chart data.

Burndown charts can be combined with another information radiator (*i.e.*, generic term for any of a number of handwritten, drawn, printed or electronic displays which presents useful information at a glance (AGILE ALLIANCE, 2018)). In particular, the CurationMaster should propose a set of metrics to guide and inspire the contributors in the fulfillment of their goals. It is recommended to automate metrics collection as much as possible as well as to minimize the number of metrics collected. Make the measures visible, independently verifiable, review often. The right metrics can motivate, focus and build confidence.

5.4.3 Dimensions and Processes

Besides principles and values presented in Section 5.3.1, the Louvre framework should ensure that all of the necessary stages in the metadata lifecycle are covered. In consequence, the overall curation of metadata is conceived regarding an initial set of six dimensions; each one contributes to one or more stages of metadata lifecycle.

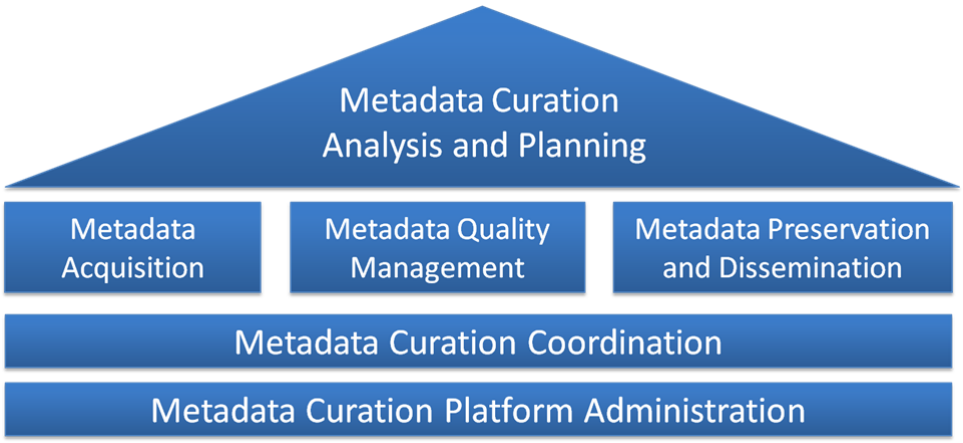

Figure 21 – Louvre Dimensions. Source: Author

The scope of each dimension shown in Figure 21 is:

- Metadata Curation Analysis and Planning: Provides the basis for the exercise of planning, monitoring the metadata curation initiative.
- Metadata Acquisition: Provides the basis for planning, control and support the acquisition and selection of metadata.
- Metadata Quality Management: Provides the basis for planning, implementation and control the quality assurance and improvement of metadata.
- Metadata Preservation and Dissemination: Provides the basis for preserving metadata and keeping them discoverable and accessible to Data Ecosystem actors.
- Metadata Curation Coordination: Provides the basis for monitoring, enabling, promoting and coordinating collaborative efforts from contributors.
- Metadata Curation Platform Administration: Provides the basis for planning, development, and management of a metadata curation platform.

The Metadata Curation Analysis and Planning dimension can be classified as strategic because its processes support the requirements elicitation as well as the planning on how the metadata curation work should be done. On the other hand, Metadata Acquisition, Metadata Quality Management, and Metadata Preservation and Dissemination dimensions are more systematic as they define processes related to the metadata life cycle. Finally, the Metadata Curation Coordination and Metadata Curation Platform Administration dimensions are orthogonal dimensions. The first one monitors activities that should align and help achieve the goals established in the Metadata Curation Analysis and Planning dimension. The latter, in turn, enables the set of tools that will be used as the basis for the execution of the metadata curation work, including the coordination of the actors.

Table 23 – Louvre Dimensions and Processes

Dimensions	Processes
Metadata Curation Analysis and Planning	Metadata Curation Requirements Engineering Metadata Curation Planning
Metadata Acquisition	Metadata Creation Management Metadata Harvesting Management Metadata Model Management Metadata Appraisal and Selection Management
Metadata Quality Management	Metadata Quality Control Metadata Quality Improvement
Metadata Preservation and Dissemination	Metadata Ingest Management Metadata Versioning Management Metadata Integration Management Metadata Access and Dissemination Management
Metadata Curation Coordination	Metadata Curation Monitoring Recruiting and Engagement Management Communication and Feedback Management Metadata Curation Controlling
Metadata Curation Platform Administration	Metadata Curation Platform Design Metadata Curation Platform Deployment Metadata Curation Platform Maintenance

Table 23 relates the dimensions and their respective processes. The whole set of dimensions and processes are presented in details in Appendix C. The following subsections focus on presenting a summary of each dimension and their processes. For each process is presented its purpose. The rest of elements is presented in Appendix C.

5.4.3.1 Metadata Curation Analysis and Planning Dimension

The Metadata Curation Analysis and Planning Dimension consists of those processes performed to define and refine the goals and requirements, and develop the course of action for curating metadata in Data Ecosystems.

The *Metadata Curation Requirements Engineering* process recommends activities for elicitation (collecting, creating), analysis (aligning, prioritizing), and validation (monitoring, enforcing) of requirements involving metadata (*e.g.*, business rules, metadata ownership, metadata classification, metadata quality, metadata usage, metadata access, authentication, entitlements, etc.). A requirement is a specification of constraints, demands, necessities or needs that a product, system or business must satisfy. In software engineering, sets of requirements are used to capture the information needed to design, build and test software. In metadata curation context, requirements are used to capture the needs related to the curation process itself and expectations about metadata.

In summary, metadata curation requirements can be classified into two categories: curation requirements and metadata requirements. The Curation Requirements represents a set of requirements that will determine what curatorial activities should be performed and how MetadataCurators should perform them. In an analogy to the functional require-

ments of software engineering, a curation requirement defines the operation or behavior of the metadata curation work. Indeed, this category of requirements determines what contributors have to do in order to curate metadata. In its turn, Metadata Requirements specify the set of metadata that should be curated, that is, such requirements define the inputs of metadata curation activities. For example, these requirements define what metadata need to be created, what metadata should have their quality assessed, what metadata needs to be preserved, and other activities.

The *Metadata Curation Planning* process recommends activities for establishing the basis for creating and maintain a metadata curation action plan that aligns with the requirements with a strategy and a set of policies, standards and procedures. The goal of a metadata curation planning is to consider the many aspects of metadata curation, including metadata curation activities, the types of metadata to be curated, the standards that would be applied, provisions for archiving and preservation, quality policies, and etc. That is, such planning should ensure that metadata are well-managed in the present and prepared for preservation in the future. In particular, some domains had proposed a myriad of standards to structure and organize metadata. These standards should be used in order to accomplish interoperability requirements.

5.4.3.2 Metadata Acquisition Dimension

The Metadata Acquisition Dimension consists of those processes required for creating, harvesting and selecting metadata. In particular, 'create' refers to original metadata generated and recorded by participants, and 'harvest' refers to pre-existing metadata collected from other sources. This dimension also includes activities to allow the selection and rejection of metadata that does not meet specified requirements and policies.

The *Metadata Creation Management* process recommends activities for creating appropriate metadata. The metadata produced by human beings is probably what most people assume when they think of a metadata source. In this process, the metadata will be created manually by the MetadataCurators.

The *Metadata Harvesting Management* process recommends activities for harvesting appropriate metadata. Not all metadata should be created from scratch. Indeed, such a notion will be unworkable. Harvesting is the process of collecting metadata from a remote or external source.

The *Metadata Model Management* process recommends activities for designing, developing and refining of a metadata model, which is the overall structure for the metadata. Such model should be derived from meta-model presented in Chapter 4. It should be an integrated, subject-oriented set of specifications defining the essential metadata produced and consumed across the Data Ecosystem. While integrated means that the concepts in the model fit together, subject-oriented means the model is divided into commonly recognized subject areas that span across different aspects related to Data Ecosystem. In

particular, subject area models are used to address specific elements needed by a business or management activity. As mentioned in Section 5.4.3.1, there are a plenty of standards defining and structuring metadata for specific domains. These standards should be used whenever possible.

The *Metadata Appraisal and Selection Management* process recommends activities for evaluating metadata and selecting for long-term curation and preservation. An appraisal is "the process of evaluating records to determine which are to be retained as archives, which are to be kept for specified periods and which are to be destroyed" (HIGGINS, 2008). Selection is a more general term, usually applied when deciding what will be added to a repository. Appraisal and selection of metadata are critical because of the limited resources that most Data Ecosystems dedicate for preservation. Given the large (and growing) volume of metadata, it is a practical necessity to choose only the most important for long-term management.

5.4.3.3 Metadata Quality Management Dimension

The Metadata Quality Management Dimension consists of those processes required for detecting and preventing inconsistencies and defects in metadata as well as for improving the quality of metadata by cleaning such defects. The quality assurance aims to ensure that curated metadata remains authentic, reliable, usable, accessible to, and understandable over the long term.

The *Metadata Quality Control* process recommends activities for measuring, assessing and ensuring the quality of metadata. Metadata Quality control is a set of procedures intended to ensure that the metadata adheres to a defined set of quality criteria, thus meeting the curation requirements. In its turn, The *Metadata Quality Improvement* process recommends activities for planning, implementation and control activities that apply quality management methods to improve and refine the fitness for the use of metadata.

5.4.3.4 Metadata Preservation and Dissemination Dimension

The Metadata Preservation and Dissemination Dimension consists of those processes required for preserving metadata and ensuring that metadata is discoverable and accessible to both contributors and non-contributors. It includes processes to guide how to integrate metadata, transfer metadata to an appropriate repository and securely store them adhering to relevant standards. In addition, it also recommends processes to guide how to make metadata accessible by displaying publicly or by exposing them to other systems.

The *Metadata Ingest Management* process recommends activities for transferring metadata to an appropriate repository for permanent (long-term) storage while maintaining and verifying the integrity of metadata.

The *Metadata Versioning Management* process recommends activities for managing metadata version. Metadata are not immutable objects; rather they are subject to changes.

Moreover, several contributors may be involved in the creation of the same metadata content. Hence, it is important to identify the different version of metadata. Version management is the ability to manage the change metadata. The versioning process should also be clear and transparent to end users, so they always know which version of the metadata and data they have acquired.

The *Metadata Integration Management* process recommends activities for combining metadata and presenting them in a unified way. Often metadata accumulates in metadata silos, which refers to metadata kept separated from other related metadata. This leads to many problems ranging from inability to access the right metadata to unnecessary replication of metadata. Metadata integration consists of a set of techniques used for connecting silos of metadata. Thus, the integration of metadata enables to gain a more comprehensive understanding from metadata.

The *Metadata Access and Dissemination Management* process recommends activities for providing access to metadata. In order to be used, metadata must be accessible for both humans and machines. This process includes activities for creating the means for enabling machine-based search and retrieval functionality that help a user identifies what metadata exist, where the metadata are located, and how can they be accessed (*e.g.*, download).

5.4.3.5 Metadata Curation Coordination Dimension

The Metadata Curation Coordination Dimension consists of those processes required to promote, engage, monitor and control the contributors efforts toward achieving common and recognized metadata curation goals. According to Kraut e Streeter (1995), coordination can be defined as "the integration or linking together of different parts of an organization to accomplish a collective set of tasks". In Louvre, this means that different contributors agree on a common definition of metadata curation, share information, and work together to achieve their goals. They must have a common view of what metadata should be curated, how metadata curation should be organized, and how it should fit with other activities already in place or undergoing parallel in Data Ecosystems. To efficiently curate metadata , they must not only monitor and control the work being done but also share detailed information about the progress of curation activities. Furthermore, the contributors must coordinate their work so that it gets done and fits together.

The *Metadata Curation Monitoring* process recommends activities for monitoring metadata curation work. Monitoring allows results, processes, and experiences to be documented and used as a basis to steer decision-making and learning processes. Monitoring is checking progress against plans, as well as it reveals mistakes and offers paths for learning and improvement. Monitoring process needs to be planned and carried out on a regular basis throughout metadata curation initiative. Such regular monitoring keeps contributors up to date on the curation progress as well as verify the adherence to the curation plan.

The *Recruiting and Engagement Management* process recommends activities for recruiting contributors and promoting engagement. A collaborative metadata curation initiative relies on a constant stream of curation activity, with multiple contributors contributing in various ways, at various stages. A pool of volunteer contributors will perform metadata curation work. Without contributors to occupy necessary roles, the metadata curation initiative would cease to function. Therefore, it is necessary to motivate, engage and retain new contributors to promote a sustainable curation initiative. High engagement means that contributors care about their work, feel like they're part of a community, are brought into the greater vision, and bring their unique strengths to their work (HASSELL, 2018). Recruiting of contributors should be seen as a continuous process that reaches the widest possible range of participants.

The *Communication and Feedback Management* process recommends activities for planning and maintain an effective communication flow between contributors as well as ensuring the ultimate disposition of metadata curation information. Contributors spend part of their time communicating with other contributors, whether they are or not working the same metadata curation tasks. Effective communication creates a bridge between diverse contributors who may have different cultural references, different skill, perspectives and expectations. All of them impact or have an influence upon the metadata curation execution or outcome. In particular, this process should promote feedback that describes a circular process where the output of an activity, a process or a system is returned as input in order to regulate or influence a further output. It helps companies and institutions to capture user/customers concerns and analyze the data for future developments. In a metadata curation environment, feedback allows contributors to become indirectly involved in the curation work. Hence, collecting and analyzing feedback is an essential step towards improving metadata and their curation tasks.

The *Metadata Curation Controlling* recommends activities for organizing, orchestrating, and leading the metadata curation work. As mentioned before, the metadata curation work will be performed by a set of contributors with assigned roles and responsibilities. Contributors may have varied skill sets, may be assigned full or part-time, and may enter or leave the initiative at any moment. This process also focuses on continuous communication with contributors to understand their needs and expectations, addressing issues as they occur, managing conflicting interests and fostering appropriate engagement in curation activities.

5.4.3.6 Metadata Curation Platform Administration Dimension

The Metadata Curation Platform Administration Dimension consists of those processes for the design, implementation/deployment, and maintenance of the metadata curation platform responsible for support the metadata curation work. Such a platform should enable standardization and integration of metadata curation tasks. Proper implementation

of a platform for supporting metadata curation environment is not just about the tools to archive metadata. Rather, it is about creating a strategy to plan, design, and construct a platform capable of addressing curation requirements and allowing supporting the metadata curation strategy. Thus, this process also recommends activities to planning and designing an integrated and holistic platform to support the metadata curation work.

The *Metadata Curation Platform Design* process recommends activities for the strategic and technical design and planning of metadata curation platform architecture. Metadata curation platform architecture is an integrated set of high-level structures that govern and define how metadata is used, stored, managed and integrated within a Data Ecosystem. Each structure comprises software elements (*i.e.*, tools, systems or applications used to curate metadata), relations among them, and properties of both elements and relations. It functions as a blueprint for the metadata curation platform, laying out the structures to be developed and maintained by PlatformManager. Moreover, such architecture provides an understanding of creating and managing the flow of metadata and how it is processed across curation systems and applications.

The *Metadata Curation Platform Deployment* process recommends activities for establishing the metadata curation platform to support metadata curation work. This process defines or develops and deploys the facilities, tools, and communications and information technology assets needed for metadata curation work with respect to the architecture designed in the Metadata Curation Platform Design process.

The *Metadata Curation Platform Maintenance* process recommends activities for the day-to-day technical supervision of the metadata curation platform. The metadata curation platform must to meet necessary conditions throughout the entire metadata lifecycle. Such management does not have one objective. There are many, including performance, efficiency, security, and privacy. The management of metadata curation environment comprises a number of proactive techniques including performance monitoring and tuning, storage and capacity planning, backup and recovery.

5.4.4 Deployment Method

In spite of the promising benefits of curating metadata in a Data Ecosystem, starting a metadata curation initiative is a challenging task. The barriers include the natural resistance to new processes, lack of resources, need for coordination and involvement, among others. In summary, it demands a great deal of adjustment from all the contributors. Thus, there is a need for specific guidelines and methods to support systematic selection, deployment, and tailoring of metadata curation processes to fit a Data Ecosystem context.

The Louvre employs the iterative and incremental development approach of agile software development methods by proposing that the metadata curation initiative is based on a cycle of iterations. Each iteration generates value to the curatorial work in the

form of new metadata curation procedures or services planned, implemented, verified and evaluated within a PDCA cycle, as presented in Figure 22.

The PDCA method is used by organizations to manage and improve their internal processes using information as a decision-making factor in order to ensure the fulfillment of established goals (FONSECA; MIYAKE, 2006). Figure 22 shows a graphical representation of the phases of the PDCA. The PLAN phase aims to define the ideal goals of the process under management as well as establishing the methods for their achievement. The DO phase comprises the execution of actions planned. The CHECK phase aims to verify the effectiveness and efficiency of actions planned by comparing the execution against the planning. In this phase, it can be noted whether the expected results were or were not achieved. The ACT phase implies corrective actions. From the results achieved, two different paths can be followed. If the verification indicates that it was not possible to achieve the expected results, corrective actions should be carried out and then start a new iteration; but if the expected results were achieved, then the process should be standardized, thus ensuring its continuity (FONSECA; MIYAKE, 2006).

In general, it can be seen that the PDCA can be used to either maintain or improve the actions planned for a process. In metadata curation context, this allows the CurationMaster to gradually implement a metadata curation process. In each iteration, a new and improved curation strategy can be implemented. This incremental and iterative approach both allows for adaptation to the Data Ecosystem context and allows the benefits of metadata curation to be achieved more quickly. In addition, the PDCA method allows the CurationMaster to continually look for optimizing the metadata curation work by improving their policies and procedures as they learn from their experiences.

Different from traditional digital curation models, which typically focus on the creation and preservation aspects of the metadata lifecycle, this approach involves all curatorial work from the starting point of the curation initiative to acquisition and quality assurance of metadata and improvement of curation processes.

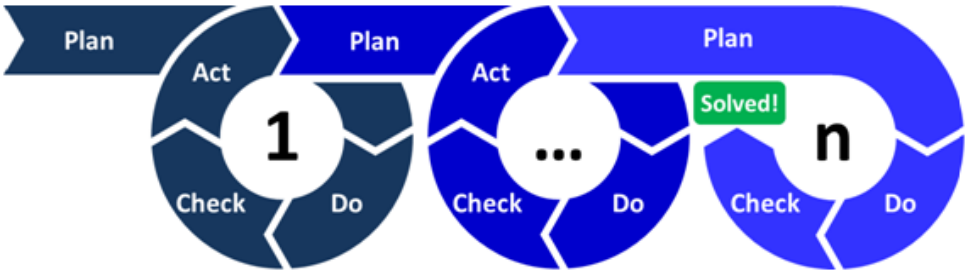

Figure 22 – Multiple iterative loops of a PDCA method. Source:(ROSER, 2016)

The four phases defined for metadata curation are:

- *Phase Plan*: This phase recommends that, before the execution of any actual curatorial activity, establishing action plan necessary to deliver results in accordance

with the expected metadata curation requirements. More specifically, the CurationMaster in conjunction with Stakeholders and MetadataCurators must identify the metadata needs and metadata curation requirements as well as the means to fulfill these needs and requirements. The Systematic Dimensions are used as a repository of practices and processes, which might be implemented. In particular, the Metadata Curation Analysis and Planning Dimension processes should be used to deliver the action plan. Moreover, an action plan will be based on the processes and activities presented in rest of the systematic dimensions.

- *Phase Do*: This phase comprises the execution of an organized and disciplined curation of metadata according to the elaborated action plan. That is, at this stage, the actual metadata curation work will be carried out, from metadata acquisition to quality assurance and preservation of the metadata activities. An essential component of this phase is the creation of the Curation Backlog and the Metadata Backlog, both of which will be used to coordinate and monitor the metadata curation work. Moreover, it is important to collect data for charting and analysis for the following Check phase. Both Metadata Curation Monitoring process and Communication and Feedback Management process can aid in the collection of such kind of data.

- *Phase Check*: This phase comprises the monitoring and evaluation of the curation processes and compares the results against the expected results (goals) to ascertain any differences. This helps to verify what strategies, policies, and procedures work better than others, and if strategies, policies and procedures can be improved as well.

- *Phase Act*: Phase in which a set of improvements will be suggested and implemented for the metadata curation initiative. In this phase, new strategy, policies, and procedures can be defined to be realized/improved by the MetadataCurators, as well as the development and implementation of new services and tools to support the curatorial activities. In addition, it may be suggested to define a new set of metadata curation goals.

Such iterative and incremental approach allows the continuous improvement of metadata curation practices. Its focus is on deploying and improving curation practices in Data Ecosystems, and it addresses the importance of utilizing the experiences of the MetadataCurators as an important source of input to the improvement of curation activities. To the best of our knowledge, all the digital curation models focus all the efforts in Plan and Do phases. In practice, most curators focus all their efforts here, neglecting, in particular, the check and act phases, and retrospective, where they should reflect on what they learned from the previous curated work and adjust their processes accordingly. In this context, this method aims to provide mechanisms to aid in an incremental process, based

on the agile practices, to aid actors to create a planned and disciplined metadata curation initiative in an adaptable and flexible manner.

5.5 REVISITING THE MOTIVATIONAL SCENARIO

In order to support effective metadata curation, it is necessary to envision a coherent strategy and a suite of services. It is also necessary to coordinate across a diverse range of actors and processes to deliver the necessary technological and human infrastructures. Moreover, metadata curation implicates in adaptations (or corrections) in the processes, methods and tools that directly and indirectly handle metadata. Such adaptations would require the need for specific recommendations related with the adaptation/changes that will be applied.

A possible solution for these problems would involve the Louvre as a repository of practices and processes to create a metadata curation initiative. In this sense, the Louvre framework acts as an action plan to achieve the desired metadata management and sharing goals.

Generally speaking, the first step towards a metadata curation initiative demands a group of actors to acknowledge the commitment for advocating, creating and maintaining such initiative. In our scenario, Alceu after receiving complaints from Elba, Fagner e Capiba decides to change how metadata is being created, maintained and shared in the Data Ecosystem he is providing datasets. Hence, he decides to start a metadata curation initiative. However, meeting all of metadata curation needs is not a single man job. It needs a commitment of a community to contribute to the initiative. So, Alceu advocates with government technicians and other actors to increase the awareness of metadata curation benefits.

Using the Louvre's Deployment Method, before initiating any creation/change/adaptation metadata management solution and services, Alceu needs to clearly set the objective for the initiative. Alceu identifies Elba, Capiba and Fagner as stakeholders and, then interviews them to envision the metadata curation requirements, as recommended by Metadata Curation Requirement Engineering process. According to the needs collected from Elba, Capiba and Fagner, the list of requirements includes:

- R1: create more metadata;
- R2: provide solutions to store metadata;
- R3: provide solutions to access metadata;
- R4: provide solutions to ingest metadata;
- R5: provide mechanisms to gather feedback;
- R6: provide mechanisms to enrich metadata;

- R7: provide solutions to versioning metadata.

All these requirements are registered at a Curation Backlog, which is shared to all contributors. These requirements help Alceu to identify any gaps between the metadata curation current position and where he hopes to reach in the future. For instance, Alceu diagnoses that the current set of solutions does not properly store and access metadata as well as does not allow to gather feedback, to ingest metadata, to version metadata and to enrich metadata.

Moreover, Alceu provides a small set of descriptive metadata. Unfortunately, this current set of metadata related to Budget Expenses and Budget Financial Statements datasets is not enough. For instance, Elba expects a set of metadata documenting quality of these datasets. As another example, Fagner demands structural metadata to describe the content of datasets. In order to address these requests, Alceu can himself produce these new set of metadata. However, some of them were already produced by other Data Ecosystem actors. For instance, Fagner generates the structural and operational metadata. The problem here can be classified in two categories: (i) how to create from scratch quality metadata? and (ii) how to receive metadata from external contributors?

Having an overarching strategy is essential to Alceu creates the means to coherently fulfill these above gaps. Hence, Alceu relates these requirements to the outcomes defined in each Louvre's processes. As a result, he identifies that Metadata Creation Management, Metadata Curation Platform Design, Metadata Curation Platform Implementation, Metadata Curation Platform Maintenance and Metadata Curation Ingest processes would help him to address the defined metadata curation requirements. Moreover, Alceu also makes use of recommendations of other processes to plan and coordinate the metadata curation initiative, such as Metadata Curation Planning and Metadata Curation Controlling processes.

For instance, based on the Metadata Curation Planning process, Alceu outlines as initial curation strategy the following steps:

- Implement a Metadata Curation Platform;
- Define means to receive metadata from external contributors;
- Create new metadata.

It is important to remark that R6 and R7 requirements are not being considered in this planning. Due to high complexity to implement such services, Alceu and the stakeholders decided to postpone them to future iterations. They prioritized the rest of requirements aiming to consolidate an essential set of services and procedures. Such rationale was inspired by the "Classifying and prioritizing requirements" activity of Metadata Curation Requirements Engineering process.

Alceu as CurationMaster is also in charge for coordinating the activities. As recommended by Metadata Curation Controlling process, Alceu creates a backlog for organizing the list of works to be performed. *A priori*, he only lists the high level requirements, such as *define/develop and provide solutions to store metadata*. Besides the backlog, Alceu needs to team up with other contributors. As mentioned before, the metadata curation work will be performed by a set of contributors with assigned roles and responsibilities.

In the following subsection, some concerns related to the planned key objectives are presented in details.

5.5.1 Metadata Curation Platform Implementation

In order to implement a Metadata Curation Platform, Alceu recruits other government technicians. Lets call them Amelinha and Ednardo. He created a clear and compelling cause by explaining to Amelinha and Ednardo how the current solutions used to publish metadata are not appropriate to store and share metadata. So, he proposes to adapt a ready-to-use catalog solution like CKAN and SOCRATA to store and publish metadata. By disseminating this new goal to other Data Ecosystem actors, Alceu recruited another contributor, called Belchior, who is an expert on the CKAN platform. As a specialist, Belchior takes the responsibility to lead the team, assuming the TeamLeader role.

As a part of the effort to implement the Metadata Curation Platform, the team performs a set of meetings to define all the tasks related to implementation of metadata curation platform. According to their analysis, CKAN only addresses the R2 and R3 requirements. They also discovered some new tasks related to the platform deployment and maintenance. In particular, they need to provide sufficient infrastructure to host the CKAN-based metadata repository as well to implement fixity checking mechanisms to ensure metadata integrity. These new tasks were inspired by both Metadata Curation Platform Deployment and Metadata Curation Platform Maintenance processes. Furthermore, the team also need to implement solutions to ingest metadata as well as to gather feedback. All these tasks are registered in a new backlog used to coordinate the team work.

Since Belchior is not co-located with the rest of team, the team creates a project at GitHub[2] platform, which includes issue management, time tracking, and collaboration tools like wikis and forums. All the communication will be disseminated through this platform. Such communication management concern is a recommendation of Communication and Feedback Management process.

A successful delivery of the outcomes will result from logical and iterative development of the tasks registered in the backlog. The team agreed to break the curation work into at least five sprints. In the first sprint, they will assemble the necessary infrastructure and properly configure it. In second sprint, the team plans to install and create a personalized

[2] http://github.org

metadata repository using CKAN. In the third sprint, the team will implement a fixity mechanism to verify the integrity of stored metadata. In the fifth, they will implement the solutions to ingest metadata and develop and to gather feedback. Alceu as the CurationMaster can monitor the progress and efficiency of metadata curation work using a burndown chart as well as tasks board.

Once the metadata curation platform is deployed, Belchior can periodically evaluate the services and solutions in order to verify their adequacy to current needs. If the current platform does not deliver the expected outcomes, it can be improved or reimplemented. Incidentally, the stakeholders can also define new metadata curation requirements that would demand a new set of services, tools and systems. Both scenarios is aligned with the PDCA cycle defined at Deployment Method. Moreover, as Belchior may leave the metadata curation initiative in the future, he prepares a maintenance plan to guide Amelinha and Ednardo to perform themselves the platform maintenance, as recommended by Metadata Curation Platform Maintenance.

5.5.2 Metadata Creation

As mentioned before, there is a lack of quality metadata. Before starting to produce metadata from scratch, Alceu needs to plan how to deliver metadata that meets some context specific quality requirements. In order to achieve such goal, Alceu decides to address the recommendations of Metadata Creation Management and Metadata Model Management processes.

His first step is to define a proper model to describe quality metadata. Instead of design his own model, Alceu analyzes the literature to identify and select a well-accepted standard to describe data quality information. Then, he selects the Data Quality Management Vocabulary [3], which provides a model for the structured representation of data requirements, data quality assessment results, data cleansing rules, and data requirement violations connected to their origin.

Once the model was chosen, Alceu also wants to ensure that all contributors can produce quality metadata. As recommended by Metadata Creation Management, he needs to document the workflow used to prepare and create data quality metadata. Such documentation would help contributors not familiar with data quality management to create related metadata. To reach a wider audience, Alceu employs multiple methods to document the workflow. For instance, he provides a very comprehensive checklist to be considered across the creation of quality metadata. He also provides a flowchart that depicts the quality metadata creation process.

Now, Alceu begins the proper creation of quality metadata. As this metadata curation tasks does not require so much effort, Alceu does not delegate the creation to another contributor. So, he will also act as a MetadataCurator. By following the previously created

[3] http://semwebquality.org/dqm-vocabulary/v1/dqm.html

workflow, Alceu creates quality metadata that allows Elba or other Data Ecosystem actors to assess completeness, timeliness and accuracy of Budget Expenses and Budget Financial Statements datasets. As final step, Alceu stores such metadata into Metadata Repository.

5.5.3 Metadata Ingestion

Often, actors produce metadata related other's resources. For instance, in our motivational scenario, both Fagner and Capiba are producing metadata related to Budget Expenses and Budget Financial Statements datasets. However, these metadata are not shared back to Alceu or other government technicians. With this in mind, Belchior and his team developed a solution to allow external ingestion of metadata. Nonetheless, such mechanism does not ensure that actors will contribute to the metadata curation initiative. Contributors need to feel in control when they hand over the fruits of their labour and should be able to rely on getting back what they put in. They can reasonably also expect to benefit from some form of added value.

So, it is necessary to create a deposit agreement that will set out terms and conditions that communicate the responsibilities and rights of metadata depositor. The agreement should give the repository rights to manipulate the data, as preservation may require migration to to new formats. It should also allow the repository to reserve the right to withdraw the data for legal or other reasons. Since none of the current contributors is not skilled on legal regulations, Alceu recruits another actor to act as a TechnicalExpert, which will produces a deposit agreement template. Such deposit agreement is a recommendation of Metadata Ingest Management process.

Moreover, once ingested into the metadata repository, all metadata deposited undergo quality checks to confirm the accuracy and usability of the metadata. Anomalies in metadata are corrected in consultation with metadata depositors. In order to perform the quality checks, the solution used to deposit metadata automatically register a set of quality review tasks into the curation backlog. Any skilled actor is compelled to accept this tasks. For instance, Amelinha or Ednardo as government technicians are prone to execute the quality review.

5.5.4 Summary of Motivational Scenario

The scenario described in this Section provided an example of some activities that can be performed to curate metadata. Through examples, we showcased how Louvre make planning the activities easier, less daunting task and making metadata curation more systematic. Of course, there are plenty of different scenarios for metadata curation. Each one would make use of a different set of processes, activities and practices.

5.6 CONSIDERATIONS OF THE CHAPTER

This Chapter presented a framework proposal, called Louvre, for metadata curation in Data Ecosystems. This framework was constructed from the state-of-the-art in data management and digital curation identified throughout the research work. The proposed framework adheres to well-known standards about the construction of processes and reference models. In addition, the development of the Louvre framework was guided by a set of key factors, as well as was based on a set of management and curation activities of data and metadata identified from a broad bibliographic review. This initial set of activities was adapted, improved, extended, refined and verified in several cycles until its presentation in the present work.

The proposed framework, in its version 2.0, was structured in 19 processes (with their respective purposes, inputs, outcomes and sets of activities), which were distributed in six dimensions. The Louvre framework also recommends a set of agile practices, a deployment process, and a set of roles to support metadata curation work. The Louvre form a comprehensive set of elements from which a group of actors can construct metadata curation model appropriate to their context. That is, a group of actors, depending on their goals and available resources, can select and apply an appropriate subset to fulfill that purpose.

In the next Chapter we present the evaluations of the proposed metadata curation framework.

6 EVALUATION OF THE LOUVRE FRAMEWORK

Evaluation is an important procedure to define the feasibility, coherence, completeness and adequacy of the proposed framework. Empirical methods based on evaluation in real environments (*e.g.*, case studies and controlled experiments) usually lead to collect a more qualified evidence about the phenomena being studied (EASTERBROOK et al., 2008). In our context, such approaches require that different Data Ecosystems implement a metadata curation initiative. Moreover, the entire metadata curation initiative must be analyzed over a long period of time.

However, such approach may be impossible or impractical due to several issues. A problem that arises when designing evaluations in real environments is the complexity in controlling the variables and experimental subjects. The independent variables cannot be fully controlled, and a proper analysis and interpretation of the validation results is very hard to achieve (GARCIA, 2010). Albuquerque (2005) and Garcia (2010) highlight other factors such as the schedule alignment among experimental unit and research work, budget limitations to invest in innovative programs, and even the intrinsic uncertainty whether the solution to be implemented will generate satisfactory results or losses. For this reason, researchers and practitioners tend to find alternative ways to evaluate and validate their proposals.

Considering the methodologies used for evaluation in the studies analyzed in the research exploratory phase of this thesis, two approaches were adopted to evaluate our proposal: survey based on the expert opinion and focus group. In the first, the framework was submitted to a group of experts (researchers and professionals of industry and academia) in the areas of consumption, publication and/or management of data, with different levels of experience in Data Ecosystems. Then, a refined version of the framework was evaluated using a focus group study. Both studies contributed to validate the framework proposed and, besides, the collected feedback were used to improve the framework. This Chapter introduces the planning of each evaluation study and presents the results obtained by them.

6.1 SURVEY BASED ON EXPERTS OPINION

The evaluation with experts corresponds to a field research through observation of a population. According to Groves et al. (2011), a survey is a "systematic method for gathering information from (a sample of) entities for the purpose of constructing quantitative descriptors of the attributes of the larger population of which the entities are members".

Expert opinion has been recognized as an important assessment tool (GROVES et al., 2011; EASTERBROOK et al., 2008; KITCHENHAM; PFLEEGER, 2002). In the literature, some

researches have identified some evidences about the relevance of this technique. For instance, Kitchenham et al. (2002) analyzed the precision of several methods based on the opinion of experts. According to them, a process of human estimative centered in the opinion of experts can overcome substantially simple models of quantitative analysis (*e.g.*, function point-based analysis). Thus, it is possible to say that the computing community in general has given more importance and credibility (reliability) to studies that use the expert opinion technique (FARIAS, 2014).

The survey evaluation approach for this study was systematically organized in a set of phases inspired in Almeida (2015) and Solingen e Berghout (1999) works. It is composed of the following five phases:

- Planning. The background and problem are defined, characterised, and planned, resulting in a clearly systematized evaluation plan;
- Definition. The questionnaire is defined, validated, and documented.
- Selection of Experts. A number of experts must be identified based on a set of criteria which should include the credibility, knowledge, ability, and dependability of experts;
- Data Collection. Ensures the conditions of conduction to an elicitation process and collects evaluation data;
- Interpretation. The collected data is processed with respect to a set of defined metrics into measurement results, that provide answers to the defined questions.

The planning phase is performed to fulfill all basic requirements for performing a successful survey evaluation, including training, management involvement and project planning. During the definition phase all survey material is developed, including the questionnaire and their related research questions and related metrics. During the selection of experts phase a number of experts is identified and selected from a target population to conduct a survey. The purpose of sampling is to reduce the cost and/or the amount of work that it would take to survey the entire target population. When all definition activities are completed, actual evaluation can start. During data collection phase, data collection forms are sent to the selected experts, filled-in and stored in a measurement database. In the final phase, the data collected is analyzed with the objective of characterizing the Louvre framework with respect to its feasibility, coherence, completeness and adequacy.

6.1.1 Survey Questionnaire

The survey consisted of a questionnaire developed with a strong influence from survey research guidelines (KITCHENHAM; PFLEEGER, 2002) and studies related to the evaluation

of reference models for software engineering projects (LUNA, 2009; ALMEIDA et al., 2015; GARCIA, 2010). The survey was defined and revised for about one month.

The questions asked issues related to: the completeness of agile practices, roles, dimensions, processes, activities and outcomes; the correctness of specification and description of the entire Louvre's elements; adequacy of activities and outcomes; the difficulty in implementing the Louvre elements.

The above aspects were evaluated using statements. For instance, for "completeness" the following statements were presented "The set of agile practices is complete" and "The set of activities is complete". For each statement, the participants were asked to indicate to what extent the statements hold in a scale with values ranging from 1 to 5, where the number 1 corresponds to "Strongly Disagree" and the number 5 corresponds to "Strongly Agree".

The final questionnaire was composed of 90 questions and it was designed to be concluded in 2 hours. A sketch of the questionnaire is available at <https://bit.ly/2UcTPt5>.

6.1.2 Survey Population and Sample

As previously described, the population or universe for this study was composed of research subjects with knowledge in: publication and consumption of data; Data Ecosystems; Data Management; and/or Data Curation. Among the positions/functions preferably performed by research subjects we can highlight: IT managers; Project Managers; Systems Analysts; Developers; Professors; Researchers; among others.

Sample sizes may be chosen considering different aspects (GIL, 2008), including:

- Population Size: The size of the sample is related to the size of the population;
- Confidence Level: Quantifies the level of confidence that a data evaluated lies in the margin of error. The larger the required confidence level, the larger the sample size;
- Margin of Error: Expresses a range of values below and above the sample statistic in a confidence interval;
- Expected proportion: Quantifies the (estimated) proportion of the population expected to find.

The population can be classified as finite or infinite. Finite population represent those whose number of elements does not exceed 100,000. Infinite population are those that are greater than 100,000 elements. In Data Ecosystem context, several recent surveys (*e.g.*, (Data World, 2018; BEAN, 2016; Stitch Data, 2015)) point out that the population of actors involved in Data Ecosystems can exceed 100,000 elements. For example, Stitch Data (2015) identified, in 2015, at least 11,400 data scientists employed by companies known to LinkedIn. They also identified at least 6,200 different companies employing

data scientists. The same report had pointed out that the number of data scientists has doubled over the last 4 years. Corroborating this affirmation, Data World (2018) also identified at least 126,093 data science job opportunities registered at Indeed.com site. It is important to remark that these numbers are a conservative estimate of data science professionals and includes only those who explicitly identify themselves as data scientists. As presented in Section 3.2.5, there are plenty of other kinds of Data Ecosystem roles. All of these numbers suggest that total of different actors involved with Data Ecosystems can exceed 100,000 elements.

Based on these aspects, a stratified random sampling from an infinite population is given by (GIL, 2008):

$$n = \frac{\sigma^2 \cdot p \cdot q}{e^2}$$

where:

- n = sample size;
- σ^2 = confidence level chosen, expressed in number of standard deviations;
- p = Expected proportion of estimated population;
- q = Complementary proportion ($100 - p$);
- e^2 = maximum permissible error.

For this study, the total set of practitioners and researchers of Data Ecosystem can be classified as an infinite population. We established that expected proportion of population would be around 3%, therefore q is equal to $100 - 3$, that is, 97. Then, a confidence level of 93% (corresponds to 1.83 standard deviations) and a maximum error of 7% was adopted. The aforementioned values were established from simulations that resulted in a viable value for carrying out the study and obtaining the desired results. Thus, by applying these values to the formula presented above, we arrive at the following result:

$$n = \frac{1.83^2 \cdot 3 \cdot 97}{7^2}$$

Therefore, to meet the requirements established by the study, it was established that the number of sample elements should count with a response of at least 19.88 respondents.

6.1.3 Survey Data Collection

During data collection phase, a questionnaire was used as an instrument. This questionnaire was developed using the Google Forms™online service. This questionnaire was available online from 1st to 31th of November, 2018.

To this thesis, about 160 potential respondents were selected, composing the group of experts for this research. They represent a group of experts with capacity to evaluate and contribute to the improvement of the Louvre framework. From the selected sample, only 25 experts participated of this study. The next session will describe the process of their selection. In the course of this study, we made it clear the data collected would be used for research purposes. Moreover, the data collected from the respondents remain confidential. It means that have no way of associating any survey results with the person who submitted that response.

Exclusion criteria were used to identify respondents who would not be included or who would have his evaluation discarded from the study. In particular, were excluded respondents who did not fit the profile of the research, those who refused to participate voluntarily and those who have not fully responded to the questionnaire. From a total of 25 questionnaires were answered, only 20 were considered valid and 5 were discarded, as they met the criteria presented above. The main criterion for discarding was the partial completion of the questionnaire. Thus, the total sample for this study ended up being composed of 20 respondents. This sample was considered statistically significant, since it was necessary at least 19.88 participants according to the probability theory presented in Section 6.1.2.

6.1.4 Survey Results

This section presents the analysis of data collected, discussing in details the main questions and pointing some correlation points that must be taken in consideration.

6.1.4.1 Experts Characterization

In order to better understand the respondents' context and background, some questions were provided to classify each of the respondents. During the data analysis, it was observed that most of the respondents had a master's degree (65%), the remainder had a bachelor's degree (25%) and a doctorate degree (10%) (Figure 23). In addition, the majority of the respondents classified their current position/function either as System Analyst (25%) or Developer (15%). It was also observed a considerable participation (40%) of professors and students (master and doctorate candidates). Regarding educational background, it was observed that 95% of respondents have a background in Computer Science or related areas (Information Systems or Computer Engineering). In addition, 50% of the respondents develop their activities in the academic context, 20% work in industry and 30% work in both academic and industry context (*cf.* Figure 24).

Regarding respondent's experience, it was observed that 100% of respondents had directly or indirectly participated in Data Ecosystems initiatives. It was observed that 55% of respondents had "between 1 and 5 years" and 25% of respondents had "between 6 and 10 years" experience in Data Ecosystems (Figure 25). On a scale from 1 to 5, where

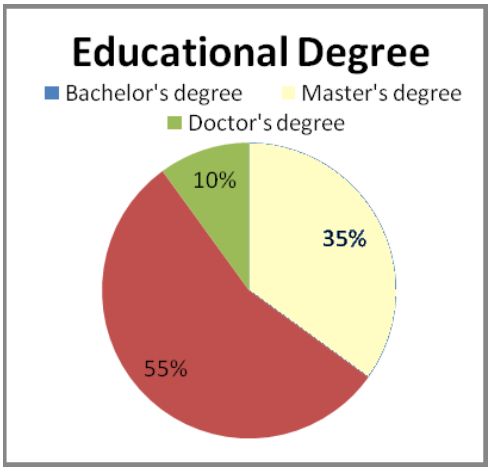

Figure 23 – Experts Educational Degree. Source: Author

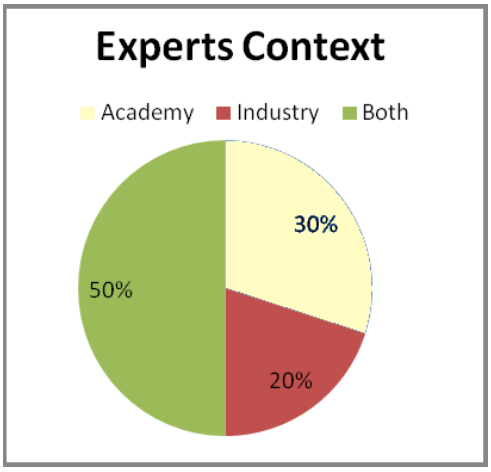

Figure 24 – Experts Working Context. Source: Author

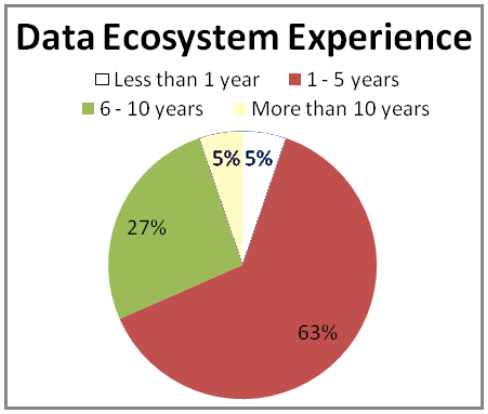

Figure 25 – Experts Experience on Data Ecosystems. Source: Author

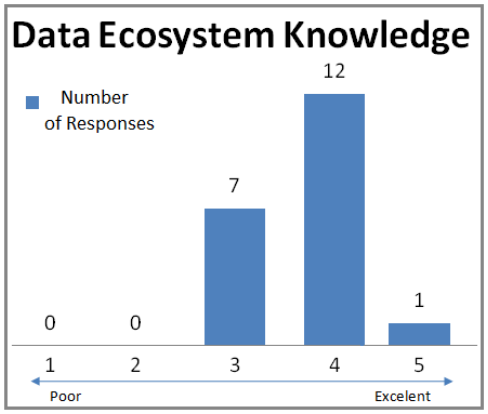

Figure 26 – Experts Knowledge about Data Ecosystems. Source: Author

the number 1 corresponds to "Poor" and the number 5 corresponds to "Excellent", 65% of respondents rated their knowledge on Data Ecosystems between 4 and 5. Yet, 35% of participants ranked their knowledge on Data Ecosystems as 5 (*cf.* Figure 26).

Regarding the experience on metadata curation, 55% of the respondents, on a scale from 1 to 5, where the number 1 corresponds to "Poor" and the number 5 corresponds to "Excellent", classified their knowledge in metadata curation as at least 3 (*cf.* Figure 27). Moreover, 85% of the respondents have curated metadata related to Data Ecosystems. The respondents were also asked to classify their metadata curation practices, on a scale from 1 to 5, where the number 1 corresponds to "Ad-hoc" and the number 5 corresponds to "Well-organized". None of the respondents classified their metadata curation practices as structured or well-organized (*cf.* Figure 29). Furthermore, two respondents had not curated any metadata. Finally, 90% of the respondents classified as important and very important the curation of metadata in Data Ecosystems (*cf.* Figure 28).

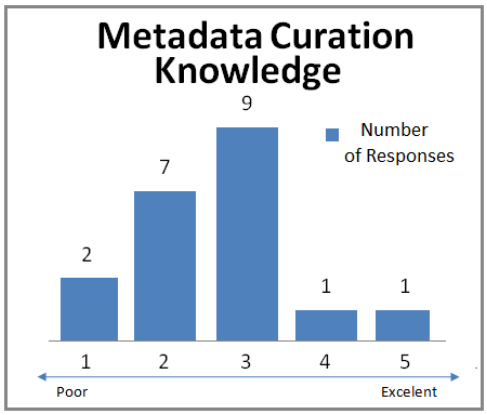

Figure 27 – Experts Knowledge about Metadata Curation. Source: Author

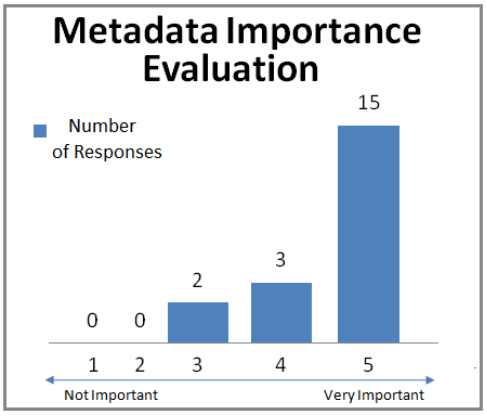

Figure 28 – Experts Rating about Metadata Importance for Data Ecosystems. Source: Author

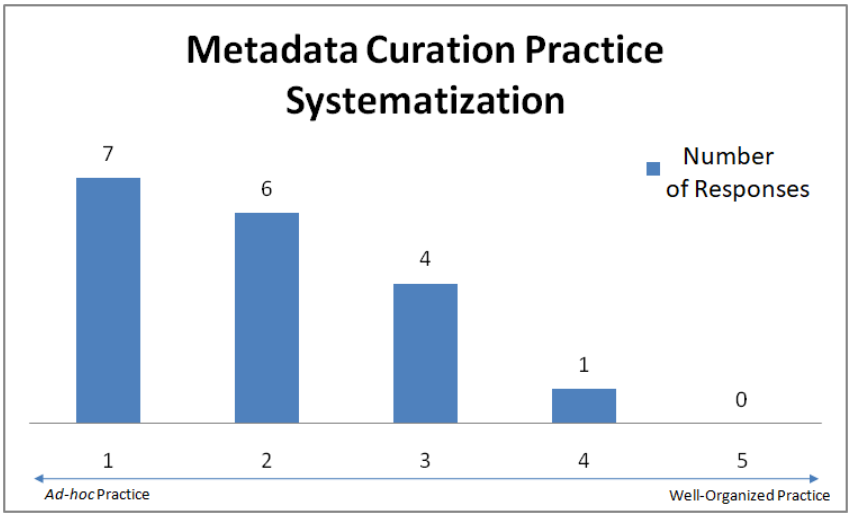

Figure 29 – Experts Classification of their Metadata Curation Practice. Source: Author

6.1.4.2 Louvre Evaluation

Table 24 presents the results obtained from the evaluation of the agile practices, actors and roles, processes, activities and outcomes. Individually, the set of averages also indicates positive results in terms of correctness, completeness, adequacy and feasibility. The respondents' evaluation was positive. In addition, over 85% of responses correspond to higher values of the scale (Agree and Strongly Agree). According to Table 24, the correctness and adequacy of processes, activities and outcomes were scored with the highest values.

In its turn, completeness and feasibility aspects were scored with lower values. This seems to reflect constraints related to lack of fully knowledge on metadata curation or even on general curation practices. Without such a background, turns difficult to measure of the completeness and feasibility of metadata curation activities and their outcomes compared to the total universe of available practices. This hypothesis is grounded on

Table 24 – Evaluation Results

Evaluation Statements	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The set of roles is correctly described	0.00%	0.00%	0.00%	35.00%	65.00%
The set of roles is complete	0.00%	0.00%	10.00%	55.00%	35.00%
The set of roles is appropriate	0.00%	0.00%	5.00%	40.00%	55.00%
The set of agile practices is correctly described	0.00%	0.00%	0.00%	30.00%	70.00%
The set of agile practices is complete	0.00%	0.00%	20.00%	55.00%	25.00%
The set of agile practices is appropriate	0.00%	0.00%	10.00%	50.00%	40.00%
The set of agile practices is feasible to be implemented	0.00%	0.00%	20.00%	50.00%	30.00%
The process is correctly described	0.00%	0.53%	1.84%	28.68%	68.95%
The process is enough for its purpose	0.00%	0.53%	3.42%	31.32%	64.74%
The set of activities is correctly described	0.00%	0.53%	3.16%	28.16%	68.16%
The set of activities is complete	0.00%	1.32%	9.74%	38.95%	50.00%
The set of activities is appropriate	0.00%	0.53%	2.89%	30.26%	66.32%
The implementation of the set of activities is feasible	0.00%	0.00%	3.42%	42.63%	53.95%
The set of outcomes is correctly described	0.00%	0.00%	2.37%	31.32%	66.32%
The set of outcomes is complete	0.00%	0.79%	10.26%	35.26%	53.68%
The set of outcomes is appropriate	0.00%	0.26%	3.68%	29.47%	66.58%
The set of outcomes is capable of being achieve	0.00%	0.00%	6.05%	35.26%	58.68%

the minimum set of suggestions or comments received from respondents. Except for agile practices and the set of roles, none suggestion to include new activities or outcomes was collected. Regarding the evaluation of feasibility, respondents agreed that the activities and outcomes were likely to be implemented and achieved, respectively. However, we again believe that the lack of experience in systematized metadata curation practices has impacted their confidence on strongly agreeing on the feasibility of activities and outcomes.

The following Subsections present a detailed analysis of each of Louvre element evaluation.

Roles Evaluation

We asked if the set of roles is correctly described in the version alpha of the Louvre framework (*cf.* Table 24). 35% and 65% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of roles is complete. 55% and 35% of respondents agreed and strongly agreed with this affirmation, respectively. Finally, we asked if the set of roles is appropriate. 40% and 55% of respondents agreed and strongly agreed with this affirmation, respectively. From results, it can be seen that, on average,

95% of respondents evaluated the set of roles as correct, complete and appropriate.

Two respondents evaluated as neutral the statement "The set of roles is complete". These two respondents suggested the creation of specialized roles. In its turn, another respondent also evaluated as neutral the statement "The set of roles is appropriate". This respondent did not suggest a specification a new role nor the modification or exclusion of a existing one. However, he questioned the motive for create roles based on the Scrum framework.

A respondent suggested the inclusion of a new role. According to him, the "*it is important to specify a metadata promoter role that would be responsible for the dissemination, communication and doubts related to the consumption of metadata by the grand public.*" Similarly, another respondent suggested "*Like the PlatformAdministrator role [the Louvre] should specify more specialized roles, such as metadata designers and metadata validators, should be recommended.*"

In response to these comments, the set of roles was defined based on the Scrum approach, which does not recommend the creation of specialized roles. Often, several actors do not have full autonomy to perform all the metadata curatorial activities. Hence, it is difficult to assume that in all curation teams is possible to define specialized roles. The Louvre define only the MetadataCurator roles. All specialized curation tasks should be solved through teamwork. The intention is also to reduce complexity, dissolving unnecessary complex organizational solutions, and coordinate metadata curation in a simpler way. According to Grgić (2015), less roles, less management, less organizational structures.

Agile Practices Evaluation

We asked if the set of agile practices is correctly described in the version alpha of the Louvre framework (Table 24). 30% and 70% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of agile practices is complete. 55% and 25% of respondents agreed and strongly agreed with this affirmation, respectively. Moreover, we asked if the set of agile practices is appropriate. 50% and 40% of respondents agreed and strongly agreed with this affirmation, respectively. Finally, we asked if the set of agile practices is feasible to be implemented. 50% and 30% of respondents agreed and strongly agreed with this affirmation, respectably. From results, it can be seen that, on average, 87.5% of respondents evaluated the set of agile practices as correct, complete, appropriate and feasible.

Four respondents evaluated as neutral the statement "The set of agile practices is complete". These four respondents suggested the tailoring of new agile practices. Actually, two of these four respondents also evaluated as neutral the statement "The set of agile practices is appropriate". Four respondents evaluated as neutral the statement "The set of agile practices is feasible to be implemented". One of these four respondents questioned the feasibility of automated tests to verify quality proprieties.

Some comments from respondents are presented below:

- **Comment 1**: *[..] would be interesting to include the stand up meetings that could be carried out by the team leader to follow the process and update the progress of the activities (burndown chart)*
- **Comment 2**: *Regular meetings are a well-known characteristic in agile methods. Maybe they could be explicitly represented in the Louvre Framework.*
- **Comment 3**: *I missed the daily meetings*
- **Comment 4**: *[would be interesting] a planning activity, or the equivalent of "planning poker", to prioritize the curatorial activities.*
- **Comment 5**: *The "automated test" practice could be more detailed. It is missing the description of how to verify/measure the metadata quality.*

In response to Comment 5, the mentioned practice was refined. In addition, with respect to comments 1,2 and 3, we analyzed the suitability of several agile practices for the metadata curation context. During the construction of the alpha version of the Louvre, we were not confident about the possibility of performing regular meetings. As a result of the comments received, we adapted and incorporated to the Louvre framework the practice of coordination meetings. However, due to the distributed and autonomous nature of the actors involved in Data Ecosystems, we believe it is not feasible to hold daily meetings with curation teams. Regarding Comment 4, we also incorporated this new practice to the Louvre.

Metadata Curation Planning Dimension Evaluation

Table 25 presents the average of the evaluations made by the respondents related to all processes belonging to the Metadata Curation Planning dimension (current Metadata Curation Analysis and Planning dimension) and their elements. In particular, Table 25 presents for each evaluation grade a percentage of the grand total of all the responses collected.

In particular, we asked to the respondents to evaluate a set of statements related to each process and to the sets of activities and outcomes associated to each process. With regards to the statement "The process is correctly described", 37.5% and 72.5% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the process is enough for its purpose. 50.00% and 47.50% of respondents agreed and strongly agreed with this affirmation, respectively.

Moreover, we asked to evaluate the set of activities of each process. With regards to the statement "The set of activities is correctly described", 40% and 60% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of

Table 25 – Metadata Curation Analysis and Planning Dimension Evaluation

Evaluation Statements	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The process is correctly described	0.00%	0.00%	0.00%	37.50%	62.50%
The process is enough for its purpose	0.00%	0.00%	2.50%	50.00%	47.50%
The set of activities is correctly described	0.00%	0.00%	0.00%	40.00%	60.00%
The set of activities is complete	0.00%	0.00%	10.00%	40.00%	50.00%
The set of activities is appropriate	0.00%	0.00%	7.50%	37.50%	55.00%
The implementation of the set of activities is feasible	0.00%	0.00%	2.50%	42.50%	55.00%
The set of outcomes is correctly described	0.00%	0.00%	0.00%	37.50%	62.50%
The set of outcomes is complete	0.00%	0.00%	15.00%	37.50%	47.50%
The set of outcomes is appropriate	0.00%	0.00%	10.00%	30.00%	60.00%
The set of outcomes is capable of being achieved	0.00%	0.00%	2.50%	47.50%	50.00%

activities is complete. 40.00% and 50% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of activities is appropriate". 37.50% and 55.00% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the implementation of the set of activities is feasible. 42.50% and 55.00% of respondents agreed and strongly agreed with this affirmation, respectively.

We also asked to evaluate the set of outcomes of each process. With regards to the statement "The set of outcomes is correctly described", 37.50% and 62.50% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of outcomes is complete. 37.50% and 47.50% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of outcomes is appropriate". 30.00% and 60.00% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the set of outcomes is capable of being achieved. 47.50% and 50.00% of respondents agreed and strongly agreed with this affirmation, respectively.

From results, it can be seen that, on average, 95% of respondents evaluated the set of processes correct and complete as well as their sets of activities and outcomes as correct, complete, appropriate and feasible.

One respondent evaluated as neutral the statement "The process is enough for its purpose". He recommended the creation of more activities related to metadata curation planning and analysis. For the same reason, four respondents evaluated as neutral the statement "The set of activities is complete". Four and five respondents evaluated as neutral the statement "The set of outcomes is complete" and "The set of outcomes is appropriate", respectively. Despite evaluating as neutral, these respondents did not provide any comment about their evaluation.

Some comments from respondents are presented below:

- **Comment 1**: *I could not fully understand why the requirements engineering dimension comes along with the planning dimension.*
- **Comment 2**: *There are some activities that could be integrated into a single activity. For example, "defining and prioritizing metadata needs" and "defining and reviewing metadata standards".*
- **Comment 3**: *There should be a process for analyzing the use of data / licenses.*

In response to Comment 2, the mentioned activities were refined after the evaluation in order to better differentiate them. In addition, with respect to Comment 1, processes related to requirements management and planning are typically separated into distinct phases in software development processes. However, during the literature review, we have identified that some data management models (*e.g.*, (MOSLEY et al., 2010)) group these processes into a single dimension or phase. In any case, the description of the dimension as well as its title were refined in order to ease the understanding about the dimension's purpose. In relation to Comment 3, we believe that the creation of an exclusive process for the analysis of data and licenses is beyond the scope of Louvre framework. Such kind of process is more focused on data management rather than metadata.

Metadata Acquisition Evaluation

Table 26 presents the average of the evaluations made by the respondents related to all processes belonging to the Metadata Acquisition Management dimension and their elements. In particular, Table 26 presents for each evaluation grade a percentage of the grand total of all the responses collected.

We asked to the respondents to evaluate a set of statements related to each process and to the sets of activities and outcomes associated to each process. With regards to the statement "The process is correctly described", 33.75% and 65.00% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the process is enough for its purpose. 28.75% and 70.00% of respondents agreed and strongly agreed with this affirmation, respectively.

Moreover, we asked to evaluate the set of activities of each process. With regards to the statement "The set of activities is correctly described", 23.75% and 75.00%of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of activities is complete. 40.00% and 50.00% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of activities is appropriate". 37.50% and 61.25% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the implementation of the set of

Table 26 – Metadata Acquisition Dimension Evaluation

Evaluation Statements	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The process is correctly described	0.00%	0.00%	1.25%	33.75%	65.00%
The process is enough for its purpose	0.00%	0.00%	1.25%	28.75%	70.00%
The set of activities is correctly described	0.00%	0.00%	1.25%	23.75%	75.00%
The set of activities is complete	0.00%	0.00%	10.00%	40.00%	50.00%
The set of activities is appropriate	0.00%	0.00%	1.25%	37.50%	61.25%
The implementation of the set of activities is feasible	0.00%	0.00%	6.25%	47.50%	46.25%
The set of outcomes is correctly described	0.00%	0.00%	2.50%	31.25%	66.25%
The set of outcomes is complete	0.00%	0.00%	11.25%	37.50%	51.25%
The set of outcomes is appropriate	0.00%	0.00%	1.25%	33.75%	65.00%
The set of outcomes is capable of being achieved	0.00%	0.00%	6.25%	33.75%	60.00%

activities is feasible. 47.50% and 46.25% of respondents agreed and strongly agreed with this affirmation, respectively.

We also asked to evaluate the set of outcomes of each process. With regards to the statement "The set of outcomes is correctly described", 31.25% and 66.25% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of outcomes is complete. 37.50% and 51.25% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of outcomes is appropriate". 33.75% and 65.00% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the set of outcomes is capable of being achieved. 33.75% and 60.00% of respondents agreed and strongly agreed with this affirmation, respectively.

From results, it can be seen that, on average, 95.64% of respondents evaluated the set of processes correct and complete as well as their sets of activities and outcomes as correct, complete, appropriate and feasible.

The Metadata Acquisition dimension received a considerable number of neutral evaluations. In particular, eight and nine respondents evaluated as neutral the statement "The set of activities is complete", considering the four processes specified in this dimension. The Metadata Creation management received the greater number of neutral responses. Unfortunately, one only commented was provided for the entire dimension. We believe that in this case some respondents resented a lack of knowledge about metadata creation. Because of the lack of knowledge, they did not feel confident in agreeing or disagreeing with the statements.

The comment received was "*I lacked the activity of mapping, among the Data Ecosystem, which repositories can be consumed*". We believe that this concern is already addressed by some activities of the "Metadata Harvesting Management" process. However,

we refined the activity "Identifying, analyzing, and selecting metadata sources" to make it clear that metadata sources can be found among the resources provided and exchanged in Data Ecosystems.

Metadata Quality Management Evaluation

Table 27 presents the average of the evaluations made by the respondents related to all processes belonging to the Metadata Quality Management dimension and their elements. In particular, Table 27 presents for each evaluation grade a percentage of the grand total of all the responses collected.

We asked to the respondents to evaluate a set of statements related to each process and to the sets of activities and outcomes associated to each process. With regards to the statement "The process is correctly described", 27.50% and 70.00% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the process is enough for its purpose. 25.00% and 67.50% of respondents agreed and strongly agreed with this affirmation, respectively.

Table 27 – Metadata Quality Management Dimension Evaluation

Evaluation Statements	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
The process is correctly described	0.00%	0.00%	2.50%	27.50%	70.00%
The process is enough for its purpose	0.00%	0.00%	7.50%	25.00%	67.50%
The set of activities is correctly described	0.00%	0.00%	0.00%	30.00%	70.00%
The set of activities is complete	0.00%	2.50%	7.50%	37.50%	52.50%
The set of activities is appropriate	0.00%	0.00%	2.50%	27.50%	70.00%
The implementation of the set of activities is feasible	0.00%	0.00%	0.00%	30.00%	70.00%
The set of outcomes is correctly described	0.00%	0.00%	2.50%	27.50%	70.00%
The set of outcomes is complete	0.00%	2.50%	12.50%	27.50%	57.50%
The set of outcomes is appropriate	0.00%	0.00%	2.50%	32.50%	65.00%
The set of outcomes is capable of being achieved	0.00%	0.00%	2.50%	32.50%	65.00%

Moreover, we asked to evaluate the set of activities of each process. With regards to the statement "The set of activities is correctly described", 30.00% and 70.00%of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of activities is complete. 37.50% and 52.50% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of activities is appropriate". 27.50% and 70.00% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the implementation of the set of activities is feasible. 30.00% and 70.00% of respondents agreed and strongly agreed with this affirmation, respectively.

We also asked to evaluate the set of outcomes of each process. With regards to the statement "The set of outcomes is correctly described", 27.50% and 70.00% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of outcomes is complete. 27.50% and 57.50% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of outcomes is appropriate". 32.50% and 65.00% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the set of outcomes is capable of being achieved. 32.50% and 65.00% of respondents agreed and strongly agreed with this affirmation, respectively.

From results, it can be seen that, on average, 95.5% of respondents evaluated the set of processes correct and complete as well as their sets of activities and outcomes as correct, complete, appropriate and feasible.

The Metadata Quality Management dimension received a considerable number of neutral evaluations. However, one respondent evaluated as disagree two statements: "The set of activities is complete" and "The set of outcomes is complete". Both statements were related to the Metadata Quality Control process. According to this respondent, this process was lacking of more activities related to the management of quality issues.

Some comments from respondents are presented below:

- **Comment 1**: *I'm not sure if the activity of "Managing metadata quality issues" should be in the process of Metadata Quality Improvement. From what I could understand, the activity should be part of the Metadata Quality Control.*
- **Comment 2**: *Maybe, it might have some aspect related to tools or technologies related to the Quality management*

In response to Comment 1, the mentioned activity was moved to the "Metadata Quality Control". Regarding to Comment 2, we refined the "Developing metadata quality metrics and measurement methods" and "Measuring and monitoring metadata quality" activities to include the management of quality tools"

Metadata Preservation and Dissemination Dimension

Table 28 presents the average of the evaluations made by the respondents related to all processes belonging to the Metadata Preservation and Dissemination dimension and their elements. In particular, Table 28 presents for each evaluation grade a percentage of the grand total of all the responses collected.

We asked to the respondents to evaluate a set of statements related to each process and to the sets of activities and outcomes associated to each process. With regards to the statement "The process is correctly described", 31.25% and 63.75% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the process is

enough for its purpose. 36.25% and 60.00% of respondents agreed and strongly agreed with this affirmation, respectively.

Table 28 – Metadata Preservation and Dissemination Dimension Evaluation

Evaluation Statements	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The process is correctly described	0.00%	0.00%	5.00%	31.25%	63.75%
The process is enough for its purpose	0.00%	1.25%	2.50%	36.25%	60.00%
The set of activities is correctly described	0.00%	0.00%	10.00%	30.00%	60.00%
The set of activities is complete	0.00%	1.25%	10.00%	33.75%	55.00%
The set of activities is appropriate	0.00%	0.00%	3.75%	28.75%	67.50%
The implementation of the set of activities is feasible	0.00%	0.00%	7.50%	36.25%	56.25%
The set of outcomes is correctly described	0.00%	0.00%	3.75%	32.50%	63.75%
The set of outcomes is complete	0.00%	1.25%	7.50%	36.25%	55.00%
The set of outcomes is appropriate	0.00%	0.00%	3.75%	26.25%	70.00%
The set of outcomes is capable of being achieved	0.00%	0.00%	3.75%	36.25%	60.00%

Moreover, we asked to evaluate the set of activities of each process. With regards to the statement "The set of activities is correctly described", 30.00% and 60.00%of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of activities is complete. 33.75% and 55.00% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of activities is appropriate". 28.75% and 67.50% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the implementation of the set of activities is feasible. 36.25% and 56.25% of respondents agreed and strongly agreed with this affirmation, respectively.

We also asked to evaluate the set of outcomes of each process. With regards to the statement "The set of outcomes is correctly described", 32.50% and 63.75% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of outcomes is complete. 36.25% and 55.00% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of outcomes is appropriate". 26.25% and 70.00% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the set of outcomes is capable of being achieved. 36.25% and 60.00% of respondents agreed and strongly agreed with this affirmation, respectively.

From results, it can be seen that, on average, 93.87% of respondents evaluated the set of processes correct and complete as well as their sets of activities and outcomes as correct, complete, appropriate and feasible.

The Metadata Preservation and Dissemination dimension received a considerable number of neutral evaluations. However, one respondent evaluated as disagree three state-

ments: "The process is enough for its purpose", "The set of activities is complete" and "The set of outcomes is complete". Such evaluations were related to the Metadata Versioning Management process. According to this respondent, this process should be removed from dimension. In addition, he stated that versioning process is lacking of more detailed activities.

Some comments from respondents are presented below:

- **Comment 1**: *According to description of "Identify and resolving conflicts" activity, deduplicating metadata would be a type of error and could be handled in this activity.*
- **Comment 2**: *Maybe the version control process could be deleted.*
- **Comment 3**: *I did not identify any missing activities. But I feel that the set of processes could be reduced, there are overlaps. Example: Authentication and Access Control are usually related to activities of the same process.*

Historically, literature on data warehouse and data cleansing have considered duplicated data as an error/inconsistency type. However, recent studies on data integration and big data analytic have reconsidered such issue as multiple instance of the same information, which does not necessarily represent an error. A hypothetical example could be a metadata that identify a person's name. This name could be stored differently in multiple sources. For instance, "Marcelo Iury", "Marcelo Iury S. Oliveira" or "Oliveira, M.I.S". All these values represent the same person and none are false or incorrect. The conflict resolution activity is aimed at identifying which of these values represent the metadata to be preserved. With regards to Comment 2, due to these multiple values, as well as the possibility of metadata change over time, it may be necessary to perform the version control of metadata. In addition to these two aspects, it is important to remark that metadata creation in many cases is a subjective task. Different metadata creators can generate different metadata about a same resource. Thus, version control can help to maintain the different versions as well as assist in conflict resolution. In response to Comment 3, the process of "Metadata Preservation and Dissemination" were refined after the evaluation in order to better differentiate them and reduce the overlaps.

Metadata Curation Monitoring and Controlling Dimension Evaluation[1]

Table 29 presents the average of the evaluations made by the respondents related to all processes belonging to the Metadata Curation Monitoring and Controlling dimension (current Metadata Curation Coordination dimension) and their elements. In particular, Table 29 presents for each evaluation grade a percentage of the grand total of all the responses collected.

[1] This is the former name for Metadata Curation Coordination Dimension.

We asked to the respondents to evaluate a set of statements related to each process and to the sets of activities and outcomes associated to each process. With regards to the statement "The process is correctly described", 23.75% and 76.25% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the process is enough for its purpose. 28.75% and 68.75% of respondents agreed and strongly agreed with this affirmation, respectively.

Table 29 – Metadata Curation Monitoring and Controlling Dimension Evaluation

Evaluation Statements	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
The process is correctly described	0.00%	0.00%	0.00%	23.75%	76.25%
The process is enough for its purpose	0.00%	0.00%	2.50%	28.75%	68.75%
The set of activities is correctly described	0.00%	0.00%	2.50%	25.00%	72.50%
The set of activities is complete	0.00%	1.25%	8.75%	43.75%	46.25%
The set of activities is appropriate	0.00%	0.00%	2.50%	17.50%	80.00%
The implementation of the set of activities is feasible	0.00%	0.00%	0.00%	52.50%	47.50%
The set of outcomes is correctly described	0.00%	0.00%	2.50%	32.50%	65.00%
The set of outcomes is complete	0.00%	1.25%	8.75%	36.25%	53.75%
The set of outcomes is appropriate	0.00%	0.00%	2.50%	27.50%	70.00%
The set of outcomes is capable of being achieved	0.00%	0.00%	11.25%	35.00%	53.75%

Moreover, we asked to evaluate the set of activities of each process. With regards to the statement "The set of activities is correctly described", 25.00% and 72.50% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of activities is complete. 43.75% and 46.25% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of activities is appropriate". 17.50% and 80.00% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the implementation of the set of activities is feasible. 52.50% and 47.50% of respondents agreed and strongly agreed with this affirmation, respectively.

We also asked to evaluate the set of outcomes of each process. With regards to the statement "The set of outcomes is correctly described", 32.50% and 65.00% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of outcomes is complete. 36.25% and 53.75% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of outcomes is appropriate". 27.50% and 70.00% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the set of outcomes is capable of being achieved. 35.00% and 53.75% of respondents agreed and strongly agreed with this affirmation, respectively.

From results, it can be seen that, on average, 95.62% of respondents evaluated the set of processes correct and complete as well as their sets of activities and outcomes as correct, complete, appropriate and feasible.

The Metadata Curation Monitoring and Controlling dimension received a considerable number of neutral evaluations. However, one respondent evaluated as disagree two statements: "The set of activities is complete" and "The set of outcomes is complete". Such evaluations were related to the Metadata Feedback and Communication Management process. According to this respondent, this process is ignoring activities related to dissemination of metadata curation assets.

Some comments from respondents are presented below:

- **Comment 1**: *I do not see difference between the "Assigning Stewardship" and "Forming Curation Team" activities. [...They appear be] the same thing.*
- **Comment 2**: *The coordination process is a bit out of context.*
- **Comment 3**: *[...Create] a 'public relations' role responsible to promote the consumption of information [would help the communication process].*

In response to Comment 1, the mentioned activities were refined after the evaluation in order to better differentiate them. Regards to Comment 2, we refined the entire process in order to ease its understanding. In relation to Comment 3, we believe the "Recruiting and Engagement Management" process already recommends a set of activities aimed at promoting metadata curation as well as the curated resources. Therefore, we believe it is not necessary to create a new activity for this purpose.

Metadata Curation Platform Administration Dimension Evaluation

Table 30 presents the average of the evaluations made by the respondents related to all processes belonging to the Metadata Curation Platform Administration dimension and their elements. In particular, Table 30 presents for each evaluation grade a percentage of the grand total of all the responses collected.

We asked to the respondents to evaluate a set of statements related to each process and to the sets of activities and outcomes associated to each process. With regards to the statement "The process is correctly described", 20.00% and 75.00% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the process is enough for its purpose. 23.33% and 68.33% of respondents agreed and strongly agreed with this affirmation, respectively.

Moreover, we asked to evaluate the set of activities of each process. With regards to the statement "The set of activities is correctly described", 26.67% and 68.33%of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of activities is complete. 38.33% and 46.67% of respondents agreed and strongly agreed with

Table 30 – Metadata Curation Platform Administration Dimension Evaluation

Evaluation Statements	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The process is correctly described	0.00%	3.33%	1.67%	20.00%	75.00%
The process is enough for its purpose	0.00%	1.67%	6.67%	23.33%	68.33%
The set of activities is correctly described	0.00%	3.33%	1.67%	26.67%	68.33%
The set of activities is complete	0.00%	3.33%	11.67%	38.33%	46.67%
The set of activities is appropriate	0.00%	3.33%	1.67%	36.67%	58.33%
The implementation of the set of activities is feasible	0.00%	0.00%	1.67%	40.00%	58.33%
The set of outcomes is correctly described	0.00%	0.00%	1.67%	26.67%	71.67%
The set of outcomes is complete	0.00%	0.00%	10.00%	33.33%	56.67%
The set of outcomes is appropriate	0.00%	1.67%	5.00%	28.33%	65.00%
The set of outcomes is capable of being achieved	0.00%	0.00%	6.67%	30.00%	63.33%

this affirmation, respectively. Another affirmation evaluated was "The set of activities is appropriate". 36.67% and 58.33% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the implementation of the set of activities is feasible. 40.00% and 58.33% of respondents agreed and strongly agreed with this affirmation, respectively.

We also asked to evaluate the set of outcomes of each process. With regards to the statement "The set of outcomes is correctly described", 26.67% and 71.67% of respondents agreed and strongly agreed with this affirmation, respectively. We also asked if the set of outcomes is complete. 33.33% and 56.67% of respondents agreed and strongly agreed with this affirmation, respectively. Another affirmation evaluated was "The set of outcomes is appropriate". 27.50% and 70.00% of respondents agreed and strongly agreed with this affirmation, respectively. Furthermore, we also asked if the set of outcomes is capable of being achieved. 30.00% and 63.33% of respondents agreed and strongly agreed with this affirmation, respectively.

From results, it can be seen that, on average, 93.5% of respondents evaluated the set of processes correct and complete as well as their sets of activities and outcomes as correct, complete, appropriate and feasible.

The Metadata Curation Platform Administration was the dimension evaluated with the greater number of disagree. In particular, one respondent criticized the Metadata Curation Platform Maintenance process. According to him, this process was focusing on risk control ignoring practical maintenance activities. Another respondent criticized the Metadata Curation Platform Maintenance process. According to him, such process specification is shallow.

Some comments from respondents are presented below:

- **Comment 1**: *I believe that the dimension is not very consistent with the activities performed. The first two activities are related to risk management, while the latter is more related to the maintenance of the platform.*
- **Comment 2**: *There is a lack of maintenance activities.*
- **Comment 3**: *Perhaps the implementation process could be incorporated into another process.*

We agree with Comment 1 and Comment 2. Hence, the mentioned process was refined after the evaluation in order to better specify maintenance activities. Moreover, we refined the entire process in order to ease its understanding. Regards to Comment 3, In fact, the two activities of the "Metadata Curation Platform Implementation" process could be incorporated into the other two processes of this dimension. However, we believe in the importance of highlighting that these activities represent a distinct phase between the design and maintenance of the platform. In addition, the implementation and deployment of the platform requires a significant effort from the contributors responsible for the metadata curation initiative. The lack of a implementation process may give the false impression that implementation activities are trivial.

6.1.5 Summary of the findings

The respondents were also asked about a high-level evaluation of Louvre's dimensions. More specifically, it was asked if either a new process should be incorporated into the dimension or if some process should be updated, moved or excluded. For the most part, 94% of the evaluations indicated that the processes were correctly described, with no need for any modification. Among the recommendations for modification is the creation of a new process for data analysis and license management. Another recommendation was the creation of a new process responsible for metadata dissemination. Although important, we believe that the first recommendation is partially covered by the planning process. As for the second recommendation, we created a new activity for publicizing the metadata curation initiative as well as the curated resources.

Furthermore, the respondents were also asked about a high-level evaluation of the Louvre Framework. In particular, they were asked to indicate to what degree the Louvre is adaptable and flexible. Such evaluation was performed on a scale of 1 to 5, where the number 1 corresponds to "Not Completely" and the number 5 corresponds to "Completely". For the most part, 95% and 85% classified the Louvre as flexible and adaptable, respectively. Finally, at least 95% of the respondents also indicated that a Data Ecosystem will obtain benefits from metadata curation and from Louvre.

In general, it was observed that, in the opinion of the respondents, the framework presents relevant level of completeness, correctness, adequacy and feasibility for metadata

curation in Data Ecosystems. This indicates that the main objectives can be achieved through fairly reliable processes and that these processes are written clearly and easily understood. Moreover, these respondents indicated, from their evaluations, a capacity to use the proposed processes, activities and other Louvre's elements without significant difficulties.

6.2 FOCUS GROUP

This second evaluation focused on the beta version of the proposed framework. To this end, we conducted an empirical study using focus group research method. A focus group is a technique that involves the use of group of participants brought together to focus on a particular issue or evaluate a particular topic (KONTIO; LEHTOLA; BRAGGE, 2004; KONTIO; BRAGGE; LEHTOLA, 2008). These participants are selected because of their knowledge of the study area and do not necessarily represent a representative sample of specific population.

Focus group are becoming popular in several research areas, such as software engineering and health research (SINGER; SIM; LETHBRIDGE, 2008; KONTIO; BRAGGE; LEHTOLA, 2008). According to Kontio, Lehtola e Bragge (2004), focus group is a fast and economical empirical method for collecting evidence and conducting evaluations using participants. This method can also provide qualitative data and rich information as well as reveal insights that are difficult or expensive to capture with other methods (KONTIO; LEHTOLA; BRAGGE, 2004). Moreover, according to Tremblay, Hevner e Berndt (2010), focus groups can be effectively applied to design research studies in order to promote the refinement of the proposed artifact and evaluation of its utility.

Furthermore, different from surveys, focus group is more dynamic. The type and range of data generated through the social interaction of the group are often richer and deeper than those collected from one-to-one data collection methods, like interviews and surveys (KONTIO; BRAGGE; LEHTOLA, 2008). In fact, according to Krueger e Casey (2014) data produced from the interaction between group members is richer than those collected from individual interviews. Lederman (1990) proposes that, as a result of the synergistic interaction which takes place, focus groups generate more than the sum of individual inputs.

6.2.1 Focus Group Protocol

Our process for conducting a focus group was inspired by Almeida (2015). As first steps, we defined the criteria for selecting participants, decided the session length, designed the sequence of questions to ask during the session, and prepared documents to provide the participants with the study background and objectives.

In order to gain insights about improvements and gaps in the Louvre framework, we used the following criteria for selecting the participants:

- knowledge and expertise on Data Ecosystems;
- knowledge and expertise on metadata management and/or metadata curation;
- willingness to share their experiences and candid opinion.

In addition, to ensure proper discussion and interaction during the session, another criterion was to invite participants who knew each other as friends or co-workers. According to these selection criteria, our study needed practitioners in consumption and publication of data, Data Ecosystems and/or digital curation or digital preservation. Such practitioners are usually very busy and are not likely to respond to invitations from unfamiliar sources. Thus, a random sampling was not viable. As matter of fact, recruiting participants is a significant challenge for any research project. Ten practitioners were considered as candidates for this study and were contacted. Two said that they did not have time to contribute with the research and one did not answer the invitation. Seven accepted the invitation. However, only five practitioners participated of focus group session. Participants were guaranteed anonymity and all data has been anonymized.

Table 31 – Focus Group Participants Characterization

Participant	**Current Position**	**Position Context**	**Educational Background**	**Educational Degree**	**Data Ecosystem Experience**	**Metadata Curation/ Metadata Management Experience**
Participant A	Professor	Industry and Academy	Computing	Doctorate degree	15 years	3 years
Participant B	Professor	Academy	Computing	Doctorate degree	3 years	2 years
Participant C	Professor	Industry and Academy	Computing	Doctorate candidate	5 years	5 years
Participant D	Digital Content Management Coordinator	Academy	Information Science	Master's degree	6 years	6 years
Participant E	System Analyst	Academy	Computing	Doctorate candidate	4 years	4 years

Table 31 presents the general characteristics of participants of this study. All the participants are resident in the state of Pernambuco, Brazil. Still about the characterization of the participants, three work exclusively in the academy, while two work in both academy and industry. From the participants of the research, four have the higher education in

computing and one in information science. Furthermore, one participant have the master degree, three are doctors and two are PhD students. All the participants had worked in the last years in the Data Ecosystem area. Regarding their current positions, three are professors, one is a Digital Content Management Coordinator and one is a System Analyst. All of them have different backgrounds and with different skills discuss in Data Ecosystem and Metadata Curation.

The focus group session was held in December 2018, lasting approximately four hours. All the participants had at least five days to analyze the Louvre documentation. Such documentation was sent by email. Moreover, participants were asked to read and analyze the documentation.

The flow of the discussion was designed to be as natural and as easy to follow as possible. The discussions followed a semi-structured guide, which was based on the approach used by Almeida (2015).

The session started with a brief introduction of the participants and researchers. The introduction included the name of the participants, their organizations, current position, experience, and application domain. Then, the discussion flowed through a predefined sequence of specific topics.

In particular, an evaluation questionnaire was distributed to each of the participants. Such questionnaire contained the information that should be evaluated for each of the main elements of the Louvre framework. Moreover, the questionnaire consisted on nine closed questions and four opened questions. The questions were adapted from (GARCÍA-CASTRO; CORCHO; HILL, 2012; FARIAS, 2014). This questionnaire had two objectives. At first, it aimed to present the framework reducing the bias that would be created from an oral presentation performed by the researchers. Second, it was used to collect quality evaluation data as well as general comments. These quality evaluation data supported the discussions during the study. In addition, such questionnaire was also used for a later analysis of the information collected during the study. The questionnaire sketch is available at <https://bit.ly/2HthF20>.

At the beginning of session, the participants were encouraged to freely present their opinions. In addition, the participants were informed about ground rules needed throughout the session. These rules included: there are no right and wrong answers; everyone's ideas and experiences are valuable; it is important to hear all sides, including both positives and negatives. It was also remarked that divergences among participants would not classified as a problem. All comments are considered valid.

The session was audio recorded with the participants' consent. Two researchers conducted the focus group, in which one moderated the discussion and the other took extensive notes. Moreover, the session was held at the Federal University of Pernambuco.

6.2.2 Focus Group Results

This section presents the analysis of the data collected from focus group, discussing in details the main questions and pointing some issues that must be taken in consideration.

6.2.2.1 Actors and Roles Evaluation

We asked if the set of roles is correctly described. The idea was to identify if the set of roles is feasible, complete and appropriate for the curation of metadata in Data Ecosystems. We also wanted to identify gaps, errors and/or possibilities for improvement. Figure 30 shows the distribution.

Participant C selected "No, one or more roles need to be updated" and affirm that "*I do not know if the TechnicalExpert role makes sense. In a software development context, it makes sense due to teams typically have different technical capabilities. But in the context of data curation, I believe it is expected that anyone who is doing data curation is already an expert in the subject, right? [In addition,] Maybe PlatformOwner role is better represented as PlatformManager or something. And, I also classify the first sentence of "TeamLeader" ("because of a passion for the subject") too superficial. Often, (especially in software companies) a team leader has more human management skills rather than passion for the subject.*"

Participant D selected "No, one or more roles need to be inserted" and affirm that "*I recommend include a validator role. Perhaps it is contemplated by the MetadataCurator*".

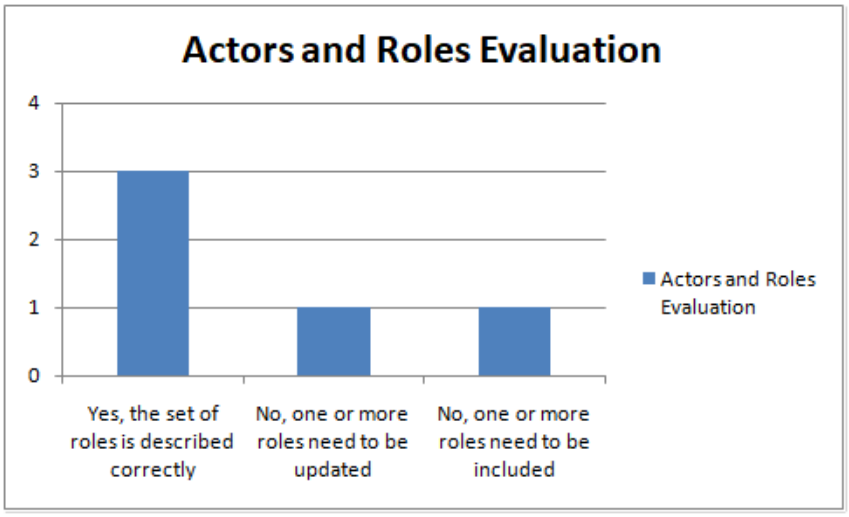

Figure 30 – Actors and Roles Evaluation
Source: Author

The other participants selected "Yes, the set of roles is described correctly". However, the *Participant B* made an observation: "*Regarding the SCRUM perspective, I considered the set of role described correctly. But, I feel myself insecure to point out the inclusion or deletion of roles as the set of roles may vary depending on the context of application*

or use of the framework.". The *Participant E* also made an observation: "*With regards to the flexibility of the framework, the set of roles is adequately described. Although in some cases there are overlap between some roles.*".

Participant A did not make any consideration about this question.

Some improvements were suggested, such as the specification of new roles. These improvement was also identified during survey evaluation. However, in a minor perspective, we observed some disagreeing positions in relation to the propositions of new roles. We had four participants agreeing on the Louvre's recommendation for not overspecializing the set of roles. It is important to remark that Louvre does not forbid the creation of specialized curation roles. These specialized roles could be specified to address specific contexts. For instance, a validator role could be created in scenarios where the compliance with legal policies is critical.

Concerning the suggestion to exclude the role of TechnicalExpert, we partially agree with the comment. Although MetadataCurators are expected to be able to learn and perform the curation activities, in some cases, a specialist can be consulted to boost the learning process. However, this role does not represent a primary role. Hence, we refined the description of the set of roles to highlight that the Louvre recognizes the existence of some secondary roles, which are typically introduced on a temporary basis or in specific cases. In summary, the rest of suggestions were related to the improvement of the nomenclature.

6.2.2.2 Agile Practices Evaluation

We asked if the set of agile practices is correctly described. The idea was to identify if the set of agile practices is feasible, complete and appropriate for the curation of metadata in Data Ecosystems. We also wanted to identify gaps, errors and/or possibilities for improvement. Figure 31 shows the distribution.

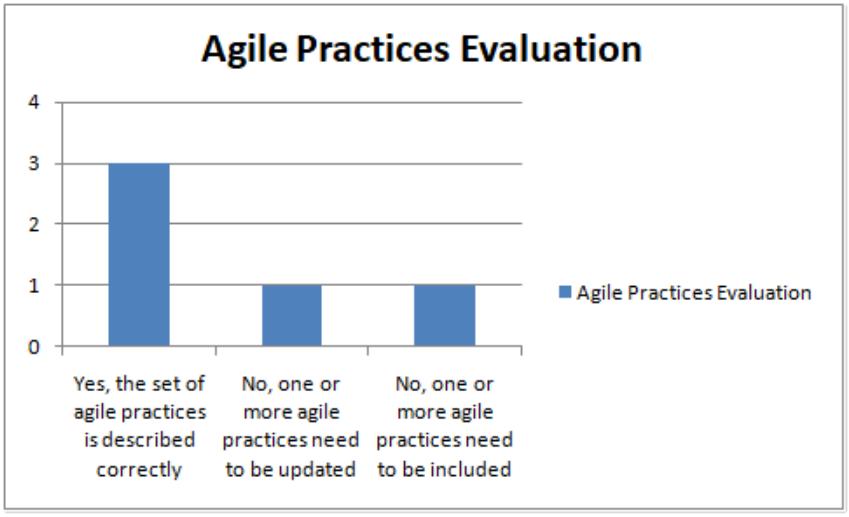

Figure 31 – Agile Practices Evaluation
Source: Author

Participant A selected "No, one or more agile practices need to be updated" and affirm that "*I believe that the practice "Collective Ownership" could be better described. In particular, it should state that not all metadata can be accessed / edited by all. It's important to affirm this in some way.*" *Participant C* selected "No, one or more agile practices need to be inserted" and affirm that "*As you describe the use of backlog and continuous integration, perhaps the use of sprints could be considered.*". The other participants selected "Yes, the set of roles is described correctly". They did not make any consideration about this question.

In response to the comments, we refined the set of agile practices to address both suggestions.

6.2.2.3 Metadata Curation Analysis and Planning Dimension Evaluation

We asked if the set of processes of this dimension is correctly described. The idea was to identify if the set of processes is feasible, complete and appropriate for the analysis and planning of metadata curation in Data Ecosystems. We also wanted to identify gaps, errors and/or possibilities for improvement. Figure 32 shows the distribution.

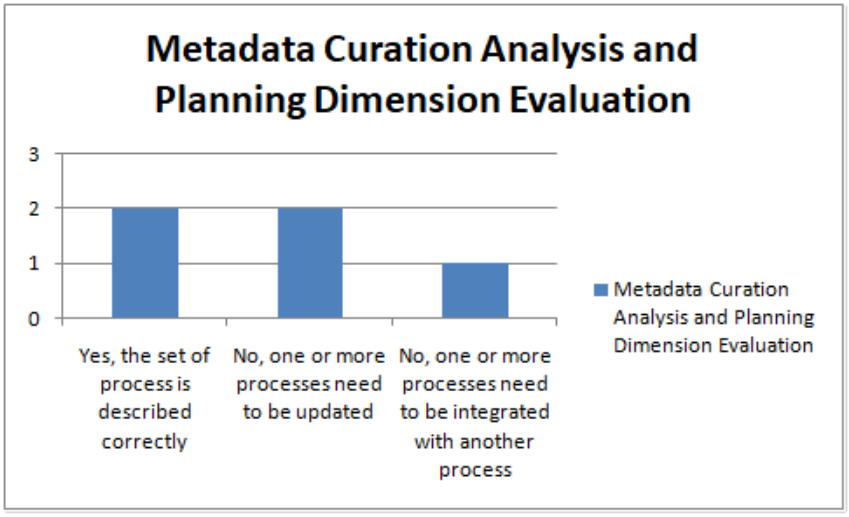

Figure 32 – Metadata Curation Analysis and Planning Dimension Evaluation
Source: Author

Participant A selected "No, one or more processes need to be updated" and affirm that "*From the point of view of nomenclature, I suggest using a standard, or noun or verb [for specifying the activities titles].* ".

Participant B selected "No, one or more processes need to be updated" and affirm that "*I suggest specifying inputs for each process. What do I need to have before starting to carry out the activities?*".

Participant C also selected "No, one or more processes need to be updated" and agreed with Participant B. According to *Participant C*, "*I think it would be interesting to make process' inputs clear. Since the activities and outcomes are well defined, it is natural to look for which are the incomes of the activities. Perhaps the outcome "the needs and expectations of stakeholders concerning metadata curation are collected" is actually an input.*". *Participant C* also suggested to "*move or refine the activity "define and approving Platform Curation Architecture*".

The other participants selected "Yes, the dimension is described correctly". They did not make any consideration about this question.

We had two experts acknowledging this dimension is well-defined, coherent and feasible. We could also include the acceptance of *Participant A*, since his recommendation was related to the improvement of the nomenclature. *Participant B* and *Participant C* also did not suggest a specific change for any process. They recommended an improvement of the structure of the Louvre, which we also considered important.

6.2.2.4 Metadata Acquisition Dimension Evaluation

We asked if the set of processes of this dimension is correctly described. The idea was to identify if the set of processes is feasible, complete and appropriate for the acquisition

and selection of metadata. We also wanted to identify gaps, errors and/or possibilities for improvement. Figure 33 shows the distribution.

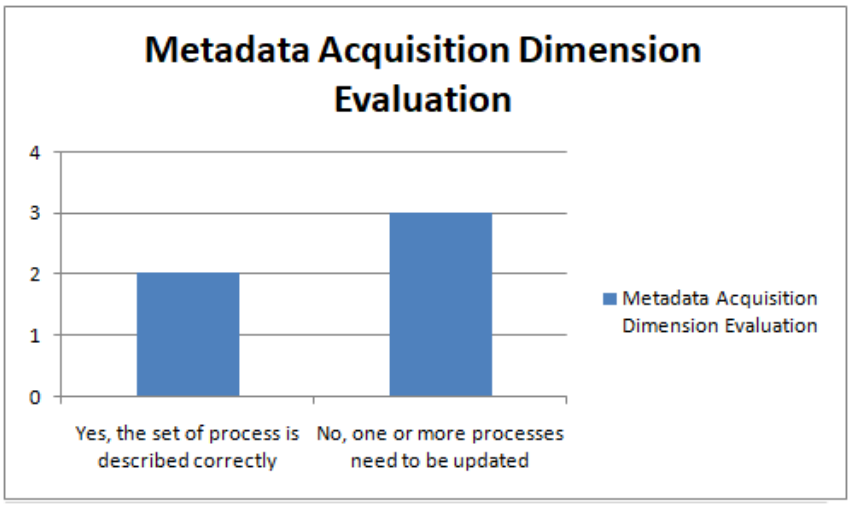

Figure 33 – Metadata Acquisition Dimension Evaluation
Source: Author

Participant A selected "No, one or more processes need to be updated" and affirm that "*I believe that aspects of security, privacy and property should be covered by this dimension, like these aspects were presented in the previous dimension.*".

Participant B selected "No, one or more processes need to be updated" and affirm that "*The activity of defining modeling language must be refined since the "defining" term implies implementing a language from scratch. Also, in Appraisal and Selection process, the identification of policies to guide the selection of metadata should not be a step earlier?*".

Participant C selected "No, one or more processes need to be updated" and affirm that "*In the Metadata Creation Management process, I recommend a correction in the title of the "selecting **AND** developing a model" activity. It might be more appropriate "selecting **OR** developing a model". There is also an overlap between the processes "Metadata Creation Management" and "Metadata Model Management"*".

Participant E selected "Yes, the dimension is described correctly", however made an observation: "*From the metadata acquisition perspective, the provenance of the metadata is very important. However, there is no activity in this dimension covering this concern. Such activity is important for final quality.*".

The *Participant D* selected "Yes, the dimension is described correctly". He did not make any consideration about this question.

We had two experts acknowledging this dimension is well-defined, coherent and feasible. *Participant A* and *Participant E* recommended some new activities. *Participant B* and *Participant C* recommended improvements of description of some activities and processes. *Participant C* highlighted an overlap between two processes. All of mentioned activities and processes were refined after evaluation of the proposed improvements.

6.2.2.5 Metadata Quality Management Dimension Evaluation

We asked if the set of processes of this dimension is correctly described. The idea was to identify if the set of processes is feasible, complete and appropriate for the metadata quality management. We also wanted to identify gaps, errors and/or possibilities for improvement. Figure 34 shows the distribution.

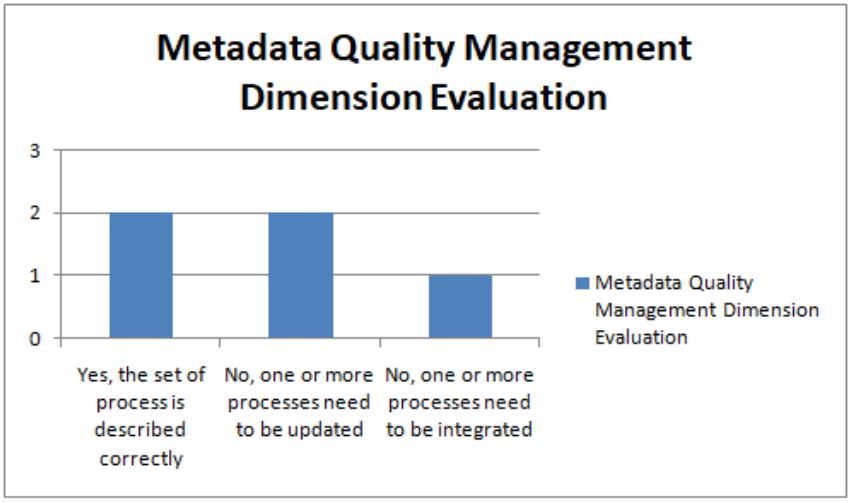

Figure 34 – Metadata Quality Management Dimension Evaluation
Source: Author

Participant A selected "No, one or more processes need to be updated" and affirm that "*I believe it would be important to provide the monitoring and updating of the goals, deadlines, indicators and plan of the metadata curation processes as a whole and not just focus on the quality of the metadata.*".

Participant B selected "No, one or more processes need to be updated" and affirm that "*The quality management dimension should be linked to the activity of "ensuring basic quality control" in the "Metadata Creation Management" process. A simple mention would resolve this issue.*".

Participant C selected "No, one or more processes need to be integrated" and affirm that "*A possible solution to my comment about the "Metadata Curation Monitoring and Controlling" dimension might be to move to "Metadata Quality Management" dimension all technical issues.*".

The other participants selected "Yes, the dimension is described correctly". They did not make any consideration about this question.

We agree with *Participant C*'s comment. In order to address such improvement, we refined both dimensions to separate the human and technical concerns. This improvement partially address *Participant A*'s recommendation. *Participant B* recommended improvements of description of some activities and processes. All of mentioned activities and processes were refined after evaluation of the proposed improvements.

6.2.2.6 Metadata Preservation and Dissemination Dimension Evaluation

We asked if the set of processes of this dimension is correctly described. The idea was to identify if the set of processes is feasible, complete and appropriate for the preservation and dissemination of metadata in Data Ecosystems. We also wanted to identify gaps, errors and/or possibilities for improvement. Figure 35 shows the distribution.

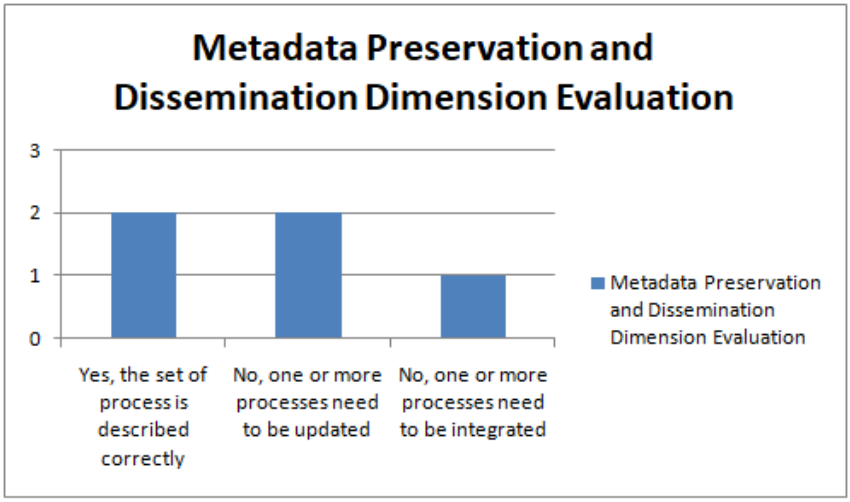

Figure 35 – Metadata Curation Monitoring and Controlling Dimension Evaluation
Source: Author

Participant A selected "No, one or more processes need to be updated" and affirm that "*In relation to dissemination, I did not identify anything related to the publicity of some metadata, nor did I identify the policies to encourage the use of the metadata to disseminated.*".

Participant B selected "No, one or more processes need to be moved" and affirm that "*The provenance activity, in ingest, must be taken into the acquisition process. Also, outcomes' descriptions should be adjusted. Some outcomes descriptions looks like activities.*".

Participant C selected "No, one or more processes need to be updated" and affirm that "*There is an overlap between authentication in 8.1 and all 8.4 which is just about access management.*".

Participant D selected "Yes, the dimension is described correctly", however made an observation: "*There is a lack of activities related to ensuring the integrity and reliability of the stored metadata. These activities are important to ensure the preservation of metadata.*".

We had two experts acknowledging this dimension is well-defined, coherent and feasible. *Participant A* and *Participant D* recommended some new activities. *Participant B* recommended to move some activities to another dimension. *Participant C* highlighted an overlap between two processes. All of mentioned activities and processes were refined after evaluation of the proposed improvements.

6.2.2.7 Metadata Curation Monitoring and Controlling Dimension Evaluation

We asked if the set of processes of this dimension is correctly described. The idea was to identify if the set of processes is feasible, complete and appropriate for the monitoring and controlling of metadata curation in Data Ecosystems. We also wanted to identify gaps, errors and/or possibilities for improvement. Figure 36 shows the distribution.

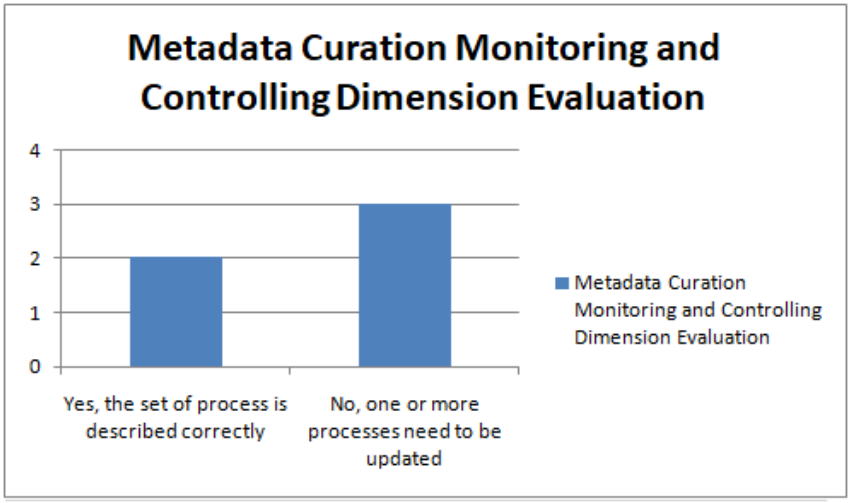

Figure 36 – Metadata Curation Monitoring and Controlling Dimension Evaluation
Source: Author

Participant A selected "No, one or more processes need to be updated" and affirm that "*It would be important to present examples in the description of activities. What examples of monitoring indicators?.* ".

Participant B selected "No, one or more processes need to be updated" and affirm that "*a notion of relationship between the steps/processes would help to understand the flow of the execution of the activities. Also, a better definition of the inputs and outputs of each process could help to ease the understanding of relationships and dependencies between the steps. Some inputs may be the outcomes of the previous steps"*".

Participant C selected "No, one or more processes need to be updated" and affirm that "*It was a bit confusing for me to differentiate between monitoring and controlling. Some of the activities/processes control and monitor people (e.g., "recruitment and engagement"), while other activities/process monitor and control technical things (e.g., burndown). [...] I suggest to separate these two concerns in different dimensions.*".

The other participants selected "Yes, the dimension is described correctly". They did not make any consideration about this question.

We had two experts acknowledging this dimension is well-defined, coherent and feasible. *Participant A* recommended some description improvements. *Participant B* suggested again the specification of inputs for each process. However, we disagree with the suggestion for specifying dependencies between activities. This suggestion is contrary to the premise of adaptability used in the construction of the Louvre. The definition of dependencies can

impose to create a very large set of activities to be implemented by any metadata curation initiatives, regardless of their context. Meanwhile, we observed that both *Participant A* and *Participant B* also classified this dimension as well-defined, coherent and feasible, since they did not recommended any major modification related to processes and their elements.

However, we agree with *Participant C* comments. This dimension was addressing two distinct concerns. Thus, because of this problem, we refined this dimension, as presented in Chapter 5. In its final version, the processes recommended by the Louvre for this dimension cover only the coordination of the contributors as well as the metadata curation work. All activities related to technical control have been moved to Metadata Quality Management dimension.

6.2.2.8 Metadata Curation Platform Administration Dimension Evaluation

We asked if the set of processes of this dimension is correctly described. The idea was to identify if the set of processes is feasible, complete and appropriate for the design, implementation and maintenance of metadata curation platform. We also wanted to identify gaps, errors and/or possibilities for improvement. Figure 37 shows the distribution.

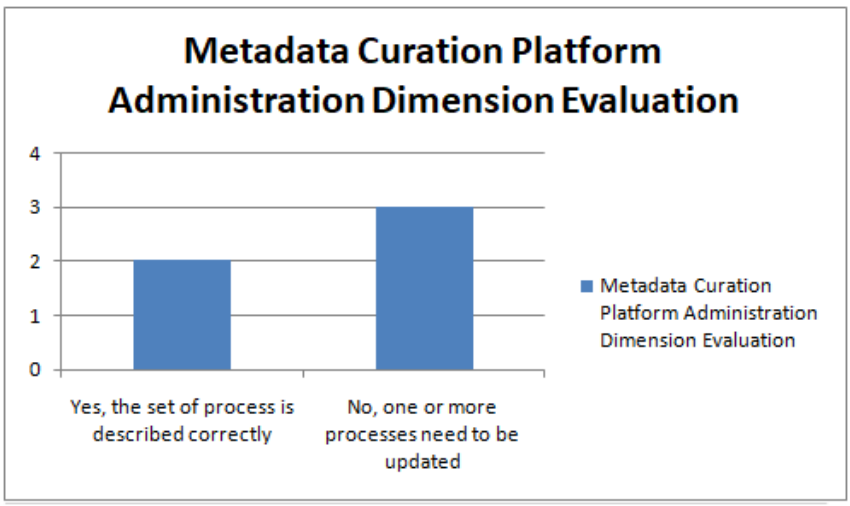

Figure 37 – Metadata Curation Platform Administration Dimension Evaluation
Source: Author

Participant A selected "No, one or more processes need to be updated" and affirm that "*I was unable to clearly identify the difference and key aspects that need to be considered in the activities listed within the context of metadata curation. The descriptions of some activities seem quite similar to those already adopted in traditional software engineering. It is not clear which aspects would be more relevant to the context under study.*".

Participant B selected "No, one or more processes need to be updated" and affirm that "*The outcomes of Metadata Curation Platform Implementation are confusing. How would*

it be possible installing systems before developing them? What is the difference between technologies and platform?".

Participant C selected "No, one or more processes need to be updated" and affirm that "*In the description of this dimension, it is defined that it consists of the design, implementation and maintenance processes of the platform. So you are considering that everyone will develop your curatorial platform. That does not make much sense. Are there signs? Very soon the process can simply be summarized choice of platform. It might be interesting to define processes of choice, or configuration, or deploy, or something ...* ".

The other participants selected "Yes, the dimension is described correctly". They did not make any consideration about this question.

We had two experts acknowledging this dimension is well-defined, coherent and feasible. Regarding to the "genericness" pointed by *Participant A*, in fact, during the construction of the Louvre we tried to specify its elements in a generic manner. Although many potential Louvre users work with software development, not necessarily all actors participating in Data Ecosystems have knowledge on design, implementation, deployment, or maintenance of systems. Therefore, we have chosen to give a generic or high-level view of the common activities related to the metadata curation platform administration. However, during activities description, we try to give some examples of specific metadata curation concerns. For example, the selection of storage technologies. In addition, the more generic specification of the activities was inspired by both the DMBOK and the ISO 8000 family of standards.

With regards to *Participant B*'s comment, as stated in 5.3.2, the Louvre does not imply any order for its processes or activities. However, we refined the description to mitigate the mentioned confusion. Similarly, we agree with *Participant C*. Again, we refined the description to mitigate the mentioned problem.

6.2.2.9 General evaluation of the Louvre framework

We asked if the Louvre framework is appropriate to guide and support the curation of metadata. The participants could select Yes or No. If the participants selected No, we asked him to justify his choice.

All participants selected "Yes", however some of then made some observations. *Participant B* affirmed "*Yes, it is. The components are appropriate. I can see its application in different contexts, but in the thesis it would help a lot to have a guideline. This guide might have to be defined by the team that will use the framework itself. In believe that such guide will be defined by the curation team that will use the framework*". According to *Participant B*, "*It supports, but it is complex to affirm that it guides, since you have not defined guidelines. If you apply this to a person who does not know how to curate metadata, does the framework guide this person? I think not. He supports, but does not guide.*". *Participant A* agreed with *Participant B*. *Participant E* affirmed "*Yes, it will help*

guide and support. But, in practice, a user may find some challenges to implement all the dimensions and processes, but several parts will be implemented in a practical manner.". *Participant D* affirmed "*Yes, it will help guide and support. But, in practice, an user may find some challenges to implement all the dimensions and processes, but several parts will be implemented in a practical manner.*". *Participant D* affirmed "*I would say yes, but depending on applicability. For example, in open scientific data scenarios, I believe that it would guide and support. But I would like to to see a practical case.*"

Table 32 – Focus Group - General Evaluation of Louvre Framework

	Louvre is adaptable	**Louvre is flexible**
Participant A	4	4
Participant B	4	4
Participant C	4	5
Participant D	4	4
Participant E	4	5

We asked to the participants to indicate to what extent the following statements hold:

- Louvre is adaptable
- Louvre is flexible

They had to indicate a value ranging from 1 to 5, where the number 1 corresponds to "Strongly Disagree" and the number 5 corresponds to "Strongly Agree". Table 32 shows the distribution of answers.

We asked if the elements of the Louvre support the definition of a systematic curation of metadata, *i.e.*, Louvre provides elements to create a path from informal *ad-hoc* curation to systematic metadata curation. The participants could select Yes or No. If the participants selected No, we asked him to justify his choice.

Four participants answered "Yes", however some of then made some observations. *Participant E* did not answered neither "Yes" nor "No". He made an observation: "*I can not answer. I believe it will improve the metadata curation. But, I do not know if applied in different contexts, it will achieve the same results. Compared to informal approaches, it is certainly is a better alternative.*"

6.2.3 Summary of the findings

The aim of this study was to evaluate the Louvre feasibility, completeness and adequacy based on focus group session. Another goal of this study was to evolve the framework. Under this motivation, we were interested in detecting the improvements and gaps. So, as a result of this study, we evolved the model to include the most important considerations and suggestions pointed out by the participants. This new version was presented in Chapter 5.

Summarizing the focus group analysis of the Louvre framework can be noticed that some processes need some improvements, ranging from nomenclature and description refinements to creation of new activities and reorganization of two dimensions. Some of the suggestions were included, while others were registered to be considered later, serving as basis for future work related to the model. Additionally, the participants did not agree in all suggestions.

The major improvements suggested are related to the specification of inputs for each processes and the provision of guidelines. We accept the first suggestion. In this sense, the structure of the Louvre was refined to define a set of inputs for each process. As well as, Work Products were also specified for each process. A Work Product produced by a process can be used as input by another process. Both work products and input are presented in Appendix C.

Regarding the provision of guidelines, we partially accepted this suggestion. Although we recognize the importance of guidelines, we believe that the proposition of guidelines is beyond the scope of this thesis. These guidelines could be defined by another research work. For instance, a maturity model for metadata curation based on Louvre processes would provide guidelines to implement a metadata curation initiative. Another alternative would be the creation of Louvre-based metadata curation processes, *i.e.*, Louvre implementations for specific contexts. Meanwhile, in order to ease the creation of metadata curation initiatives based on the Louvre, we introduced the Deployment Process, which was presented in Chapter 5. This Deployment Process recommends a set of steps to guide how a metadata curation initiative can be undertaken.

6.3 CONSIDERATIONS OF THE CHAPTER

This Chapter presented the definition, planning, operation, analysis and interpretation of two empirical studies that evaluated the feasibility, completeness and adequacy of the Louvre framework. In general, the Louvre had a good acceptance. The evaluation has shown that the Louvre is complete and adequate, and can be feasible, according to the opinion of the participants of both studies. In fact, these results represent a positive evidence in relation to the Louvre. Both studies also identified some directions for improvements. We tried to include as many as these as possible in the version of the framework presented in this thesis.

However, it is still not possible to guarantee that the proposed framework effectively achieves the desired results. This confirmation can only be obtained in further researches through the implementation and execution of the Louvre in scenarios of varied nature and of different sizes. Only then it will be possible to observe the real practical benefits achieved by the application of the Louvre.

Next Chapter presents the final considerations, research contributions and future work of this thesis.

7 CONCLUSION AND FUTURE WORKS

In this work, the metadata curation problem was presented. While state-of-the-art works had already proven the benefit of digital curation in many areas, in Data Ecosystems the majority of actors simply publishes or catalogs datasets and fails to cover other important aspects related to digital curation. In many cases, the metadata is forgotten or not well managed. In addition, the most of existing curation models not specify how metadata curation should be performed. Only a few works recommend a process to manage metadata. However, even these works either classify the metadata management as an atomic process or under specify metadata management. Disregarding the complex nature of metadata curation may harm the creation and progress of metadata curation initiatives.

From the literature review in the beginning of this research to the present moment, this work proposed the first metadata curation framework as well as the first meta-model for describing Data Ecosystems.

This thesis begins with the presentation of its general context, the research problem and its relevance to the area. In addition, the objectives and methodology applied to the research are also presented. Then, a review of the literature is presented, in which the central themes that involve the work are presented: metadata curation and Data Ecosystems. An informal review covered the metadata curation related work. A systematic mapping analyzed Data Ecosystem research field. Both literature reviews consolidated a rich and consistent theoretical material. This material supported the execution of the other research stages.

After the theoretical chapters, this doctoral thesis presented a meta-model for the description of Data Ecosystems. This meta-model allows to identify the main elements of Data Ecosystems and how they are related. Next, this doctoral thesis presents a metadata curation framework called the Louvre. The framework was structured based on a varied of works, such as the international standard ISO/IEC 12207. The Louvre adapts and recommends sets of practices and roles to support metadata curation work. Both practices and roles were inspired by agile methodologies. Thus, it also recommends a PDCA cycle-based deployment process to guide the creation and evolution of metadata curation initiatives. In addition, the Louvre proposes six dimensions, through which its processes are distributed. All processes were structured in terms of their purposes, activities, inputs and outcomes.

Finally, this thesis also presented evaluations of the meta-model and framework. The evaluation results have contributed to the evolution of the proposed framework since it also collected feedback from experts and practitioners in Data Ecosystems and data management. During this evaluation, it was possible to identify improvements in several aspects of the Louvre framework. In particular, there were suggestions for improvements

regarding clarity and organization of processes and also recommendations to create new processes, activities and agile practices. Finally, the participants pointed out the framework as promising and relevant.

7.1 RESEARCH CONTRIBUTIONS

The main contributions of this thesis can be classified into five aspects: (i) the presentation of more specifics surveys on the state of the art on metadata curation and Data Ecosystems; (ii) the formalization and definition of Data Ecosystems; (iii) the development of meta-model for describing Data Ecosystems; (iv) the construction of a metadata curation framework; (v) in addition, a case study, a survey and a focus group studies performed to evaluate the artifacts proposed.

- **Metadata Curation Literature Review**: Initially the state of the art on the data and metadata management was investigated. The study analyzed the available solutions and their open issues. Then, we investigated the digital curation area, by analyzing the main approaches for curating data and metadata in order to identify their fundamentals as well as to define a base of requirements for the approach defined in this work. The identified data management frameworks and curation models were also used as basis to construct the Louvre Framework.

- **Data Ecosystem Systematic Mapping**: A systematic mapping of the state of the art on Data Ecosystems was presented highlighting their origins, principles, features and issues. It was also discussed the Data Ecosystem landscape taken into account a few explanations of what a Data Ecosystem is, how it behaves, what that implies for the design and management of Data Ecosystems and what factors need to be considered in order for a Data Ecosystem to succeed. The systematic mapping findings have been also taken into account during the development of a meta-model and the Louvre framework.

- **Data Ecosystem Meta-Model**: A meta-model for describing Data Ecosystems was proposed. This meta-model was based on set of constructs identified from literature.

- **Louvre Framework**: The Louvre framework was proposed. Its set of roles, agile practices, dimensions and processes were based on the state of the art on data management and digital curation, and on agile practices.

- **Validation of Meta-Model and Louvre**: Three empirical studies were performed in order to evaluate both meta-model and Louvre. The first one evaluated the meta-model through a case study in a real Data Ecosystem. The other two studies conducted through a survey based on expert opinion and a focus group evaluated Louvre

elements, *i.e.*, agile practices, roles, dimensions, processes and their activities and outcomes. These studies also identified some gaps and mistakes in the framework organization and specification.

During the research of the Louvre framework, we also took up challenges on conceiving of a Platform Layer by proposing two solutions to be used as a start point (*i.e.*, Waldo and DataCollector). We also collaborated to the construction of other two solutions (*i.e.*, DataFeed and DWMS). These four solutions are presented in the following:

- Waldo (OLIVEIRA; GAMA; LÓSCIO, 2015; TAVARES; ; LÓSCIO, 2016; OLIVEIRA; LóSCIO; GAMA, 2015): Waldo is a cataloging solution for storing and maintaining metadata. Despite the focus on smart cities and IoT, the flexible metadata used by Waldo can be extended to cataloging a different kind of metadata.

- DataCollector (OLIVEIRA et al., 2016b; OLIVEIRA et al., 2016a): DataCollector is a solution for collecting metadata from multiple sources and transforming them in a uniform and standardized metadata schema to represent them. Therefore, the DataCollector aims to extract, normalize and process the metadata related to datasets and their producers.

- DataFeed (SANTOS; OLIVEIRA; LÓSCIO, 2017): DataFeed is a tool for collecting feedback about data published on the Web and for making the feedback freely available, thereby creating a channel of communication between data consumers and producers.

- Data on the Web Management System (DWMS) (OLIVEIRA et al., 2018): DWMS is a reference model that describes a collection of services to help data sharing on the Web. In general, such services aim to facilitate the definition, creation, maintenance, manipulation, and sharing of datasets on the Web across multiple users and applications.

This thesis generated also a book, two chapters of books, an award and several publications in national and international qualified periodicals, symposiums and workshops, as presented in Appendix F. This research also contributed to strength of the Aladin research group in the Informatics Center of the Federal University of Pernambuco

7.2 LIMITATIONS

Some limitations could be observed in the study, even with all the mitigation promoted by the researchers. First of all, it was not easy to find researchers and practitioners willing to participate with the evaluation studies. As a consequence, we cannot create a major representative population sample. We only achieved the minimum valid statistic for validate the work.

In relation to the research method, the limitations are typical of empirical studies, particularly in the generalization of the results. Data extraction and analysis could be biased by the personal opinions of researcher executing the process. To mitigate this threat, we dedicated adequate time to performing several refinement-evaluation cycles for construct the proposed artifacts. Another mitigation action was to frequently consult other researchers from Aladdin research group to address and resolve conflicts that came with research progress. Finally, whenever possible, we consolidated the partial results as papers and submitted then to conferences in order to collect external feedback.

In relation to the systematic review, one of the greatest concerns is related to the selection of relevant studies. We defined research questions in advance and devised the inclusion and exclusion criteria in order to ensure an unbiased selection. However, some important fundamental works might be excluded. This threat was mitigated by selecting different terms to represent Data Ecosystems as well as by not using any filter predicate in the search query string used in the automatic searches. Moreover, the search for studies was conducted in five search engines, even though it is possible we missed some relevant studies. Nevertheless, this threat was mitigated by selecting the search engines which have been considered the most relevant scientific sources for the computer science community and therefore prone to containing the majority of important studies. We also looked for related studies referenced from studies already in the pool, in order to decrease the risk of missing relevant studies.

With regards to the meta-model, more studies must be performed with different Data Ecosystems in order to verify the adequacy and completeness of the meta-model in different contexts. Moreover, it is important to test the alignment of meta-model with other meta-models and models that cover specific aspects of Data Ecosystem. Furthermore, the meta-model evaluation was influenced by the skill and knowledge of case study participants who participated in the evaluation. There is a need for a guidelines to guide the use of the meta-model.

7.3 FUTURE WORK

The focus of this thesis was to define a metadata curation framework for Data Ecosystems. However, the framework is not applicable under all circumstances and in every context. As future works, we intend perform new empirical studies to implement and evaluate the proposed framework, considering multiples scenarios (*i.e.*, different Data Ecosystems). Such studies would generate new empirical data and also collecting more improvements issues for the framework, as well as obtaining greater validation of the framework. The same applies for the meta-model, which still lacks of more validation.

We also intend to propose a maturity model based on the Louvre processes and elements. Maturity Models became popular from when the Capability Maturity Model (CMM)(PAULK et al., 1993) was first proposed. According to Fisher (2004), maturity mod-

els are used to compare and evaluate improvements, thus allowing the degree of evolution in certain domains to be measured. In the business environment, they aim to help organizations identify ways to improve the quality of their processes and reduce their execution time, thereby providing them with competitive advantages. In metadata curation context, maturity models can be applied to provide clear recommendations on how to drive improvements based on current context in which a metadata curation initiative lies.

Moreover, we intend to organize workshops to promote the use of the Louvre framework. These workshop are meant to be short-time case studies. Thus, the workshop participants can provide new insight and feedback to improve and refine the framework.

In order to contribute to digital curation field, we intend integrate the Louvre framework with both other digital curation models and data management. Such integration is aligned with a major requirement of Louvre, which advocates for Louvre to be used as complement to other solutions available in the literature.

Another future work is to develop a CASE (Computer-aided software engineering) tool to help practitioners to construct concrete models derived from our meta-model. In fact, such CASE tool would also help to demonstrate the feasibility, the expressiveness and usefulness of our meta-model.

Considering the Data Ecosystem scope, we identified many opportunities to the continuity of the developed studies. For instance, most of the studies present and analyze potential business models to define how an ecosystem creates and delivers value for the actors. Despite being important, these kinds of models cover only a small fragment of how a Data Ecosystem works. There is also a lack of modeling languages for representing Data Ecosystems at a high level of abstraction. For instance, business actors, or even technical users, may face some difficulties when evaluating a business model or the structural organization of a Data Ecosystem. A new modeling language based on a graphical notation can provide the capability of understanding important aspects and processes of a Data Ecosystem as well as giving actors the ability to communicate these aspects and processes in a standard manner.

Assessment models and solutions for validating the health of Data Ecosystems is another gap in the Data Ecosystem literature. These models should provide the means to evaluate the functionality and status of elements in a Data Ecosystem. The health of an ecosystem depends in part on a variety of factors, including the actors and how they act, relationships, policies, and the infrastructure available. In fact, part of defining effectiveness and success for a Data Ecosystem lies in determining and utilizing metrics to measure its health.

Developing engineering methods for effective Data Ecosystems governance and control is another gap in the Data Ecosystem research. Data Ecosystem functioning depends on the activity and interaction of a set of different actors. Such diffuse performance comes at the expense of decreased control and the resulting increase in challenges associated

with planning and maintenance. Therefore, Data Ecosystem engineering methods supply a common structure in the form of well-defined rules, procedures, protocols and processes to develop, manage and evolve Data Ecosystems. Engineering methods can also increase the ecosystem's health.

Additional work would be needed to formulate guidelines for different Data Ecosystems domains. Future work that needs to be carried out to develop methodologies for Data Ecosystem. These methodologies should provide the responsible actors with clear guidance on how the Data Ecosystem initiatives should be implemented and how the known challenges and risks should be addressed. If the challenges are not properly tackled, it might prevent the generation of expected benefits. Furthermore, a proper methodology must drive to a clear project with well-defined objectives and actions, as well as clear planning.

The management and coordination of Data Ecosystems is another research gap. A management framework can be defined as a set of various methods to discover, model, analyze and measure the coordination of actors to improve their activities as well as to deliver more benefits. Specifying a management framework faces some fundamental challenges. One of these is the link between the plethora of Data Ecosystem elements and their contribution to value generation. Moreover, a Data Ecosystem management framework must encompass regulations applicable to industry, policies and quality standards. In short, the distributed and socio-technical nature of Data Ecosystem requires a systematic, holistic approach to management that aligns actors' capabilities and resources to their needs.

There are several technical/non-technical challenges related to using data in a Data Ecosystem, including the complexity of activities needed to identify, understand, and use data, lack of capabilities and technical knowledge among actors (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012; ZUIDERWIJK; JANSSEN, 2014). Additional challenges encompass problems with the provenance of data, data management, and quality (*e.g.*, validity, completeness and timeliness), metadata provision, and interoperability, as well as concerns for privacy and confidentiality (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012; ZUIDERWIJK; JANSSEN, 2014). Generally speaking, the proposal of solutions to address some of these challenges may ease the burden on actors, mainly on data consuming actors, and consequently promote their participation in Data Ecosystems.

REFERENCES

ABBOTT, D. *What is digital curation*. 2013. <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation/>. Accessed on 20 December, 2018.

ABOWD, G. D.; DEY, A. K.; BROWN, P. J.; DAVIES, N.; SMITH, M.; STEGGLES, P. Towards a better understanding of context and context-awareness. In: SPRINGER. *International symposium on handheld and ubiquitous computing*. [S.l.], 1999. p. 304–307.

ABU-MATAR, M. Towards a software defined reference architecture for smart city ecosystems. In: IEEE. *Smart Cities Conference (ISC2), 2016 IEEE International*. [S.l.], 2016. p. 1–6.

AGILE ALLIANCE. *Agile Glossary*. [S.l.]: AGILE ALLIANCE, 2018. <https://www.agilealliance.org/glossary/>. Accessed on 20 December, 2018.

ALAN, R. H. V.; MARCH, S. T.; PARK, J.; RAM, S. Design science in information systems research. *MIS quarterly*, Springer, v. 28, n. 1, p. 75–105, 2004.

ALBUQUERQUE, C. A. M. d. *Qualidade ágil de software*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2005.

ALLARD, S. Dataone: Facilitating escience through collaboration. *Journal of eScience Librarianship*, v. 1, n. 1, p. 3, 2012.

ALMEIDA, H. R.; MAGALHÃES, E. M. C. de; MOURA, H. P. de; FILHO, J. G. d. A. T.; CAPPELLI, C.; MARTINS, L. M. F. Evaluation of a maturity model for agile governance in ict using focus group. In: BRAZILIAN COMPUTER SOCIETY. *Proceedings of the annual conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective-Volume 1*. [S.l.], 2015. p. 3.

ALMEIDA, H. R. d. *Um modelo de maturidade para governança ágil em tecnologia da informação e comunicação*. Tese (Doutorado) — Universidade Federal de Pernambuco, 2015.

AMBLER, S.; LINES, M. Disciplined agile delivery: an introduction. *IBM Software, Somers, NY*, 2011.

ARKSEY, L. O. H. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, v. 8, p. 19–32, 2005.

ATKINSON, C.; KUHNE, T. Model-driven development: a metamodeling foundation. *IEEE software*, IEEE, v. 20, n. 5, p. 36–41, 2003.

ATTARD, J.; ORLANDI, F.; AUER, S. Data value networks: Enabling a new data ecosystem. In: IEEE. *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*. [S.l.], 2016. p. 453–456.

BALL, A. *Review of the State of the Art of the Digital Curation of Research Data*. [S.l.]: University of Bath, 2012. <http://opus.bath.ac.uk/18774/2/erim1rep091103ab11.pdf>. Accessed on 20 December, 2018.

BARBOSA, L.; PHAM, K.; SILVA, C.; VIEIRA, M. R.; FREIRE, J. Structured open urban data: understanding the landscape. *Big data*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 2, n. 3, p. 144–154, 2014.

BARNAGHI, P.; SHETH, A.; HENSON, C. From data to actionable knowledge: Big data challenges in the web of things. *IEEE Intelligent Systems*, IEEE, v. 28, n. 6, p. 6–11, 2013.

BEAN, R. *Variety, Not Volume, Is Driving Big Data Initiatives*. 2016. <https://sloanreview.mit.edu/article/variety-not-volume-is-driving-big-data-initiatives/>. Accessed on 20 December, 2018.

BECK, K.; BEEDLE, M.; BENNEKUM, A. V.; COCKBURN, A.; CUNNINGHAM, W.; FOWLER, M.; GRENNING, J.; HIGHSMITH, J.; HUNT, A.; JEFFRIES, R. et al. *Manifesto for agile software development*. 2001. <https://agilemanifesto.org/>. Accessed on 20 December, 2018.

BELHAJJAME, K.; WOLSTENCROFT, K.; CORCHO, O.; OINN, T.; TANOH, F.; WILLIAM, A.; GOBLE, C. Metadata management in the taverna workflow system. In: IEEE. *Cluster Computing and the Grid, 2008. CCGRID'08. 8th IEEE International Symposium on*. [S.l.], 2008. p. 651–656.

BERCZUK, S. Back to basics: The role of agile principles in success with an distributed scrum team. In: IEEE. *Agile Conference (AGILE), 2007*. [S.l.], 2007. p. 382–388.

BISHOP, W.; GRUBESIC, T. H. Data lifecycle. In: SPRINGER INTERNATIONAL PUBLISHING. *Geographic Information: Organization, Access, and Use*. Cham: Springer International Publishing, 2016. p. 169–186. ISBN 978-3-319-22789-4. Disponível em: <http://dx.doi.org/10.1007/978-3-319-22789-4_9>.

BOSCH, T.; CYGANIAK, R.; WACKEROW, J.; ZAPILKO, B. Leveraging the ddi model for linked statistical data in the social, behavioural, and economic sciences. In: *International Conference on Dublin Core and Metadata Applications*. [S.l.: s.n.], 2012. p. 46–55.

BOUCHARAS, V.; JANSEN, S.; BRINKKEMPER, S. Formalizing software ecosystem modeling. In: ACM. *Proceedings of the 1st international workshop on Open component ecosystems*. [S.l.], 2009. p. 41–50.

BOURNE, P. E.; LORSCH, J. R.; GREEN, E. D. Perspective: Sustaining the big-data ecosystem. *Nature*, Nature Research, v. 527, n. 7576, p. S16–S17, 2015.

BRAMBILLA, M.; CABOT, J.; WIMMER, M. Model-driven software engineering in practice. *Synthesis Lectures on Software Engineering*, Morgan & Claypool Publishers, v. 1, n. 1, p. 1–182, 2012.

BUDHATHOKI, N. R.; HAYTHORNTHWAITE, C. Motivation for open collaboration: Crowd and community models and the case of openstreetmap. *American Behavioral Scientist*, Sage Publications Sage CA: Los Angeles, CA, v. 57, n. 5, p. 548–575, 2013.

BURTON, A.; TRELOAR, A. Designing for discovery and re-use: The 'ands data sharing verbs' approach to service decomposition. *International Journal of Digital Curation*, v. 4, n. 3, p. 44–56, 2009.

CARLSON, J. The use of life cycle models in developing and supporting data services. *Research data management: Practical strategies for information professionals*, Purdue University Press West Lafayette, IN, p. 63–86, 2014.

CARROLL, J. M. Five reasons for scenario-based design. *Interacting with computers*, Oxford University Press, v. 13, n. 1, p. 43–60, 2000.

CHEN, M.; MAO, S.; LIU, Y. Big data: a survey. *Mobile Networks and Applications*, Springer, v. 19, n. 2, p. 171–209, 2014.

CHESSELL, M. *The case for open metadata and governance*. 2016. <http: //www.ibmbigdatahub.com/blog/insightout-case-open-metadata-and-governance>. Accessed on 20 December, 2018.

CHRISTENSEN, H. B.; HANSEN, K. M.; KYNG, M.; MANIKAS, K. Analysis and design of software ecosystem architectures–towards the 4S telemedicine ecosystem. *Information and Software Technology*, Elsevier, v. 56, n. 11, p. 1476–1492, 2014.

CHUN, S. A.; SHULMAN, S.; SANDOVAL, R.; HOVY, E. Government 2.0: Making connections between citizens, data and government. *Information Polity*, v. 15, n. 1, p. 1, 2010.

CLEGG, C. W. Sociotechnical principles for system design. *Applied Ergonomics*, Elsevier, v. 31, n. 5, p. 463–477, 2000.

COCKBURN, A. *Agile software development*. [S.l.]: Addison-Wesley Boston, 2002. v. 177.

COMMITTEE, J. I. S. et al. *An Invitation for Expressions of Interest to Establish a New Digital Curation Centre for Research intoand Support of the Curation and Preservation of Digital Data and Publications*. 2003.

CONBOY, K.; FITZGERALD, B. Method and developer characteristics for effective agile method tailoring: A study of xp expert opinion. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, ACM, v. 20, n. 1, p. 2, 2010.

COOPER, A.; REIMANN, R.; CRONIN, D. Modeling users: Personas and goals. *Erschienen in About Face*, v. 2, p. 55–74, 2003.

CORTI, L.; EYNDEN, V. Van den; BISHOP, L.; WOOLLARD, M. *Managing and sharing research data: a guide to good practice*. [S.l.]: Sage, 2014.

COX, A. M.; TAM, W. W. T. A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, Emerald Publishing Limited, v. 70, n. 2, p. 142–157, 2018.

CRESWELL, J. W. Projeto de pesquisa métodos qualitativo, quantitativo e misto. In: *Projeto de pesquisa métodos qualitativo, quantitativo e misto*. [S.l.]: Artmed, 2010.

CRISTANI, M.; CUEL, R. A survey on ontology creation methodologies. *International Journal on Semantic Web and Information Systems (IJSWIS)*, IGI Global, v. 1, n. 2, p. 49–69, 2005.

CROWSTON, K.; QIN, J. A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the American Society for Information Science and Technology*, Wiley Online Library, v. 48, n. 1, p. 1–9, 2011.

CURRY, E.; FREITAS, A.; O'RIÁIN, S. The role of community-driven data curation for enterprises. In: *Linking enterprise data*. [S.l.]: Springer, 2010. p. 25–47.

DALLAS, C. Digital curation beyond the "wild frontier": A pragmatic approach. *Archival Science*, Springer, v. 16, n. 4, p. 421–457, 2016.

DANIEL, E. M.; WILSON, H. N. The role of dynamic capabilities in e-business transformation. *European Journal of Information Systems*, Springer, v. 12, n. 4, p. 282–296, 2003.

Data World. *Data Science and Open Data By the Numbers*. 2018. <https://data.world/data-science-by-the-numbers>. Accessed on 20 December, 2018.

DAVIES, T. Open data: infrastructures and ecosystems. *Open Data Research*, 2011.

DAWES, S. S.; VIDIASOVA, L.; PARKHIMOVICH, O. Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, Elsevier, v. 33, n. 1, p. 15–27, 2016.

DDI. *Data Documentation Initiative Specification*. [S.l.]: Data Documentation Initiative, 2012. <http://www.ddialliance.org/Specification>. Accessed on 20 December, 2018.

DIEBOLD, P. *List of Agile Practices*. 2015. <https://www.researchgate.net/publication/281638090_List_of_Agile_Practices>. Accessed on 20 December, 2018.

DING, L.; LEBO, T.; ERICKSON, J. S.; DIFRANZO, D.; WILLIAMS, G. T.; LI, X.; MICHAELIS, J.; GRAVES, A.; ZHENG, J. G.; SHANGGUAN, Z. et al. Twc logd: A portal for linked open government data ecosystems. *Web Semantics: Science, Services and Agents on the World Wide Web*, Elsevier, v. 9, n. 3, p. 325–333, 2011.

DINTER, B.; SCHIEDER, C.; GLUCHOWSKI, P. A stakeholder lens on metadata management in business intelligence and big data – results of an empirical investigation. In: *Proceedings of the American Conference on Information Systems (AMCIS'2015)*. Puerto Rico: Association for Information Systems, 2015. p. 1–12.

DONKER, F. W.; LOENEN, B. van. How to assess the success of the open data ecosystem? *International Journal of Digital Earth*, Taylor & Francis, v. 10, n. 3, p. 284–306, 2017.

EASTERBROOK, S.; SINGER, J.; STOREY, M.-A.; DAMIAN, D. Selecting empirical methods for software engineering research. Springer, p. 285–311, 2008.

FARIAS, I. H. *C2M - A Communication Maturity Model for Distributed Software Development*. Tese (Doutorado) — Universidade Federal de Pernambuco, 2014.

FEICHENG, M.; JUNCHENG, W. A literature review of studies on information lifecycle: the perspective of value. *Journal of the China Society for Scientific and Technical Information*, v. 5, 2010.

FERNANDEZ, D. J.; FERNANDEZ, J. D. Agile project management—agilism versus traditional approaches. *Journal of Computer Information Systems*, Taylor & Francis, v. 49, n. 2, p. 10–17, 2008.

FISCHER, G.; HERRMANN, T. Socio-technical systems: a meta-design perspective. *IGI Global*, p. 1–33, 2011.

FISHER, D. M. The business process maturity model: A practical approach for identifying opportunities for optimization. *BPTrends*, Sept 2004.

FONSECA, A. V. da; MIYAKE, D. I. Uma análise sobre o ciclo pdca como um método para solução de problemas da qualidade. *XXVI Encontro Nacional de Engenharia de Produção, pages 1-9, Fortaleza, CE*, 2006.

FOREHAND, M. Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, v. 41, p. 47, 2010.

FORTE, A.; LAMPE, C. Defining, understanding, and supporting open collaboration: Lessons from the literature. *American Behavioral Scientist*, Sage Publications Sage CA: Los Angeles, CA, v. 57, n. 5, p. 535–547, 2013.

FOWLER, M. *Continuous Integration*. [S.l.]: Fowler, Martin, 2006. <https://martinfowler.com/articles/continuousIntegration.html>. Accessed on 20 December, 2018.

FOWLER, M.; BECK, K.; BRANT, J.; OPDYKE, W.; ROBERTS, D. *Refactoring: improving the design of existing code*. [S.l.]: Addison-Wesley Professional, 1999.

FREITAS, A.; CURRY, E. Big data curation. In: *New Horizons for a Data-Driven Economy*. [S.l.]: Springer, 2016. p. 87–118.

GAMA, K.; LÓSCIO, B. F. Towards ecosystems based on open data as a service. In: *ICEIS (2)*. [S.l.: s.n.], 2014. p. 659–664.

GARCÍA-CASTRO, R.; CORCHO, O.; HILL, C. A core ontological model for semantic sensor web infrastructures. *International Journal on Semantic Web and Information Systems*, IGI Global, v. 8, n. 1, p. 22–42, 2012.

GARCIA, V. C. *RiSE reference model for software reuse adoption in Brazilian companies*. Tese (Doutorado) — Universidade Federal de Pernambuco, 2010.

GERMAN, D. M.; ADAMS, B.; HASSAN, A. E. The evolution of the r software ecosystem. In: IEEE. *2013 17th European Conference on Software Maintenance and Reengineering*. [S.l.], 2013. p. 243–252.

GIL, A. C. *Métodos e técnicas de pesquisa social*. [S.l.]: 6. ed. Ediitora Atlas SA, 2008.

GLAIEL, F. S.; MOULTON, A.; MADNICK, S. E. Agile project dynamics: A system dynamics investigation of agile software development methods. Massachusetts Institute of Technology. Engineering Systems Division, 2014.

GOBLE, C.; STEVENS, R.; HULL, D.; WOLSTENCROFT, K.; LOPEZ, R. Data curation+ process curation= data integration+ science. *Briefings in bioinformatics*, Oxford Univ Press, v. 9, n. 6, p. 506–517, 2008.

GRGIĆ, V. *Descaling Organizations with LeSS*. 2015. <https://less.works/blog/2015/05/08/less-scaling-descaling-organizations-with-less.html>. Accessed on 20 December, 2018.

GROVES, R. M.; JR, F. J. F.; COUPER, M. P.; LEPKOWSKI, J. M.; SINGER, E.; TOURANGEAU, R. *Survey methodology*. [S.l.]: John Wiley & Sons, 2011. v. 561.

GRUNZKE, R.; MULLER-PFEFFERKORN, R.; JAKEL, R.; HESSER, J.; KEPPER, N.; HAUSMANN, M.; STAREK, J.; GESING, S.; HARDT, M.; HARTMANN, V. et al. Device-driven metadata management solutions for scientific big data use cases. In: IEEE. *Parallel, Distributed and Network-Based Processing (PDP), 2014 22nd Euromicro International Conference on*. [S.l.], 2014. p. 317–321.

HA, S.; LEE, S.; LEE, K. Standardization requirements analysis on big data in public sector based on potential business models. *International Journal of Software Engineering and Its Applications*, v. 8, n. 11, p. 165–172, 2014.

HANSSEN, G. K. A longitudinal case study of an emerging software ecosystem: Implications for practice and theory. *Journal of Systems and Software*, Elsevier, v. 85, n. 7, p. 1455–1466, 2012.

HANSSEN, G. K.; DYBÅ, T. Theoretical foundations of software ecosystems. In: CITESEER. *IWSECO@ ICSOB*. [S.l.], 2012. p. 6–17.

HARRISON, T. M.; PARDO, T. A.; COOK, M. Creating open government ecosystems: A research and development agenda. *Future Internet*, MDPI AG, v. 4, n. 4, p. 900, 2012.

HASSELL, D. *Open Communication: Vital to Business Success*. [S.l.]: American Management Association(AMA), 2018. <https://www.amanet.org/training/articles/open-communication-vital-to-business-success.aspx>. Accessed on 20 December, 2018.

HEIMSTÄDT, M.; SAUNDERSON, F.; HEATH, T. Conceptualizing open data ecosystems: A timeline analysis of open data development in the UK. In: MV-VERLAG. *CeDEM14: Conference for E-Democracy an Open Government*. [S.l.], 2014. p. 245.

HELBIG, N.; CRESSWELL, A. M.; BURKE, G. B.; LUNA-REYES, L. The dynamics of opening government data. *Center for Technology in Government.[Online]. Available: http://www. ctg. albany. edu/publications/reports/opendata*, 2012.

HEVNER, A. R. A three cycle view of design science research. *Scandinavian journal of information systems*, v. 19, n. 2, p. 4, 2007.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. Design science in information systems research. *Management Information Systems Quarterly*, v. 28, n. 1, p. 6, 2008.

HIGGINS, S. The dcc curation lifecycle model. *International journal of digital curation*, v. 3, n. 1, p. 134–140, 2008.

HIGHSMITH, J. R. *Agile project management: creating innovative products*. [S.l.]: Pearson Education, 2009.

IANSITI, M.; LEVIEN, R. *The keystone advantage: what the new dynamics of business ecosystems mean for strategy, innovation, and sustainability.* [S.l.]: Harvard Business Press, 2004.

IMMONEN, A.; PALVIAINEN, M.; OVASKA, E. Requirements of an open data based business ecosystem. *IEEE Access*, IEEE, v. 2, p. 88–103, 2014.

ISO. *ISO 8000 – the international standard for data quality.* 2011. Accessed on 20 December, 2018. Disponível em: <https://www.iso.org/standard/50798.html>.

ISO. *ISO 14721 – Space data and information transfer systems – Open archival information system (OAIS) – Reference model.* 2012. Accessed on 20 April, 2019. Disponível em: <https://www.iso.org/standard/57284.html>.

ISO, I. *IEEE 12207-2008 - ISO/IEC/IEEE International Standard - Systems and software engineering – Software life cycle processes.* 2008. Accessed on 20 December, 2018. Disponível em: <https://www.iso.org/standard/43447.html>.

JALALI, S.; WOHLIN, C. Systematic literature studies: database searches vs. backward snowballing. In: ACM. *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement.* [S.l.], 2012. p. 29–38.

JANSEN, S.; BRINKKEMPER, S.; FINKELSTEIN, A. Business network management as a survival strategy: A tale of two software ecosystems. In: *Proceedings of the 1st International Workshop on Software Ecosystems.* [S.l.: s.n.], 2009. p. 34.

JANSEN, S.; BRINKKEMPER, S.; FINKELSTEIN, A. Business network management as a survival strategy: A tale of two software ecosystems. *Iwseco@ Icsr*, v. 2009, 2009.

JANSEN, S.; BRINKKEMPER, S.; SOUER, J.; LUINENBURG, L. Shades of gray: Opening up a software producing organization with the open software enterprise model. *Journal of Systems and Software*, Elsevier, v. 85, n. 7, p. 1495–1510, 2012.

JANSEN, S.; CUSUMANO, M. A.; BRINKKEMPER, S. *Software ecosystems: analyzing and managing business networks in the software industry.* [S.l.]: Edward Elgar Publishing, 2013.

JANSSEN, M.; CHARALABIDIS, Y.; ZUIDERWIJK, A. Benefits, adoption barriers and myths of open data and open government. *Information systems management*, Taylor & Francis, v. 29, n. 4, p. 258–268, 2012.

JANZEN, D.; SAIEDIAN, H. Test-driven development concepts, taxonomy, and future direction. *Computer*, IEEE, v. 38, n. 9, p. 43–50, 2005.

JONES, K. Research360: Managing data across the institutional research lifecycle. In: *Poster Presented the 7th International Digital Curation Conference, Bristol, UK, 5–8 Dec.* [S.l.: s.n.], 2011.

KAREL, R.; MINES, C.; WANG, R.; MCNABB, K.; BARNETT, J. Introducing master data management. *Forrester Research, Cambridge*, 2006.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing systematic literature reviews in software engineering.* [S.l.], 2007.

KITCHENHAM, B.; PFLEEGER, S. L.; MCCOLL, B.; EAGAN, S. An empirical study of maintenance and development estimation accuracy. *Journal of systems and software*, Elsevier, v. 64, n. 1, p. 57–77, 2002.

KITCHENHAM, B. A.; PFLEEGER, S. L. Principles of survey research part 2: designing a survey. *ACM SIGSOFT Software Engineering Notes*, ACM, v. 27, n. 1, p. 18–20, 2002.

KONTIO, J.; BRAGGE, J.; LEHTOLA, L. The focus group method as an empirical tool in software engineering. In: *Guide to advanced empirical software engineering*. [S.l.]: Springer, 2008. p. 93–116.

KONTIO, J.; LEHTOLA, L.; BRAGGE, J. Using the focus group method in software engineering: obtaining practitioner and user experiences. In: IEEE. *Empirical Software Engineering, 2004. ISESE'04. Proceedings. 2004 International Symposium on*. [S.l.], 2004. p. 271–280.

KÖSTER, V.; SUÁREZ, G. Open data for development: Experience of uruguay. In: ACM. *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*. [S.l.], 2016. p. 207–210.

KOUPER, I.; KONKIEL, S. R.; LISS, J. A.; HARDESTY, J. L. Collaborate, automate, prepare, prioritize: Creating metadata for legacy research data. In: *International Conference on Dublin Core and Metadata Applications*. Libon, PT: Dublin Core Metadata Initiave, 2013. p. 41–46.

KOZNOV, D.; ANDREEVA, O.; NIKULA, U.; MAGLYAS, A.; MUROMTSEV, D.; RADCHENKO, I. A survey of open government data in Russian Federation. In: *IC3K 2016 - Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. [S.l.: s.n.], 2016. v. 3, p. 173–180.

KRAUT, R. E.; STREETER, L. A. Coordination in software development. *Communications of the ACM*, Association for Computing Machinery, Inc., v. 38, n. 3, p. 69–82, 1995.

KRUEGER, R. A.; CASEY, M. A. *Focus groups: A practical guide for applied research*. [S.l.]: Sage publications, 2014.

LEBIED, M. *12 Examples of Big Data Analytics In Healthcare That Can Save People*. 2018. <https://www.datapine.com/blog/big-data-examples-in-healthcare/>. Accessed on 19 March, 2019.

LEDERMAN, L. C. Assessing educational effectiveness: The focus group interview as a technique for data collection. *Communication education*, Taylor & Francis, v. 39, n. 2, p. 117–127, 1990.

LEE, C. A. Open archival information system (oais) reference model. *Encyclopedia of Library and Information Sciences*, Taylor & Francis, p. 4020–4030, 2010.

LEE, C. C.; YANG, J. Knowledge value chain. *Journal of Management Development*, MCB UP Ltd, v. 19, n. 9, p. 783–794, 2000.

LEE, D. Building an open data ecosystem: an Irish experience. In: ACM. *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*. [S.l.], 2014. p. 351–360.

LEE, S.; YONG, H.-S. Agile software development framework in a small project environment. *Journal of Information Processing Systems*, Korea Information Processing Society, v. 9, n. 1, p. 69–88, 2013.

LEE, Y. W.; STRONG, D. M.; KAHN, B. K.; WANG, R. Y. Aimq: a methodology for information quality assessment. *Information & management*, Elsevier, v. 40, n. 2, p. 133–146, 2002.

LINDMAN, J.; KINNARI, T.; ROSSI, M. Business roles in the emerging open-data ecosystem. *IEEE Software*, IEEE, v. 33, n. 5, p. 54–59, 2016.

LOPEZ-HERREJON, R. E.; LINSBAUER, L.; EGYED, A. A systematic mapping study of search-based software engineering for software product lines. *Information and software technology*, Elsevier, v. 61, p. 33–51, 2015.

LORD, P.; MACDONALD, A. *Data curation for e-science in the UK: An audit to establish requirements for future curation and provision*. 2003. <https://www.cs.york.ac.uk/ftpdir/pub/leo/york-msc-2007/information/vsr-curation/science-dc-report.pdf>. Accessed on 20 December, 2018.

LORD, P.; MACDONALD, A. *E-Science curation report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. [S.l.]: Digital Archiving Consultancy Limited, 2003.

LUDÄSCHER, B.; LIN, K.; BOWERS, S.; JAEGER-FRANK, E.; BRODARIC, B.; BARU, C. Managing scientific data: From data integration to scientific workflows*. *Geological Society of America Special Papers*, Geological Society of America, v. 397, p. 109–129, 2006.

LUNA, A. J. H. d. O. *MAnGve: Um Modelo para Governança Ágil em TIC*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2009.

LUNDELL, B.; FORSSTEN, B.; GAMALIELSSON, J.; GUSTAVSSON, H.; KARLSSON, R.; LENNERHOLT, C.; LINGS, B.; MATTSSON, A.; OLSSON, E. Exploring health within OSS ecosystems. In: *First International Workshop on Building Sustainable Open Source Communities (OSCOMM 2009), Skövde, Sweden*. [S.l.: s.n.], 2009. p. 1–5.

MADHAVAN, J.; JEFFERY, S. R.; COHEN, S.; DONG, X. L.; KO, D.; YU, C.; HALEVY, A. Web-scale data integration: You can only afford to pay as you go. *Conference on Innovative Data Systems Research*, 2007.

MAGALHAES, G.; ROSEIRA, C.; MANLEY, L. Business models for open government data. In: ACM. *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*. [S.l.], 2014. p. 365–370.

MANIKAS, K.; HANSEN, K. M. Reviewing the health of software ecosystems–a conceptual framework proposal. In: *Proceedings of the 5th International Workshop on Software Ecosystems (IWSECO)*. [S.l.: s.n.], 2013. p. 33–44.

MANIKAS, K.; HANSEN, K. M. Software ecosystems–a systematic literature review. *Journal of Systems and Software*, Elsevier, v. 86, n. 5, p. 1294–1306, 2013.

MARCONI, M. d. A.; LAKATOS, E. M. Fundamentos de metodologia cientifica. In: *Fundamentos de metodologia cientifica*. [S.l.]: Atlas, 2010.

MAUTHNER, N. S.; PARRY, O. Open access digital data sharing: Principles, policies and practices. *Social Epistemology*, Taylor & Francis, v. 27, n. 1, p. 47–67, 2013.

MCLEOD, S. *Qualitative vs. quantitative research*. 2017. <https://www.simplypsychology.org/qualitative-quantitative.html>. Accessed on 20 December, 2018.

MERCADO-LARA, E.; GIL-GARCIA, J. R. Open government and data intermediaries: the case of AidData. In: ACM. *Proceedings of the 15th Annual International Conference on Digital Government Research*. [S.l.], 2014. p. 335–336.

MISSIER, P.; ALPER, P.; CORCHO, O.; DUNLOP, I.; GOBLE, C. Requirements and services for metadata management. *IEEE Internet Computing*, IEEE Computer Society, v. 11, n. 5, p. 17, 2007.

MOISO, C.; MINERVA, R. Towards a user-centric personal data ecosystem the role of the bank of individuals' data. In: IEEE. *Intelligence in Next Generation Networks (ICIN), 2012 16th International Conference on*. [S.l.], 2012. p. 202–209.

MÖLLER, K. Lifecycle models of data-centric systems and domains. *Semantic Web*, IOS Press, v. 4, n. 1, p. 67–88, 2013.

MONIRUZZAMAN, A.; HOSSAIN, D. S. A. Comparative study on agile software development methodologies. *arXiv preprint arXiv:1307.3356*, 2013.

MOORE, J. F. Creating value in the network economy. Harvard Business School Press, Boston, MA, USA, p. 121–141, 1999. Disponível em: <http://dl.acm.org/citation.cfm?id=303444.303452>.

MOSLEY, M.; BRACKETT, M. H.; EARLEY, S.; HENDERSON, D. *DAMA guide to the data management body of knowledge*. [S.l.]: Technics Publications, 2010.

MUNRO, R. Actor-network theory. *The SAGE handbook of power. London: Sage Publications Ltd*, p. 125–39, 2009.

MURDOCH, T. B.; DETSKY, A. S. The inevitable application of big data to health care. *Jama*, American Medical Association, v. 309, n. 13, p. 1351–1352, 2013.

NACHIRA, F.; DINI, P.; NICOLAI, A. A network of digital business ecosystems for Europe: roots, processes and perspectives. *European Commission, Bruxelles, Introductory Paper*, 2007.

NUSEIBEH, B. Weaving together requirements and architectures. *Computer*, IEEE, v. 34, n. 3, p. 115–119, 2001.

OLIVEIRA, L. E. R.; OLIVEIRA, M. I. S.; SANTOS, W. C. d. R.; LÓSCIO, B. F. Data on the web management system: a reference model. In: ACM. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. [S.l.], 2018. p. 2.

OLIVEIRA, L. E. R. de A.; OLIVEIRA, M. I. S.; LÓSCIO, B. F. Um survey sobre soluções para publicação de dados na web sob a perspectiva das boas práticas do w3c. In: *32rd Brazilian Symposium on Databases*. [S.l.: s.n.], 2017.

OLIVEIRA, M. I. S.; GAMA, K. S. da; LÓSCIO, B. F. Waldo: Serviço para publicação e descoberta de produtores de dados para middleware de cidades inteligentes. *XI Simpósio Brasileiro de Sistemas de Informação*, 2015.

OLIVEIRA, M. I. S.; LIMA, G. d. F. B.; LÓSCIO, B. F. Investigations into data ecosystems: a systematic mapping study. *Knowledge and Information Systems*, Springer, p. 1–42, 2019.

OLIVEIRA, M. I. S.; LÓSCIO, B. F. Metadata curation framework for supporting data ecosystems. In: *Proceedings of Workshop on Thesis and Dissertations at 32rd Brazilian Symposium on Databases*. [S.l.: s.n.], 2017.

OLIVEIRA, M. I. S.; LÓSCIO, B. F. What is a data ecosystem? In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. New York, NY, USA: ACM, 2018. (dg.o '18), p. 74:1–74:9. ISBN 978-1-4503-6526-0. Disponível em: <http://doi.acm.org/10.1145/3209281.3209335>.

OLIVEIRA, M. I. S.; LÓSCIO, B. F. Louvre: A framework for metadata curation in data ecosystem. In: SBC. *Submmited to : 15rd Brazilian Symposium on Information Systems*. [S.l.], 2019.

OLIVEIRA, M. I. S.; LóSCIO, B. F.; GAMA, K. S. da. Análise de desempenho de catálogo de produtores de dados para internet das coisas baseado em sensorml e nosql. *XIV Workshop em Desempenho de Sistemas Computacionais e de Comunicaçção*, 2015.

OLIVEIRA, M. I. S.; OLIVEIRA, H. R. de; OLIVEIRA, L. A.; LÓSCIO, B. F. Open government data portals analysis: The brazilian case. In: ACM. *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*. [S.l.], 2016. p. 415–424.

OLIVEIRA, M. I. S.; OLIVEIRA, L. A.; LIMA, G. F. B.; LÓSCIO, B. F. Enabling a unified view of open data catalogs. In: *18th International Conference on Enterprise Information Systems (ICEIS)*. [S.l.: s.n.], 2016.

OLIVEIRA, M. I. S.; OLIVEIRA, L. E. R. A.; BATISTA, M. G. R.; LóSCIO, B. F. Towards a meta-model for data ecosystems. In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. New York, NY, USA: ACM, 2018. (dg.o '18), p. 72:1–72:10. ISBN 978-1-4503-6526-0. Disponível em: <http://doi.acm.org/10.1145/3209281.3209333>.

OMG. *Meta Object Facility (MOF) Core Specification v2.5.1*. [S.l.]: Object Management Group, 2016. <http://www.omg.org/spec/MOF/2.5.1/PDF>. Accessed on 20 December, 2018.

OPEN GOVERNMENT WORKING GROUP. *Eight principles of open government data*. [S.l.]: Open Government Working Group, 2007. <https://opengovdata.org/>. Accessed on 20 December, 2018.

ORDANINI, A.; POL, A. Infomediation and competitive advantage in B2B digital marketplaces. *European Management Journal*, Elsevier, v. 19, n. 3, p. 276–285, 2001.

Oxford Dictionary. *Definition of Cognition*. [S.l.]: Oxford Dictionary. Available: www.oxforddictionaries.com, 2019.

Oxford Dictionary. *Definition of Framework*. [S.l.]: Oxford Dictionary. Available: www.oxforddictionaries.com, 2019.

PATEL, M. *I2S2 idealised scientific research activity lifecycle model*. University of Bath, 2011. Accessed on 20 December, 2018. Disponível em: <https://researchportal.bath.ac.uk/en/publications/i2s2-idealised-scientific-research-activity-lifecycle-model>.

PAULK, M. C.; CURTIS, B.; CHRISSIS, M. B.; WEBER, C. V. The capability maturity model for software. *Software engineering project management*, v. 10, p. 1–26, 1993.

PENNOCK, M. Digital curation: A life-cycle approach to managing and preserving usable digital information. *Library & Archives*, v. 1, p. 34–45, 2007.

PETERSEN, K.; FELDT, R.; MUJTABA, S.; MATTSSON, M. Systematic mapping studies in software engineering. In: *12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*. [S.l.: s.n.], 2008. v. 8, p. 68–77.

PETTERSSON, O.; SVENSSON, M.; GIL, D.; ANDERSSON, J.; MILRAD, M. On the role of software process modeling in software ecosystem design. In: ACM. *Proceedings of the Fourth European Conference on Software Architecture: Companion Volume*. [S.l.], 2010. p. 103–110.

PFEFFER, J.; SALANCIK, G. R. *The external control of organizations: A resource dependence perspective*. [S.l.]: Stanford University Press, 2003.

PICHLER, R. *10 Tips for Creating Agile Personas*. [S.l.]: Pichler, Roman, 2013. <https://www.romanpichler.com/blog/10-tips-agile-personas/>. Accessed on 20 December, 2018.

POIKOLA, A.; KOLA, P.; HINTIKKA, K. *Public Data: an introduction to opening information resources*. 2011. 24 p. <https://bit.ly/2DeOgVe>. Accessed on 20 December, 2018.

POLLOCK, R. *Building the (open) data ecosystem*. 2011. <https://bit.ly/2DddJ1k>. Accessed on 20 December, 2018.

PRIES-HEJE, J.; BASKERVILLE, R.; VENABLE, J. Strategies for design science research evaluation. *ECIS 2008 proceedings*, p. 1–12, 2008.

QIN, J.; BALL, A.; GREENBERG, J. Functional and architectural requirements for metadata: Supporting discovery and management of scientific data. In: *International Conference on Dublin Core and Metadata Applications*. Kuching, Sarawak, Malaysia: Dublin Core Metadata Initiative, 2012. p. 62–71.

ROSER, C. *All About Lean*. [S.l.]: Roser, Christoph, 2016. <https://commons.wikimedia.org/wiki/File:PDCA-Multi-Loop.png>. Accessed on 20 December, 2018.

RUNESON, P.; HÖST, M. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, Springer, v. 14, n. 2, p. 131–164, 2009.

RUSBRIDGE, C.; BURNHILL, P.; ROSS, S.; BUNEMAN, P.; GIARETTA, D.; LYON, L.; ATKINSON, M. The digital curation centre: a vision for digital curation. *IEEE International Symposium on Mass Storage Systems and Technology*, IEEE, 2005.

RUSSOM, P. Managing big data. *TDWI Best Practices Report, TDWI Research*, p. 1–40, 2013.

SANTANA, A. F. L. *BPMG–Um modelo conceitual para governança em BPM*. Tese (Doutorado) — PhD's thesis, Universidade Federal de Pernambuco, 2015.

SANTANA, A. F. L.; ALVES, C. F. Bpmg–um modelo conceitual para governança em bpm–aplicação numa organização pública. *iSys-Revista Brasileira de Sistemas de Informação*, v. 9, n. 1, p. 139–167, 2016.

SANTOS, H. D. A.; OLIVEIRA, M. I. S.; LÓSCIO, B. F. Datafeed: Uma ferramenta para coleta e visualização de feedbacks sobre dados publicados na web. In: *In Procedeeding of 5rd Simpósio Brasileiro de Tecnologia da Informação*. [S.l.: s.n.], 2017.

SANTOS, H. D. A. d.; OLIVEIRA, M. I. S.; LIMA, G. d. F. A. B.; SILVA, K. M. da; MUNIZ, R. I. V. C. S.; LÓSCIO, B. F. Investigations into data published and consumed on the web: a systematic mapping study. *Journal of the Brazilian Computer Society*, v. 24, n. 1, p. 14, Nov 2018. ISSN 1678-4804. Disponível em: <https://doi.org/10.1186/s13173-018-0077-z>.

SCHALKWYK, F. V.; WILLMERS, M.; MCNAUGHTON, M. Viscous open data: The roles of intermediaries in an open data ecosystem. *Information Technology for Development*, Taylor & Francis, v. 22, n. sup1, p. 68–83, 2016.

SCHWABER, K.; SUTHERLAND, J. *The scrum guide*. [S.l.]: Scrum Alliance, 2013. <http://www.scrumguides.org/docs/scrumguide/v1/scrum-guide-us.pdf>. Accessed on 20 December, 2018.

SCRUM ALLIANCE. *What is Scrum? An Agile Framework for Completing Complex Projects*. [S.l.]: SCRUM ALLIANCE, 2016. <https://www.scrumalliance.org>. Accessed on 20 December, 2018.

SEIDEWITZ, E. What models mean. *IEEE software*, IEEE, v. 20, n. 5, p. 26–32, 2003.

SEN, A. Metadata management: past, present and future. *Decision Support Systems*, Elsevier, v. 37, n. 1, p. 151–173, 2004.

SHANKARANARAYANAN, A. E. G. Managing metadata in data warehouses: Pitfalls and possibilities. *Communications of the Association for Information Systems*, v. 14, n. 1, p. 247–274, 2004.

SHANKARANARAYANAN, G.; EVEN, A. The metadata enigma. *Communications of the ACM*, ACM, v. 49, n. 2, p. 88–94, 2006.

SHIN, D.-H. Demystifying big data: Anatomy of big data developmental process. *Telecommunications Policy*, Elsevier, v. 40, n. 9, p. 837–854, 2016.

SHIN, D.-H.; CHOI, M. J. Ecological views of big data: Perspectives and issues. *Telematics and Informatics*, Elsevier, v. 32, n. 2, p. 311–320, 2015.

SILVA, A. R. da. Model-driven engineering: A survey supported by the unified conceptual model. *Computer Languages, Systems & Structures*, Elsevier, v. 43, p. 139–155, 2015.

SILVA, E. C. G.; OLIVEIRA, M. I. S.; OLIVEIRA, E.; GAMA, K. S. da; LÓSCIO, B. F. Um survey sobre plataformas de mediação de dados para internet das coisas. p. 1–12, 2015.

SINGER, J.; SIM, S. E.; LETHBRIDGE, T. C. Software engineering data collection for field studies. In: *Guide to Advanced Empirical Software Engineering*. [S.l.]: Springer, 2008. p. 9–34.

SMITH, G.; OFE, H. A.; SANDBERG, J. Digital service innovation from open data: exploring the value proposition of an open data marketplace. In: IEEE. *System Sciences (HICSS), 2016 49th Hawaii International Conference on*. [S.l.], 2016. p. 1277–1286.

SOLINGEN, D. R. van; BERGHOUT, E. W. *The Goal/Question/Metric Method: a practical guide for quality improvement of software development*. [S.l.]: McGraw-Hill, 1999.

SOOKLAL, R.; PAPADOPOULOS, T.; OJIAKO, U. Information systems development: a normalisation process theory perspective. *Industrial Management & Data Systems*, Emerald Group Publishing Limited, v. 111, n. 8, p. 1270–1286, 2011.

Stitch Data. *The State of Data Science*. 2015. <https://www.stitchdata.com/resources/the-state-of-data-science/>. Accessed on 20 December, 2018.

STOCK, D.; WINTER, R. The value of business metadata: Structuring the benefits in a business intelligence context. In: *Information Technology and Innovation Trends in Organizations*. [S.l.]: Springer, 2011. p. 133–141.

STRONG, D. M.; LEE, Y. W.; WANG, R. Y. Data quality in context. *Communications of the ACM*, ACM, v. 40, n. 5, p. 103–110, 1997.

TAVARES, A. T.; , M. I. S. ; LÓSCIO, B. F. Data producer catalogs for the web of things: a study on nosql solutions. In: ACM. *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. [S.l.], 2016. p. 980–985.

TEECE, D. J. Business models, business strategy and innovation. *Long Range Planning*, Elsevier, v. 43, n. 2-3, p. 172–194, 2010.

THWINK ORGANIZATION. *Sustainability*. 2014. <www.thwink.org/sustain/glossary/Sustainability.htm>. Accessed on 20 December, 2018.

TREMBLAY, M. C.; HEVNER, A. R.; BERNDT, D. J. Focus groups for artifact refinement and evaluation in design research. *Cais*, v. 26, p. 27, 2010.

UBALDI, B. *Open government data: Towards empirical analysis of open government data initiatives*. [S.l.]: Organisation for Economic Cooperation and Development (OECD), 2013.

VIEIRA, V.; TEDESCO, P.; SALGADO, A. C. Designing context-sensitive systems: An integrated approach. *Expert Systems with Applications*, Elsevier, v. 38, n. 2, p. 1119–1138, 2011.

VNUK, L.; KORONIOS, A.; GAO, J. *Managing Metadata Towards Enhanced Data Quality in Asset Management*. [S.l.]: Springer, 2012.

WELSER, H. T.; COSLEY, D.; KOSSINETS, G.; LIN, A.; DOKSHIN, F.; GAY, G.; SMITH, M. Finding social roles in wikipedia. In: ACM. *Proceedings of the 2011 iConference*. [S.l.], 2011. p. 122–129.

WIKIPEDIA. *Ecosystem*. 2001. <https://en.wikipedia.org/wiki/Ecosystem>. Accessed on 20 December, 2018.

WITT, M. Institutional repositories and research data curation in a distributed environment. *Library Trends*, The Johns Hopkins University Press, v. 57, n. 2, p. 191–201, 2008.

XIAO, B.; ZHANG, C.; MAO, Y.; QIAN, G. Review and exploration of metadata management in data warehouse. In: IEEE. *Industrial Electronics and Applications (ICIEA), 2015 IEEE 10th Conference on*. [S.l.], 2015. p. 928–933.

YIN, R. K. Case study research design and methods third edition. *Applied social research methods series*, Sage Publications Inc, v. 5, 2003.

YIN, R. K. *Case study research: Design and methods*. [S.l.]: Sage publications, 2013.

YOO, Y.; HENFRIDSSON, O.; LYYTINEN, K. Research commentary—the new organizing logic of digital innovation: an agenda for information systems research. *Information Systems Research*, INFORMS, v. 21, n. 4, p. 724–735, 2010.

ZELETI, F. A.; OJO, A. Capability matrix for open data. In: SPRINGER. *Working Conference on Virtual Enterprises*. [S.l.], 2014. p. 498–509.

ZELETI, F. A.; OJO, A. Open data value capability architecture. *Information Systems Frontiers*, Springer, p. 1–24, 2016.

ZELETI, F. A.; OJO, A. K. Critical factors for dynamic capabilities in open government data enabled organizations. In: *DG. O*. [S.l.: s.n.], 2016. p. 86–96.

ZOTT, C.; AMIT, R. Business model design: an activity system perspective. *Long Range Planning*, Elsevier, v. 43, n. 2-3, p. 216–226, 2010.

ZUBCOFF, J. J.; VAQUER, L.; MAZÓN, J.-N.; MACIÁ, F.; GARRIGÓS, I.; FUSTER, A.; CARCEL, J. V. The university as an open data ecosystem. *International Journal of Design & Nature and Ecodynamics*, WIT Press, v. 11, n. 3, p. 250–257, 2016.

ZUIDERWIJK, A.; JANSSEN, M. Barriers and development directions for the publication and usage of open data: A socio-technical view. In: *Open Government*. [S.l.]: Springer, 2014. p. 115–135.

ZUIDERWIJK, A.; JANSSEN, M.; CHOENNI, S.; MEIJER, R.; ALIBAKS, R. S.; SHEIKH_ALIBAKS, R. Socio-technical impediments of open data. *Electronic Journal of e-Government*, v. 10, n. 2, p. 156–172, 2012.

ZUIDERWIJK, A.; JANSSEN, M.; DAVIS, C. Innovation with open data: Essential elements of open data ecosystems. *Information Polity*, IOS Press, v. 19, n. 1, 2, p. 17–33, 2014.

ZUIDERWIJK, A.; JANSSEN, M.; KAA, G. van de; POULIS, K. The wicked problem of commercial value creation in open data ecosystems: Policy guidelines for governments. *Information Polity*, IOS Press, 2016.

ZUIDERWIJK, A.; JANSSEN, M.; POULIS, K.; KAA, G. van de. Open data for competitive advantage: insights from open data use by companies. In: ACM. *Proceedings of the 16th Annual International Conference on Digital Government Research*. [S.l.], 2015. p. 79–88.

APPENDIX A – META-MODEL SPECIFICATION

The Data Ecosystem Meta-model source files are available at <https://svn.riouxsvn.com/dataecosystem/DataEcosystemModel>. In the following sections, all the classes are defined.

A.1 CLASS: NAMED MODEL ELEMENT

Class Name:	**NamedModelElement**
Definition:	A NamedModelElement is an element in the meta-model that may have a name and a description. NamedModelElement supports using a string expression to specify its name as well as to describe it.
Label:	Named Model Element
subclassOf:	EClass

A.1.1 Properties

The following properties are recommended for use on this class: name and description and domain.

Property Name:	**name**
Definition:	A name given to the Data Ecosystem
Label:	name
Range:	EString

Property Name:	**description**
Definition:	Free-text describing the Data Ecosystem.
Label:	description
Range:	EString

A.2 CLASS: DATA ECOSSYSTEM

Class Name:	**DataEcosystem**
Definition:	Data Ecosystem is a socio-technical complex network that enables collaboration between autonomous actors such as enterprises, institutions and individuals
Label:	Data Ecosystem
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, domain, subDataEcosystem, subjectedBy and composedOf.

Relationship Name:	**composedOf**
Definition:	Specifies DataEcosystemElement instances that compose a Data Ecosystem
Label:	Composed Of
Range:	DataEcosystemElement
Relationship Type:	Aggregation
Cardinality:	1..*

A.2.1 Properties

Property Name:	**domain**
Definition:	The main category of the Data Ecosystem. A Data Ecosystem can have multiple domains
Label:	Domain
Range:	EString

A.2.2 Relationships

Relationship Name:	**subDataEcosystem**
Definition:	Refers to a grouping of sub DataEcosystem instances.
Label:	Sub Data Ecosystem
Range:	DataEcosystem
Relationship Type:	Aggregation
Cardinality:	0..*

Relationship Name:	**subjectedBy**
Definition:	ContextSet related to Data Ecossytem
Label:	Composed Of
Range:	ContextSet
Relationship Type:	Reference
Cardinality:	0..1

A.3 CLASS: DATA ECOSSYSTEM ELEMENT

Class Name:	**DataEcosystem**
Definition:	Data Ecosystem Element is an abstract class that represents a basic unit/construct of Data Ecosystem
Label:	Data Ecosystem Element
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, hasFeature and subjectedBy.

A.3.1 Relationships

Relationship Name:	**hasFeature**
Definition:	Refers to a grouping of feature that may characterize a DataEcosystemElement
Label:	Has Feature
Range:	Feature
Relationship Type:	Aggregation
Cardinality:	0..*

Relationship Name:	**subjectedBy**
Definition:	ContextSet related to Data Ecosystem Element
Label:	Subjected By
Range:	ContextSet
Relationship Type:	Reference
Cardinality:	0..1

A.4 CLASS: FEATURE

Class Name:	**Feature**
Definition:	Feature represents a physical or non-physical property or characteristic of a DataEcosystemElement that can be observed and measured. A feature may or may not directly affect the DataEcosystemElement behavior. In adition, The feature properties may or may not be considered as information suitable for the discovery process.
Label:	Feature
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, definition, dataType, measureUnit, derivedFeature, subFeature and hasValue.

A.4.1 Properties

Property Name:	**definition**
Definition:	Reference to a vocabulary or ontology that explains the meaning of the feature
Label:	Definition
Range:	EString

Relationship Name:	**derivedFeature**
Definition:	A derivation is a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.
Label:	Derived Feature
Range:	Feature
Relationship Type:	Reference
Cardinality:	0..*

Property Name:	**dataType**
Definition:	A classification that specifies which type of value a feature has and what type of mathematical, relational or logical operations can be applied to it without causing an error
Label:	Data Type
Range:	DataTypeEnum

Property Name:	**measureUnit**
Definition:	A quantity used as a standard of measurement
Label:	Measure Unit
Range:	EString

A.4.2 Relationships

Relationship Name:	**hasValue**
Definition:	Refers to a grouping of Feature Value instances. A Feature instance might be associated to zero or more instances of FeatureValue.
Label:	Has Value
Range:	FeatureValue
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**subFeature**
Definition:	There can be Feature instances composed by other sub-Feature instances.
Label:	Sub Feature
Range:	Feature
Relationship Type:	Aggregation
Cardinality:	0..*

A.5 CLASS: FEATURE VALUE

Class Name:	**FeatureValue**
Definition:	The value of a feature
Label:	Feature Value
subclassOf:	EClass

The following properties are recommended for use on this class: value and timestamp.

A.5.1 Properties

Property Name:	**value**
Definition:	The value of the feature value
Label:	Value
Range:	EJavaObject

Property Name:	**timestamp**
Definition:	Sequence of characters or encoded information identifying when a certain the feature value was recorded, usually giving date and time of day, sometimes accurate to a small fraction of a second
Label:	Timestamp
Range:	ELong

A.6 CLASS: CONTEXT SET

Class Name:	**ContextSet**
Definition:	Identifies the relevant context elements related to a DataEcosystem or DataEcosystemElement.
Label:	Context Set
subclassOf:	EClass

The following properties and relationships are recommended for use on this class: name and composedOf.

A.6.1 Properties

Property Name:	**name**
Definition:	A name given to the ContextSet
Label:	Name
Range:	EString

A.6.2 Relationships

Relationship Name:	**composedOf**
Definition:	Specifies Context Elements that compose a ContextSet
Label:	Composed Of
Range:	ContextElement
Relationship Type:	Aggregation
Cardinality:	1..*

A.7 CLASS: CONTEXT ELEMENT

Class Name:	**ContextElement**
Definition:	Contains information that helps actors understand the background and purpose of a DataEcosystem or DataEcosystemElement.
Label:	Context Element
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, definition, derivedContextElement, subContextElement, hasValue and providedBy.

A.7.1 Properties

Property Name:	**definition**
Definition:	Reference to a vocabulary or ontology that explains the meaning of the context element
Label:	Definition
Range:	EString

A.7.2 Relationships

Relationship Name:	**derivedContextElement**
Definition:	A derivation is a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.
Label:	Derived Context Element
Range:	ContextElement
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**subContextElement**
Definition:	There can be ContextElement instances composed by other sub-ContextElement instances.
Label:	Sub Context Element
Range:	ContextElement
Relationship Type:	Aggregation
Cardinality:	0..*

Relationship Name:	**hasValue**
Definition:	Refers to a grouping of Context Element Value instances.
Label:	Has Value
Range:	ContextElementValue
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**providedBy**
Definition:	A Context Element can be obtained from different Context Providers.
Label:	Provided By
Range:	ContextElementValue
Relationship Type:	Reference
Cardinality:	0..*

A.8 CLASS: CONTEXT ELEMENT VALUE

Class Name:	**ContextElementValue**
Definition:	The value of a Context Element
Label:	Context Element Value
subclassOf:	EClass

The following properties are recommended for use on this class: value and timestamp.

A.8.1 Properties

Property Name:	**value**
Definition:	The value of the Context Element value
Label:	Value
Range:	EJavaObject

Property Name:	**timestamp**
Definition:	Sequence of characters or encoded information identifying when a certain the Context Element value was recorded, usually giving date and time of day, sometimes accurate to a small fraction of a second.
Label:	timestamp
Range:	ELong

A.9 CLASS: CONTEXT PROVIDER

Class Name:	**ContextProvider**
Definition:	Context Provider represents a provider of a Context Element
Label:	Context Provider
subclassOf:	NamedModelElement

This class has as recommended properties: name and description.

A.10 CLASS: RELATIONSHIP

Class Name:	**Relationship**
Definition:	Represents the way in which two actors are connected. The relationships are often based on a common interest or are also related to the role each actor serves in the ecosystem.
Label:	Relationship
subclassOf:	DataEcosystemElement

The following properties and relationships are recommended for use on this class: name, description, state, includes and produces.

A.10.1 Properties

Property Name:	**state**
Definition:	Condition or position that identifies whether a relationship is active.
Label:	State
Range:	StateEnum

A.10.2 Relationships

Relationship Name:	**includes**
Definition:	Links carried out transactions within a Relationship.
Label:	Includes
Range:	Transaction
Relationship Type:	Aggregation
Cardinality:	1..*

Relationship Name:	**produces**
Definition:	Refers to values produced by a relationship. Examples of values are transparence, innovation, financial resources.
Label:	Produces
Range:	Value
Relationship Type:	Reference
Cardinality:	0..*

A.11 CLASS: BUSINESS MODEL

Class Name:	**BusinessModel**
Definition:	Describes the rationale of how a relationship creates, delivers, and captures value. The process of business model construction is part of business strategy.
Label:	BusinessModel
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description and influences.

A.11.1 Relationships

Relationship Name:	**influences**
Definition:	Links the business models to Relationship
Label:	Influences
Range:	Relationship
Relationship Type:	Reference
Cardinality:	0..*

A.12 CLASS: TRANSACTION

Class Name:	**Transaction**
Definition:	Describes an exchange or transfer of Data Ecosystem resources
Label:	Transaction
subclassOf:	EClass
hasSubClass:	Usage, Collaboration

The following properties and relationships are recommended for use on this class: id, timestamp, transactionType, transactionState, hasCost, obeysTo, trades and includes.

A.12.1 Properties

Property Name:	**identifier**
Definition:	A name given to identify a specific transaction
Label:	Identifier
Range:	EString

Property Name:	**timestamp**
Definition:	Sequence of characters or encoded information identifying when a certain transaction happened, usually giving date and time of day, sometimes accurate to a small fraction of a second
Label:	Timestamp
Range:	ELong

Property Name:	**type**
Definition:	Represents a particular group of transaction that share similar characteristics and form a smaller division of a larger set.
Label:	Type
Range:	EString

Property Name:	**state**
Definition:	Condition or position that identifies whether a transaction is active.
Label:	State
Range:	StateEnum

A.12.2 Relationships

Relationship Name:	**hasCost**
Definition:	Group of Costs associated with a transaction
Label:	Has Cost
Range:	Cost
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**obeysTo**
Definition:	Defines associated type of flow that structures how a transaction is performed
Label:	Obeys To
Range:	Flow
Relationship Type:	Reference
Cardinality:	1..*

Relationship Name:	**trades**
Definition:	Refers to Resources exchanged or transferred between Actors
Label:	Trades
Range:	Resource
Relationship Type:	Reference
Cardinality:	1..*

Relationship Name:	**includes**
Definition:	Links the carried out activities within a transaction
Label:	Includes
Range:	Activity
Relationship Type:	Aggregation
Cardinality:	1..*

A.13 CLASS: COST CONFIGURATION

Class Name:	**CostConfiguration**
Definition:	Represents a how a particular cost is defined and how it can be measured
Label:	Cost Configuration
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, definition, dataType, measureUnit and structures.

A.13.1 Properties

Property Name:	**definition**
Definition:	Reference to a vocabulary or ontology that explains the meaning of the cost measure
Label:	Definition
Range:	EString

Property Name:	**dataType**
Definition:	A classification that specifies which type of value a cost has and what type of mathematical, relational or logical operations can be applied to it without causing an error
Label:	Data Type
Range:	DataTypeEnum

Property Name:	**measureUnit**
Definition:	A quantity used as a standard of measurement
Label:	Measure Unit
Range:	EString

A.13.2 Relationships

Relationship Name:	**structures**
Definition:	Refers to how a Cost instance is defined and evaluated.
Label:	Structures
Range:	Cost
Relationship Type:	Reference
Cardinality:	1..*

A.14 CLASS: COST

Class Name:	**Cost**
Definition:	The value of a cost measure
Label:	Cost
subclassOf:	EClass

The following properties are recommended for use on this class: value and timestamp.

A.14.1 Properties

Property Name:	**value**
Definition:	The value of the cost
Label:	Value
Range:	EJavaObject

Property Name:	**timestamp**
Definition:	Sequence of characters or encoded information identifying when a certain the cost value was recorded, usually giving date and time of day, sometimes accurate to a small fraction of a second
Label:	Timestamp
Range:	ELong

A.15 CLASS: VALUE CONFIGURATION

Class Name:	**ValueConfiguration**
Definition:	Represents a how a particular value can be measured or observed
Label:	Value Configuration
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, definition, dataType, measureUnit and structure.

A.15.1 Properties

Property Name:	**definition**
Definition:	Reference to a vocabulary or ontology that explains the meaning of the value measured or observed
Label:	Definition
Range:	EString

Property Name:	**dataType**
Definition:	A classification that specifies which type of value has and what type of mathematical, relational or logical operations can be applied to it without causing an error
Label:	Data Type
Range:	DataTypeEnum

Property Name:	**measureUnit**
Definition:	A quantity used as a standard of measurement
Label:	Measure Unit
Range:	EString

A.15.2 Relationships

Relationship Name:	**structures**
Definition:	Refers to how a Value instance is defined and evaluated.
Label:	Structures
Range:	Value
Relationship Type:	Reference
Cardinality:	1..*

A.16 CLASS: VALUE

Class Name:	**Value**
Definition:	The value of a measured or observed value
Label:	Value
subclassOf:	EClass

The following properties are recommended for use on this class: value and timestamp.

A.16.1 Properties

Property Name:	**value**
Definition:	The actual observed or measured value
Label:	Value
Range:	EJavaObject

Property Name:	**timestamp**
Definition:	Sequence of characters or encoded information identifying when a certain the value was recorded, usually giving date and time of day, sometimes accurate to a small fraction of a second
Label:	Timestamp
Range:	ELong

A.17 CLASS: FLOW

Class Name:	**Flow**
Definition:	Describes an abstract workflow description structure, on which a transaction occurs
Label:	Flow
subclassOf:	EClass

The following properties are recommended for use on this class: value and timestamp.

A.17.1 Properties

Property Name:	**input**
Definition:	Represents an input parameter to transaction flow
Label:	Input
Range:	EJavaObject

Property Name:	**output**
Definition:	Represents an output parameter to transaction flow
Label:	Output
Range:	EJavaObject

A.18 CLASS: ACTOR

Class Name:	**Actor**
Definition:	Represents an autonomous entity such as enterprise, institution or individual which plays one or more specific roles in Data Ecosystem. An actor is considered as a basic element of a Data Ecosystem with an identity and that has its own distinct existence.
Label:	Actor
subclassOf:	DataEcosystemElement
hasSubclass:	Individual, Organization

The following properties and relationships are recommended for use on this class: name, description, expects, capableOf, engages, responsibleFor, actsOn, produces and plays.

A.18.1 Relationships

Relationship Name:	**expects**
Definition:	Group of Expectations associated with an Actor
Label:	Expects
Range:	Expectation
Relationship Type:	Reference
Cardinality:	1..*

Relationship Name:	**capableOf**
Definition:	Group of Capabilities associated with an Actor
Label:	Capable Of
Range:	Capability
Relationship Type:	Reference
Cardinality:	1..*

Relationship Name:	**engages**
Definition:	Links an Actor to his relationships
Label:	Engages
Range:	Relationship
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**responsibleFor**
Definition:	Links an Actor to the group of Resources he is responsible for.
Label:	Responsible For
Range:	Resource
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**actsOn**
Definition:	Links an Actor to the group of Resources he consumes.
Label:	Acts On
Range:	Resource
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**produces**
Definition:	Links an Actor to the group of Resources he produces or provides.
Label:	Produces
Range:	Resource
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**plays**
Definition:	Links an Actor to the group of Roles he plays in the Data Ecosystem.
Label:	Plays
Range:	Role
Relationship Type:	Reference
Cardinality:	1..*

A.19 CLASS: INDIVIDUAL

Class Name:	**Individual**
Definition:	A specialization of Actor that represents a single person, especially when compared to the group or set to which they belong.
Label:	Individual
subclassOf:	DataEcosystemElement

The following properties and relationships are recommended for use on this class: name, description, contact, address, gender, nationality, jobTitle, workFor and memberOf.

A.19.1 Properties

Property Name:	**contact**
Definition:	Contact information for the actor.
Label:	Contact
Range:	EString

Property Name:	**address**
Definition:	Physical address of the individual
Label:	Address
Range:	EString

Property Name:	**gender**
Definition:	Gender of the Individual
Label:	Gender
Range:	EString

Property Name:	**nationality**
Definition:	Nationality of the individual
Label:	Nationality
Range:	EString

Property Name:	**jobTitle**
Definition:	The job title of the individual (for example, Financial Manager)
Label:	Job Title
Range:	EString

A.19.2 Relationships

Relationship Name:	**memberOf**
Definition:	Organizations to which the Individual belongs.
Label:	Member of
Range:	Organization
Relationship Type:	Bi-Directional Reference
Cardinality:	0..*

Relationship Name:	**workFor**
Definition:	Organizations to which the Individual works for.
Label:	Work For
Range:	Organization
Relationship Type:	Reference
Cardinality:	0..*

A.20 CLASS: ORGANIZATION

Class Name:	**Organization**
Definition:	Organized body of individuals with a particular purpose, especially a business, society, association, etc.
Label:	Organization
subclassOf:	DataEcosystemElement

The following properties and relationships are recommended for use on this class: name, description, contact, address, location, type, scope, function, composedOf and subOrganizations.

A.20.1 Properties

Property Name:	**contact**
Definition:	Contact information for the actor.
Label:	Contact
Range:	EString

Property Name:	**address**
Definition:	Physical address of the individual
Label:	Address
Range:	EString

Property Name:	**location**
Definition:	The location of for example where an organization is located.
Label:	Location
Range:	EString

Property Name:	**type**
Definition:	Defines the type of organization. For example, public, private or non-governmental organizations
Label:	Type
Range:	EString

Property Name:	**scope**
Definition:	Extent of activity of the organization
Label:	scope
Range:	EString

Property Name:	**function**
Definition:	The most popular function(s) associated with the organization,
Label:	Function
Range:	EString

A.20.2 Relationships

Relationship Name:	**composedOf**
Definition:	Specifies Individual instances that compose the Organization
Label:	Composed Of
Range:	Individual
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**subOrganization**
Definition:	There can be Organization instances composed by other sub-Organization instances.
Label:	Sub Organization
Range:	Organization
Relationship Type:	Aggregation
Cardinality:	0..*

A.21 CLASS: CAPABILITY

Class Name:	**Capability**
Definition:	Represents a named piece of functionality (or feature) that is declared as supported or requested by an Actor.
Label:	Capability
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, definition, type, derivedCapabality and subCapability.

A.21.1 Properties

Property Name:	**definition**
Definition:	Reference to a vocabulary or ontology that explains the meaning of the capability
Label:	Definition
Range:	EString

Property Name:	**type**
Definition:	Represents a particular group of capabilities that share similar characteristics and form a smaller division of a larger set.
Label:	Type
Range:	EJavaObject

A.21.2 Relationships

Relationship Name:	**subCapability**
Definition:	There can be Capability instances composed by other sub-Capability instances.
Label:	Sub Capability
Range:	Capability
Relationship Type:	Aggregation
Cardinality:	0..*

Relationship Name:	**derivedCapability**
Definition:	A derivation is a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.
Label:	Derived Capability
Range:	Capability
Relationship Type:	Reference
Cardinality:	0..*

A.22 CLASS: EXPECTATION

Class Name:	**Expectation**
Definition:	Represents a named feeling of expecting something to happen.
Label:	Expectation
subclassOf:	NamedModelElement

The following properties are recommended for use on this class: name, description, definition, type, priority

A.22.1 Properties

Property Name:	**definition**
Definition:	Reference to a vocabulary or ontology that explains the meaning of the expectation
Label:	Definition
Range:	EString

Property Name:	**type**
Definition:	Represents a particular group of expectations that share similar characteristics and form a smaller division of a larger set.
Label:	Type
Range:	EString

Property Name:	**priority**
Definition:	Represents precedence in date or position of a given expectation in relation to other expectations. It is a preferential rating; especially one that allocates rights to goods and services usually in limited supply.
Label:	Priority
Range:	PriorityEnum

A.23 CLASS: RESOURCE

Class Name:	**Resource**
Definition:	Represents useful or valuable product or possession produced, provided, curated or consumed by Actors. In Data Ecosystems, resources range from datasets and data-based software to infrastructure.
Label:	Resource
subclassOf:	NamedModelElement
hasSubclass:	Data, Software, Infrastructure

The following properties and relationships are recommended for use on this class: name, description, conformsTo, constrainedBy and hasQualityProperty

A.23.1 Relationships

Relationship Name:	**conformsTo**
Definition:	An established standard to which the described resource conforms.
Label:	Conforms To
Range:	Standard
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**constrainedBy**
Definition:	An established license to which the described resource is constrained.
Label:	Constrained By
Range:	License
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**hasQualityProperty**
Definition:	Refers to the performed quality property. Quality property can be performed to any kind of resource.
Label:	Has Quality Property
Range:	QualityProperty
Relationship Type:	Reference
Cardinality:	0..*

A.24 CLASS: DATA

Class Name:	**Data**
Definition:	Represents data resources produced, provided, curated or consumed by Actors.
Label:	Data
subclassOf:	DataEcosystemElement

The following properties and relationships are recommended for use on this class: name, description, domain, update/modification date, accessMode and providedBy.

A.24.1 Properties

Property Name:	**domain**
Definition:	The main knowledge domain of the data resource.
Label:	Domain
Range:	EString

Property Name:	**accessMode**
Definition:	Describes the data access mode available. Each data resource might be available in different forms, these forms might represent different formats of the dataset or different endpoints. Examples of distributions include a downloadable CSV file, an API or an RSS feed.
Label:	Access Mode
Range:	EString

A.24.2 Relationships

Relationship Name:	**providedBy**
Definition:	A Data can be obtained from Infrastructure resource.
Label:	Provided By
Range:	Infrastructure
Relationship Type:	Reference
Cardinality:	0..*

A.25 CLASS: SOFTWARE

Class Name:	**Software**
Definition:	Represents data-based software resources produced, provided, curated or consumed by Actors. In particular, data-based software includes reusable assets (components and services) or software assets (applications) that are used to consume, produce or providing data.
Label:	Software
subclassOf:	DataEcosystemElement

The following properties and relationships are recommended for use on this class: name, description, domain, components, type, function, providedBy and actsOn.

A.25.1 Properties

Property Name:	**domain**
Definition:	The main knowledge domain of the data resource.
Label:	Domain
Range:	EString

Property Name:	**type**
Definition:	Represents a particular group of software resources that share similar characteristics and form a smaller division of a larger set.
Label:	Type
Range:	EString

Property Name:	**function**
Definition:	The most popular function(s) associated with the software resource. For example, data publishing, data extraction, data query, data consumption.
Label:	Function
Range:	EString

Property Name:	**components**
Definition:	List of system components
Label:	Components
Range:	EJavaObject

A.25.2 Relationships

Relationship Name:	**providedBy**
Definition:	Software can be obtained from Infrastructure resource.
Label:	Provided By
Range:	Infrastructure
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**actsOn**
Definition:	Links Software to the group of Data Resources it consumes or produces.
Label:	Acts On
Range:	Resource
Relationship Type:	Reference
Cardinality:	1..*

A.26 CLASS: INFRASTRUCTURE

Class Name:	**Infrastructure**
Definition:	Represents infrastructure resources produced, provided, curated or consumed by Actors. Infrastructure is the underlying foundation or basic framework that supports actor activities
Label:	Infrastructure
subclassOf:	DataEcosystemElement

The following properties are recommended for use on this class: name, description, domain, type, components, provider and state.

A.26.1 Properties

Property Name:	**components**
Definition:	List of system components
Label:	Components
Range:	EJavaObject

Property Name:	**type**
Definition:	Represents a particular group of infrastructure resources that share similar characteristics and form a smaller division of a larger set.
Label:	Type
Range:	EString

Property Name:	**state**
Definition:	Condition or position that identifies whether a infrastructure resource is active.
Label:	State
Range:	StateEnum

Property Name:	**provider**
Definition:	Represents a provider of a infrastructure resource
Label:	Provider
Range:	EString

A.27 CLASS: STANDARD

Class Name:	**Standard**
Definition:	Represents documents setting out specifications, procedures and guidelines. They are designed to ensure products, services and systems are safe, reliable and consistent. They are based on industrial, scientific and consumer experience and are regularly reviewed to ensure they keep pace with new technologies.
Label:	Standard
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, publisher, type and scope.

A.27.1 Properties

Property Name:	**publisher**
Definition:	An entity primarily responsible for making the standard.
Label:	Publisher
Range:	EString

Property Name:	**type**
Definition:	Represents a particular group of standard that share similar characteristics and form a smaller division of a larger set.
Label:	Type
Range:	EString

Property Name:	**scope**
Definition:	Extent of activity of the standard
Label:	Scope
Range:	EString

A.28 CLASS: LICENSE

Class Name:	**License**
Definition:	A legal document giving official permission to do something with a Resource.
Label:	License
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, publisher, version, type, scope, state, url

A.28.1 Properties

Property Name:	**publisher**
Definition:	An entity primarily responsible for making the license.
Label:	Publisher
Range:	EString

Property Name:	**version**
Definition:	Version number or other version designation of the license.
Label:	Version
Range:	EString

Property Name:	**type**
Definition:	Represents a particular group of license that share similar characteristics and form a smaller division of a larger set.
Label:	Type
Range:	EString

Property Name:	**url**
Definition:	A file that contains the license in a given format
Label:	URL
Range:	Estring

Property Name:	**scope**
Definition:	Extent of activity of the license
Label:	Scope
Range:	EString

A.29 CLASS: QUALITY PROPERTY

Class Name:	**QualityProperty**
Definition:	Quality property is a characteristic relevant to Actors (e.g., the availability of a data).
Label:	Quality Property
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, definition, type, derivedQualityProperty and subQualityProperty.

A.29.1 Properties

Property Name:	**definition**
Definition:	Reference to a vocabulary or ontology that explains the meaning of the quality property
Label:	Definition
Range:	EString

Property Name:	**type**
Definition:	Represents a particular group of quality properties that share similar characteristics and form a smaller division of a larger set.
Label:	Type
Range:	EString

A.29.2 Relationships

Relationship Name:	**hasValue**
Definition:	Refers to a grouping of Quality Property Value instances.
Label:	Has Value
Range:	QualityPropertyValue
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**derivedQualityProperty**
Definition:	A derivation is a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.
Label:	Derived Quality Property
Range:	QualityProperty
Relationship Type:	Reference
Cardinality:	Reference

Relationship Name:	**subQualityProperty**
Definition:	There can be QualityProperty instances composed by other sub-QualityProperty instances.
Label:	Sub Quality Property
Range:	QualityProperty
Relationship Type:	Aggregation
Cardinality:	0..*

A.30 CLASS: QUALITY PROPERTY VALUE

Class Name:	**QualityPropertyValue**
Definition:	The value of a quality property
Label:	Quality Property Value
subclassOf:	EClass

The following properties are recommended for use on this class: value and timestamp.

A.30.1 Properties

Property Name:	**value**
Definition:	The value of the quality property value
Label:	Value
Range:	EJavaObject

Property Name:	**timestamp**
Definition:	Sequence of characters or encoded information identifying when a certain quality property value was recorded, usually giving date and time of day, sometimes accurate to a small fraction of a second
Label:	timestamp
Range:	ELong

A.31 CLASS: ROLE

Class Name:	**Role**
Definition:	Represents a function played by an actor in a Data Ecosystem. It is related to set of duties and activities. Several roles can be identified in Data Ecosystem. Typically, at least data consumers and data producers are identified in contemporary Data Ecosystems.
Label:	Role
subclassOf:	DataEcosystemElement

The following properties and relationships are recommended for use on this class: name, description, performs and responsibleFor.

A.31.1 Relationships

Relationship Name:	**performs**
Definition:	Refers to the group of activities assigned to a Role.
Label:	Performs
Range:	Activity
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**responsibleFor**
Definition:	Links a Role to a group of duties.
Label:	Responsible For
Range:	Duty
Relationship Type:	Reference
Cardinality:	0..*

A.32 CLASS: ACTIVITY

Class Name:	**Activity**
Definition:	Represents a task played by an actor in a Data Ecosystem.
Label:	Activity
subclassOf:	NamedModelElement

The following properties and relationships are recommended for use on this class: name, description, goal, type, priority, condition, derivedActivity, subActivity, uses and produces.

A.32.1 Properties

Property Name:	**goal**
Definition:	A Goal is a purpose or objective of some task or resource that is desirable from some system or user's point of view.
Label:	Goal
Range:	EString

Property Name:	**priority**
Definition:	Represents precedence in date or position of a given activity in relation to other activity. It is a preferential rating; especially one that allocates rights to goods and services usually in limited supply.
Label:	Priority
Range:	PriorityEnum

Property Name:	**condition**
Definition:	Condition determines whether or not the activity is performed.
Label:	Condition
Range:	EJavaObject

A.32.2 Relationships

Relationship Name:	**derivedActivity**
Definition:	A derivation is a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.
Label:	Derived Activity
Range:	Activity
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**subActivity**
Definition:	There can be Activity instances composed by other sub-Activity instances.
Label:	Sub Activity Property
Range:	Activity
Relationship Type:	Aggregation
Cardinality:	0..*

Relationship Name:	**uses**
Definition:	Links an Activity to the group of Resources it uses as input.
Label:	Uses
Range:	Resource
Relationship Type:	Reference
Cardinality:	0..*

Relationship Name:	**produces**
Definition:	Links an Activity to the group of Resources it produces as output.
Label:	Produces
Range:	Resource
Relationship Type:	Reference
Cardinality:	0..*

A.33 CLASS: DUTY

Class Name:	**Duty**
Definition:	Obligatory tasks, conduct, service, or functions that arise from a role
Label:	Duty
subclassOf:	NamedModelElement

The following properties are recommended for use on this class: name, description, goal, priority, condition, type

A.33.1 Properties

Property Name:	**type**
Definition:	Represents a particular group of duties that share similar characteristics and form a smaller division of a larger set.
Label:	Type
Range:	EString

Property Name:	**terms**
Definition:	Provisions that determine the nature and scope of a duty
Label:	Terms
Range:	EString

APPENDIX B – EXAMPLE OF MODEL DERIVED FROM THE META-MODEL FOR DATA ECOSYSTEM

Figure 38 presents the model created during the study case presented in Chapter 4. The presented model is instantiation of the meta-model considering a specific Data Ecosystem. In particular, an Open Data Ecosystem conducted by a Brazilian public university. The model creation involved the preliminary study of the meta-model and, therefore, a selection of classes that would represent important information to create a catalog for documenting and promoting the ecosystem under study.

This new concrete model included both the concepts previously selected and specializations to represent the particularities of the ecosystem. For example, it was created a hierarchy of roles and also a taxonomy of organizations that are part of the ecosystem.

In particular, some meta-model classes were not selected for the creation of concrete model. For example, Relationship, Transaction and Business Model. By the way, all classes associated with Relationship were not selected. According to case study participants, during the instantiation, they recognized the importance of tracking information related to Data Ecosystem relationships. However, due to lack of documentation and the incipient state of the Data Ecosystem under study, the participants recognized their lack of knowledge to generate concrete classes to represent the types of relationships in the university's Data Ecosystem.

Meanwhile, all classes related to Actor, Resource and Role constructs were selected to derive the concrete model. For each of these major constructs, the participants tried to verify specializations existing in the context of the university. For example, Data Provider, Data Consumer, Infrastructure Provider, Regulatory Authority and Ecosystem Coordinator roles have been identified. In addition, several classes derived from Actor were created. Each of them represents different types of actors either from the academic community or external data consumers.

Finally, some classes of the meta-model were selected, but did not require specialization. For instance, Expectation, Capability, Duty and Activity. According to the participants, the attributes defined for these classes were already sufficient to describe the information they needed.

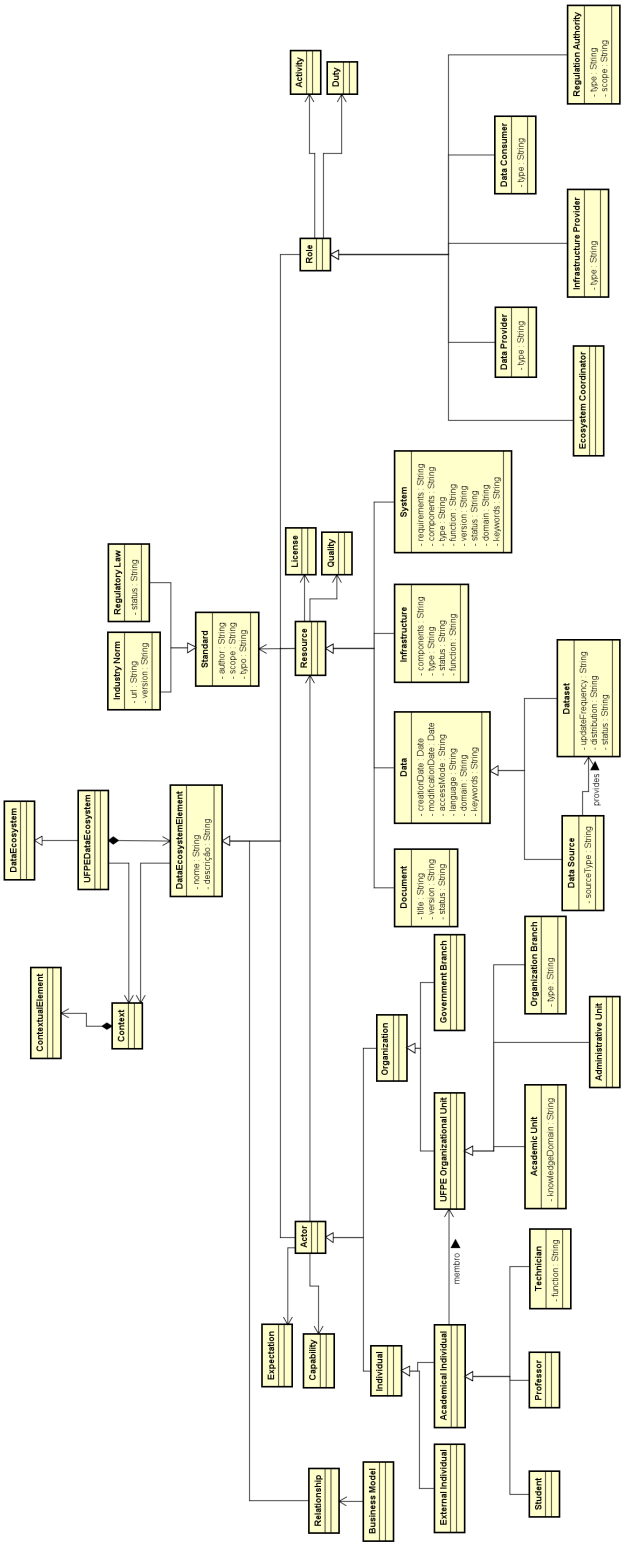

Figure 38 – Model created in the Case Study presented at Chapter 4. Source: Author

APPENDIX C – LOUVRE FRAMEWORK SPECIFICATION

C.1 ACTORS AND ROLES

Actors and Roles
Description: The people part of metadata curation involves actors who curate metadata, which they will make available to contributors and non-contributors alike (from here, all the actors who contribute to the curation of metadata will be called contributors). Louvre also supports team-building and teamwork. Contributors have varying areas of expertise, with their diverse set of skills, the contributors should be able to curate metadata as a group. Hence, teams and the characteristics of teams are central to metadata curation into Data Ecosystems. Contributors should work in collaboration with each other to curating metadata into Data Ecosystems. Several agile approaches emphasize that collaboration is the key to success for agile project delivery.
Roles clearly defined likely makes coordination easier (though non-trivial), while communication and coordination patterns often align with the structure of the software itself. Inspired by Scrum and Disciplined Agile Delivery frameworks, Louvre deemphasizes roles based strictly on skill sets in favor of primary roles that can include a variety of skills. Accordingly, the three primary roles are: • CurationMaster: Represents an authority to exist beyond the potential crowd of curators who organizes the whole metadata curation work providing a way for any actor to contribute to the common effort. The CurationMaster identifies and explains the metadata curation requirements, prioritizes requirements, provides feedback on the curatorial strategy. • Stakeholder: Represents someone who is materially impacted by the outcome of the metadata curation work. A stakeholder could be any Data Ecosystem actor who directly or indirectly consumes curated metadata. Stakeholder represents the needs and desires of the Data Ecosystem actors regarding metadata and curation needs. • MetadataCurator: Represents contributors who focus on performing the actual curation of metadata. MetadataCurators will carry out all tasks required to curate the metadata, including acquisition, appraisal and selection, quality assurance and preservation. Louvre also recognizes two secondary roles, which are typically introduced, often on a temporary basis, to address scaling issues. Accordingly, the three secondary roles are: • TechnicalExpert: Since MetadataCurators Team should perform tasks in one or more disciplines, this doesn't imply that every MetadataCurator needs to be an expert at everything. Sometimes the team needs the help of a TechnicalExpert to overcome a difficult problem and to transfer their skills to one or more MetadataCurator. • PlatformManager: Represents the authority that provides the services and functions for the overall creation and operation of the metadata curation platform. The platform used to support metadata curation tasks is a key source for successful curation and someone needs to be responsible for mitigates this risk. PlatformManager is responsible for planning, designing and maintaining the metadata curation platform and its related infrastructure. It is also responsible for establishing and maintaining preservation standards and policies, and providing curators support. • TeamLeader: Often, a team needs an informal leader, called a TeamLeader, who emerges from the team. The TeamLeader is a kind of metadata curation coach, helping to keep the team focused on curation activities work items and fulfilling their iteration goals and commitments that they have made to the Stakeholders. A TeamLeader has some additional responsibilities such as being the teacher of how metadata curation should be performed, (2) monitoring the long-term health of curated metadata, (3) organizing a team, (4) organizing design workshops, and (5) reviewing curation work done. Service providers, application developers, regulatory authorities and other roles may be involved in metadata curation. However, it would beyond the scope of Louvre to identify and defines all possible type of contributors, and all the roles that would apply.

C.2 AGILE PRACTICES

Agile Practices

Purpose: The Agile Practices provide a set of best practices that enable, encourage and guide contributors in the iterative, incremental and self-organizing curation of metadata. Best practice is a concept which refers to a set of techniques, procedures, and activities that has been shown by research and experience to produce good results and that is established or proposed as a standard suitable for widespread adoption.

Practices:

- **User Story:** It is a technique widely used by several agile methods to capture system requirements [74]. In metadata curation context, the simplicity of user stories lends itself to the articulation of metadata curation requirements. The informal language may contribute to create a shared understanding with those contributors not familiar with metadata curation and metadata management concepts.
- **Persona:** It is a detailed, synthetic biography of fictitious user. In metadata curation context, personas are useful when both MetadataCurators and CurationMaster don't have easy access to Stakeholders, helping to guide your decisions about metadata needs and curation requirements.
- **Backlog:** Backlog is a list of all the work necessary for the product. It lists all requirements, features, improvements, and fixes that constitute the changes to be made to the product in future releases. In metadata curation context, the backlog can be used to coordinate the contributors to perform tasks. In this sense, a curation backlog lists technical work related to metadata curation. The backlog items reflect tasks related to Louvre's dimensions and processes.
- **Sprint:** In Scrum methodology, the development process is divided into regular cycles over time. Each one of these cycles are called Sprint. Sprints are used to achieve some defined goals. During each Sprint a set of requirements is implemented, resulting in an increment of the product being developed. Each Sprint has a goal of what is to be built, a design and flexible plan that will guide building it, the work, and the resultant product increment. Each Sprint have time defined, meaning that the schedule for an interaction must be considered fixed and the scope of the iteration content is actively controlled to respect the schedule. In the context of metadata curation, a sprint is created from the creation of the curation backlog. Thus, each sprint would cover all the curatorial tasks related to a set of metadata. Thus, the metadata curation will be developed and delivered in small interactions, with a constant evolution of the processes or services of metadata curation, so that the curation initiative can obtain a quick return of curation benefits and minimize risks.
- **Coordination Meeting:** Coordination meetings is one of the most commonly practiced technique in software developments methods. These meetings present opportunity for a development team to get together on a regular basis to coordinate their activities. In several agile methods, it is recommended to schedule meetings in a short time basis. In distributed environments like Data Ecosystems, it is a challenge to arrange daily meetings in which all curation team members are able to attend. Hence, the Louvre recommends scheduling coordination meetings at particular points in the sprint cycle. It's important that the meeting is more about status updates than problem solving. Any impediments that are raised in the meeting would become team's responsibility. Hence, impediments can be resolved as quickly as possible.
- **Continuous Integration:** In Software Engineering, continuous integration is the practice in which developers integrate their work frequently, and testing the modifications, as early and often as possible. In metadata curation context, it is recommended that all metadata acquired should be as soon as possible transferred/ingested to a metadata repository. Once ingested in the repository, the metadata must be evaluated against compliance with quality policies, standards and other regulations. The idea is to find issues quickly; giving each MetadataCurator feedback on their work and automated mechanisms evaluate that work quickly.

Continued on next page

Table 34 – continued from previous page

Agile Practices

Practices:

- **Continuous Refactoring:** Refactoring is a controlled technique for restructuring the internal structure of the existing program's source code without changing its external behavior [52]. In Louvre, refactoring practice is viewed as the refactoring metaphor proposed by [13]. In that sense, refactoring should be seen as a restructuring and possible optimization of metadata curation work. For example, planned metadata curation actions that are not being properly performed or resulting in a poor performance should be reviewed as soon as the problem is identified.
- **Automated Tests:** In software development, an automated test is a technique to automate some repetitive but necessary tasks in a formalized testing process, or perform additional testing that would be difficult to do manually. In metadata curation context, an automated test suite enables curation teams to verify the quality of metadata safely. Moreover, such a practice can continuous assure quality.
- **Collective Ownership:** In software development projects, collective ownership is the explicit convention that any team member is allowed to make changes to any code file as necessary. In metadata curation context, the metadata can be owned by all the contributors. Everyone can have access and authorization to access, edit, and enhance any metadata. Ownership is collective and everyone is equally responsible to all parties.
- **Burndown Chart:** Burndown chart is a visual measurement tool to measure the development team progress. It shows the completed work per day against the projected rate of completion for the current project release. Burndown charts can be applied to metadata curation to measure progress over time. Thus, the curation work can be represented regarding either time or curation backlog items.

C.3 METADATA CURATION PLANNING DIMENSION

The Metadata Curation Planning Dimension consists of those processes performed to establish the total scope of the effort, define and refine the goals and requirements, and develop the course of action for curating metadata into Data Ecosystems. The Planning processes develop the metadata curation plan and the curation artifacts (e.g., policies, standards, and procedures) that will be used to carry out the curation initiative. The distributed nature of Data Ecosystems may require the use of repeated feedback loops for additional analysis. As more the experience and information are gathered and understood from metadata curation initiative, additional planning will likely be required. Significant changes occurring throughout the metadata curation initiative may trigger a need to revisit the action plan.

Process: Metadata Curation Requirements Engineering
Purpose: Recommends activities for elicitation (collecting, creating), analysis (aligning, prioritizing), and validation (monitoring, enforcing) of requirements involving metadata (e.g., business rules, metadata ownership, metadata classification, metadata quality, metadata usage, metadata access, authentication, entitlements, etc.).
Input: • Metadata needs • Curation Expectations from stakeholders • Regulatory Requirements
Outcomes: • The needs and expectations are refined into requirements. • Requirements are analyzed to determine their feasibility and priority. • Requirements are organized and available to stakeholders and other contributors.
Work Products: • Requirements Documentation: Describes metadata curation requirements for the Data Ecosystem. Requirements may start out at a high level and become progressively more detailed as more about the requirements are known.
Activities: • Envisioning requirements: Envision high-level goals for coming to a common understanding about the scope of metadata curation work. • Elaborating requirements: The information obtained during requirements envisioning is expanded and modified during elaboration. • Validating requirements: Certify that the requirements are an acceptable description of the stakeholders' expectations and needs related to metadata curation. • Classifying and prioritizing requirements: This activity grouping requirements into different priority groups with each group representing something stakeholders can relate to a concept or category. • Managing requirements: This activity organizes and manages requirements in order to communicate them between stakeholders and other contributors.

Process: Metadata Curation Planning

Purpose: Recommends activities for establishing the basis for creating and maintaining a metadata curation action plan that aligns with the requirements with a strategy and a set of policies, standards and procedures.

Input:

- Metadata Requirements
- Metadata Curation Requirements
- Regulatory Requirements

Outcomes:

- Scope and target are defined for metadata curation in accordance with metadata curation requirements.
- A metadata curation strategy is created, describing the vision, long-term goals, and an implementation road-map.
- Policies, procedures, standards and licenses are defined according to metadata curation strategy.
- Metadata curation platform architecture is established to support the activities towards the metadata curation activities.

Work Products:

- Metadata Curation Plan: Document describing the action plan for curating metadata into the Data Ecosystem. The scope of the metadata curation planning should include planning curation goals, strategy, policies, procedures, standards and licenses for curating metadata.

Activities:

- Defining and prioritizing metadata needs: Each Data Ecosystem must identify what metadata would benefit most from curation, and determine if potential returns would support the required efforts.
- Defining and reviewing metadata curation strategy: A strategy, on the other hand, is a blueprint, layout, design, or idea used to accomplish a specific goal. A strategy is very flexible and open for adaptation and change when needed. Thus, a strategy is an outline of the steps to curate metadata. A strategy is a solution that helps curators plan the metadata curation.
- Defining and reviewing metadata policies and procedures: Policies are formal, brief, and high-level statements on how the contributors must run the acquisition, quality assessment, preservation, security, and use of metadata [67]. Procedures define how to achieve requirements and are the mechanisms to enforce policies. The whole set of procedures also represent the implementation of metadata curation strategy.
- Defining and reviewing metadata standards: Metadata standards are documented agreed on properties and rules on representation, format, definition, structuring, dissemination, manipulation, use, and management of metadata.
- Approving and reviewing metadata curation platform architecture: Metadata curation platform architecture is an integrated set of high-level structures that govern and define how metadata is used, stored, managed and integrated within a Data Ecosystem.
- Defining and reviewing and licenses for use and reuse of metadata: Licenses are legal documents giving official permission to use and reuse of metadata. That is, licenses to clearly explain the conditions under which their metadata may be used.

C.4 METADATA ACQUISITION DIMENSION

The Metadata Acquisition Dimension consists of those processes required for creating, harvesting and selecting metadata. In particular, 'create' refers to original metadata generated and recorded by participants, and 'harvest' refers to pre-existing metadata collected from other sources. This dimension also includes activities to allow the selection and rejection of metadata that does not meet specified requirements and policies.

Process: Metadata Creation Management

Purpose: Recommends activities for creating appropriate metadata. The metadata produced by human beings is probably what most people assume when they think of a metadata source. In this process, the metadata will be created manually by the MetadataCurators.

Input:
- Metadata Requirements
- Metadata Strategy
- Metadata Procedures
- Metadata Policies
- Metadata Architecture
- Metadata Models
- Curation Backlog

Outcomes:
- A list of standard terminologies to apply to metadata is defined.
- A list of guidelines and documented workflows are created for guiding and support the creation of metadata.
- A set of metadata is created.
- Metadata has its basic quality policies ensured.

Work Products:
- Metadata Artifacts: The set of metadata created.
- Standardized Terminologies: All standard terminologies selected to be used in the metadata creation.
- Metadata Creation Workflow: Contains the guidelines, features and relevant information about how to create specific metadata.

Activities:
- Identifying and selecting a metadata model: Analyze and identify which metadata model should be applied to structure the metadata to be produced.
- Identifying and selecting standard terminology: The terms and phrases that are used for creating metadata content should reflect appropriate and accepted vocabularies in Data Ecosystem. Whenever possible use a controlled vocabulary, which provides a consistent way to describe metadata contents.
- Developing and documenting workflows for metadata creation or capture: Develop and document metadata creation workflows, i.e., the processes used to prepare and create metadata. These workflows act as metadata templates and provide some guidelines to ensure consistency.
- Ensuring basic quality control: Detect and correct general potential problems with metadata that could affect its use. The rejected metadata is ideally reported back to the metadata creator for further analysis to identify and to rectify the incorrect records.

Process: Metadata Harvesting Management

Purpose: Recommends activities for harvesting appropriate metadata. Not all metadata should be created from scratch. Indeed, such a notion will be unworkable. Harvesting is the process of collecting metadata from a remote or external source.

Input:

- Metadata Requirements
- Metadata Strategy
- Metadata Procedures
- Metadata Policies
- Metadata Architecture
- Metadata Models
- Curation Backlog

Outcomes:

- A list of metadata to harvest is defined.
- A list of metadata sources are identified
- Mechanisms are defined and implemented for harvesting metadata from metadata sources.
- A set of metadata is periodically harvested.

Work Products:

- Metadata Artifacts: The set of metadata harvested.
- Metadata Harvesting Mechanisms: A set of mechanisms that automate the means for harvesting metadata from metadata sources.
- Metadata Sources Inventory: A document that contains the identified sources for relevant metadata.

Activities:

- Selecting and prioritizing metadata to harvest: This activity involves the prioritization of metadata to be harvested. Hence, it is important to analyze the feasibility of metadata identified in terms of importance and other aspects, such as cost and timeliness, providing the basis on which determine harvest priority.
- Identifying, analyzing, and selecting metadata sources: Identifying, analyzing and selecting potential sources of metadata that best fit the metadata needs of the Data Ecosystem.
- Designing and developing metadata harvesting mechanisms for selected sources: Design and develop mechanisms to harvest metadata from selected metadata sources. A properly designed harvesting mechanism extracts metadata from the sources, enforces quality and consistency standards, normalizes and integrates metadata so that separate sources can be used together, and finally delivers metadata in the standardized format so that it can be loaded into the final target repository.
- Defining a harvesting schedule plan: Usually, metadata harvesting is done on a scheduled basis to reflect changes made to the source. Normally, such a schedule plan relies on static intervals, with harvesting processes occurring at daily, weekly, monthly, or other periodic intervals.

Process: Metadata Model Management

Purpose: Recommends activities for designing, developing and refining of a metadata model, which is the overall structure for the metadata. Such a model can be viewed as an integrated, subject-oriented set of specifications defining the essential metadata produced and consumed across the Data Ecosystem. While integrated means that the concepts in the model fit together, subject-oriented means the model is divided into commonly recognized subject areas that span across different aspects related to Data Ecosystem. In particular, subject area models are used to address specific elements needed by a business or management activity.

Input:

- Metadata Requirements
- Metadata Standards
- Metadata Standardized Terminologies
- Metadata Vocabularies and Ontologies

Outcomes:

- A set of subject areas are identified to categorize metadata.
- Subject area metadata models are defined to structure and specify metadata items.

Work Products:

- Metadata Models: The set of metadata models, which structure and standardize metadata.

Activities:

- Defining modeling language: Several different data modeling languages are available, each using different diagramming conventions or styles. Hence, it is essential to evaluate and select the most suitable language to design a metadata model that support curation needs.
- Determining metadata to be modeled: Determine metadata to be modeled. It is essential to collect and restructure base information that will establish metadata model design.
- Organizing needed metadata into subject areas: Organize metadata needs into high-level subject areas. All metadata produced and consumed across the Data Ecosystem will be represented within a subject area. Each subject area is a high-level classification of metadata, which represents a group of concepts about a major topic of interest.
- Defining granularity level of each subject area: Decide which aspects of metadata are crucial for stakeholders, and how granular each type of metadata need to be.
- Identifying and selecting relevant metadata standards: It is necessary to identify those standards that include the metadata items needed to describe a subject area. These standards can be used to design subject area models.
- Designing subject area models: Create a subject metadata models, using a high-level conceptual model. Designing subject area model includes the definitions of all the concepts (entities, attributes) of the subject area.
- Evaluating and refining metadata model: Metadata model refinement is an iterative process. First evaluate the resulting model using as a reference the requirement documentation, competency questions or usage scenarios. The evaluation results will guide the refinement of metadata model on how effectively and efficiently formalize and structure metadata required.

Process: Metadata Appraisal and Selection Management

Purpose: Recommends activities for evaluating metadata and selecting for long-term curation and preservation. An appraisal is "the process of evaluating records to determine which are to be retained as archives, which are to be kept for specified periods and which are to be destroyed" (HIGGINS, 2008). Selection is a more general term, usually applied when deciding what will be added to a repository. Appraisal and selection of metadata are critical because of the limited resources that most Data Ecosystems dedicate for preservation. Given the large (and growing) volume of metadata, it is a practical necessity to choose only the most important for long-term management.

Input:

- Metadata Requirements
- Metadata Strategy
- Metadata Procedures
- Metadata Policies
- Metadata Quality Reports

Outcomes:

- Policies are identified to guide the appraisal and selection of metadata
- Procedures are defined for the appraisal and selection of metadata according to the related policies.
- Procedures are automated for the appraisal and selection of metadata according to the related policies.
- Metadata are appraised and selected according to the defined policies and procedures.
- Metadata are disposed according to the defined policies and procedures.

Work Products:

- Appraisal and Selection Plan: Documents and describes the policies and procedures used for appraising and selecting metadata.
- Appraisal and Selection Automated Procedures: A set of mechanisms, tools or systems that automate the appraisal and selection procedures of metadata.

Activities:

- Defining appraisal and selection policies: Appraisal should be done according to well-defined selection policy. A policy allows informed, consistent and accountable decisions about appraisal and selection of metadata to be made in situations where judgments are subjective and speculative.
- Defining appraisal and selection procedures: Appraisal and selection procedures indicate the means and considerations that are taken into account for selecting metadata.
- Appraisal and re-appraisal of metadata: Evaluate metadata regarding appraisal and selection criteria and select for long-term curation and preservation. Moreover, re-assessment of the appraisal and selection decisions may be required in order to accommodate changing metadata and curation requirements.
- Disposing of metadata: Dispose of metadata, which has not been selected for long-term curation and preservation in accordance with appraisal and selection policies and criteria. Typically metadata is destroyed. In other cases, metadata may be transferred to another archive, repository, or other custodians.

C.5 METADATA QUALITY MANAGEMENT DIMENSION

The Metadata Quality Management Dimension consists of those processes required for detecting and preventing inconsistencies and defects in metadata as well as for improving the quality of metadata by cleaning such defects. The quality assurance aims to ensure that curated metadata remains authentic, reliable, usable, accessible to, and understandable over the long term.

Process: Metadata Quality Control

Purpose: Recommends activities for measuring, assessing and ensuring the quality of metadata. Metadata Quality control is a set of procedures intended to ensure that the metadata adheres to a defined set of quality criteria, thus meeting the curation requirements.

Input:
- Metadata Artifacts
- Metadata Requirements
- Metadata Strategy
- Metadata Procedures
- Metadata Policies

Outcomes:
- A scope is defined for metadata quality in accordance with the metadata curation requirements.
- Policies are defined for supporting metadata quality management.
- Metrics are defined for metadata quality evaluation according to the related policies.
- Measurement methods are defined by which to determine values for metadata quality metrics.
- Metadata quality measurements are collected and monitored.
- The CurationMaster, MetadataCurators and other contributors are notified about metadata quality issues.

Work Products:
- Metadata Quality Plan: Defines metadata quality policies as well as describes how the metadata quality policies will be implemented. It describes how contributors plan to meet the quality requirements set for curation of metadata.

Activities:
- Analyzing metadata quality requirements: Metadata quality requirements are also generally considered part of requirements analysis.
- Developing metadata quality policies: Develop metadata quality policies that are appropriate for metadata curation strategy, comply with metadata requirements and establish the foundation for the continual improvement of metadata quality.
- Developing metadata quality metrics and measurement methods: Develop or select the metrics and correspond measurement methods used to measure/profile the metadata quality policies.
- Measuring and monitoring metadata quality: Measure and monitor metadata quality to evaluate the compliance with defined metadata quality policies. It is also important to establish appropriate resources to measure metadata quality.
- Documenting and reporting metadata quality issues: To provide awareness of curation activities, there should be periodic reports about most frequent quality issues and common resolutions.

Process: Metadata Quality Improvement

Purpose: Recommends activities for planning, implementation and control activities that apply quality management methods to improve and refine the fitness for the use of metadata.

Input:
- Metadata Artifacts
- Metadata Quality Plan
- Metadata Requirements
- Metadata Curation Procedures
- Metadata Curation Policies
- Metadata Quality Issues Data
- Metadata Quality Report

Outcomes:
- Metadata quality issues are registered and manage.
- Metadata quality improvements are executed involving metadata cleansing and metadata enrichment.
- Metadata improvements are reviewed and homologated.
- Records are kept for all improvements made to the metadata.

Work Products:
- Improved Metadata Artifacts: The set of metadata improved.
- Metadata Improvement Log: Contains records for all improvements made to the metadata.
- Metadata Improvement Report: Details a set of steps for improving common metadata quality issues

Activities:
- Managing metadata quality issues: Register, track, evaluate and prioritize metadata quality issues and activities for resolving those incidents. A quality issue catalog may be used to log the quality evaluation, initial diagnosis, and subsequent actions associated with metadata quality issues.
- Cleaning and correct metadata: Detect and correct (or remove) errors and inconsistencies of metadata as well as identify incomplete, incorrect, inaccurate or irrelevant parts of the metadata and then replacing, modifying, or deleting the dirty metadata.
- Enriching metadata: Enhance or improve raw metadata. Although there are different ways to enrich metadata, a common metadata enrichment process could, for example, assign meaning to metadata by semantic enrichment. Another alternative is metadata linkage which is a technique for connecting pieces of metadata that are related.
- Reviewing and homologating metadata improvements: Inspect a set of modifications (cleansing and enrichment) over metadata to identify conflicts and the introduction of new defects.

C.6 METADATA PRESERVATION AND DISSEMINATION DIMENSION

The Metadata Preservation and Dissemination Dimension consists of those processes required for preserving metadata and ensuring that metadata is discoverable and accessible to both contributors and non-contributors. It includes processes to guide how to integrate metadata, transfer metadata to an appropriate repository and securely store them adhering to relevant standards. In addition, it also recommends processes to guide how to make metadata accessible by displaying publicly or by exposing them to other systems.

Process: Metadata Ingest Management

Purpose: Recommends activities for transferring metadata to an appropriate repository for permanent (long-term) storage while maintaining and verifying the integrity of metadata.

Input:
- Metadata Artifacts
- Metadata Curation Policies
- Metadata Curation Procedures
- Metadata Curation Platform

Outcomes:
- Contributors are identified and authenticated.
- Metadata deposit agreement is collected.
- Provenance information about metadata is collected and properly managed.
- Metadata are transferred to metadata repository and as a consequence are securely stored.

Work Products:
- Metadata Repository: A metadata archive that properly stores in a well-configured (regarding hardware and software) metadata.
- Metadata Deposit Agreement: Certification by the metadata creator that the metadata conforms to all policies and conditions and are fit for deposit into the repository.
- Metadata Provenance Artifacts: The set of provenance information, which tracks the steps by which the metadata was created and derived

Activities:
- Performing authentication: Confirm the identity of the contributor, who is contributing metadata to the environment. Common methods are password authentication or authorization via a digital signature.
- Collecting deposit agreement: Collect a certification by the metadata creator (i.e., the contributor responsible for creation or harvesting of metadata) that the metadata conforms to all policies and conditions (e.g., do not violate any legal restrictions placed on the metadata) and are fit for deposit into the repository.
- Tracking and maintaining metadata provenance: Track provenance is essential to the many domains where it can be used to evaluate the quality of metadata source, track the creation of intellectual property, and provide an audit trail for regulatory purposes. In particular, provenance is one kind of information which tracks the steps by which the metadata was created.
- Storing the metadata securely: Add the ingested metadata to a well-configured (in terms of hardware and software) metadata repository. Perform routine checks and provide disaster recovery capabilities as needed.

Process: Metadata Versioning Management

Purpose: Recommends activities for managing metadata version. Metadata are not immutable objects; rather they are subject to changes. Moreover, several contributors may be involved in the creation of the same metadata content. Hence, it is important to identify the different version of metadata. Version management is the ability to manage the change metadata. The versioning process should also be clear and transparent to end users, so they always know which version of the metadata and data they have acquired.

Input:
- Metadata Requirements
- Metadata Curation Procedures
- Metadata Curation Policies
- Metadata Artifacts

Outcomes:
- Metadata versioning strategy is defined for supporting metadata versioning management.
- Procedures are defined for metadata version identification according to the related policies.
- Revision control methods and tools are developed by which allow tracking and controlling over changes to metadata.

Work Products:
- Metadata Versioning Plan: Defines metadata versioning policies as well as describes how the metadata versioning policies will be implemented. It describes how contributors plan to versioning, tracking and controlling metadata change and modification.

Activities:
- Developing and maintaining the metadata versioning strategy: Develop and maintain a metadata versioning strategy to keep track of different versions of a metadata. A metadata versioning strategy should envision a logical way to organize and control versions of metadata. It should include how to organize multiple versions within a metadata repository, and how best to describe them so they can be properly discovered and accessed.
- Developing and maintaining metadata version identification procedure: Develop and maintain a metadata identification procedure to identify metadata versions in a consistent, clear and transparent manner, so users always know which version of the metadata they have acquired.
- Developing and maintaining metadata revision control: Develop and maintain a revision control, which is a kind of practice that tracks and provides control over changes to metadata. It allows reverting a metadata to a previous revision, which is critical for tracking each other's modifications and correct inconsistencies.

Process: Metadata Integration Management

Purpose: Recommends activities for combining metadata and presenting them in a unified way. Often metadata accumulates in metadata silos, which refers to metadata kept separated from other related metadata. This leads to many problems ranging from inability to access the right metadata to unnecessary replication of metadata. Metadata integration consists of a set of techniques used for connecting silos of metadata. Thus, the integration of metadata enables to gain a more comprehensive understanding from metadata.

Input:
- Metadata Artifacts
- Metadata Models
- Metadata Standardized Terminologies

Outcomes:
- Metadata is analyzed and classified.
- Metadata is linked to another metadata items.
- Metadata is deduplicated.
- Metadata conflicts are resolved.

Work Products:
- Metadata Taxonomy: Represents the structure of categories of metadata. It organizes metadata by using a hierarchy of categories to make it easier to find related metadata.
- Integrated Metadata Artifacts: The set of metadata integrated.

Activities:
- Classifying metadata: Metadata classification is broadly defined as the process of organizing metadata by relevant categories so that it may be used and discovered more efficiently.
- Linking and combining metadata: Link and combine the ingested metadata into a single population. In particular, such activity connects pieces of metadata that are thought to relate to the same resource, person, object or abstract entity/concept. Thus it brings separate metadata together to create a richer metadata repository.
- Deduplicating metadata: Eliminates redundant or repetitive information from metadata, which can reduce metadata storage costs, as well as speed up the access and discovery of metadata.
- Identifying and resolving conflicts: Since metadata may be acquired from different sources, it can provide conflicting information. Conflicts can arise because of incomplete, erroneous, or out-of-date metadata. It is thus critical for metadata curation to resolve conflicts from redundant metadata and identify true values from false ones.

Process: Metadata Access and Dissemination Management

Purpose: Recommends activities for providing access to metadata. In order to be used, metadata must be accessible for both humans and machines. This process includes activities for creating the means for enabling machine-based search and retrieval functionality that help a user identifies what metadata exist, where the metadata are located, and how can they be accessed (*e.g.*, download).

Input:

- Metadata Artifacts
- Metadata Curation Policies
- Metadata Requirements
- Metadata Curation Platform
- Metadata Curation Platform Architecture
- Metadata security mechanisms
- Risk Assessment Report

Outcomes:

- Metadata is indexed for supporting easy and fast access.
- Conditions and controls for controlling access to metadata are defined and established.
- Interfaces for enabling access to metadata are defined and established.
- Metadata is disseminated to external catalogs and platforms.

Work Products:

- Metadata Access Mechanisms: A set of mechanisms that enable access and retrieval of metadata.

Activities:

- Indexing metadata: Develop or create indexing structures in order to improve the speed of metadata retrieval operations. Indexes are used to quickly locate metadata without having to search every metadata item in a repository every time a metadata is requested.
- Determining appropriate access controls: Determine what level of metadata access is required and how stakeholders and metadata consumers may be affected by this condition. In addition, depending on the conditions for access and reuse, place access restrictions on some or all metadata.
- Designing and developing metadata access interface: Design and develop access interfaces that support the search-and-retrieval functionality required for metadata assets. The search and retrieval services should help participants to identify what metadata exist, where the metadata are located, and how can they be accessed.
- Disseminating metadata: Active dissemination of metadata to search and discovery services (e.g., external metadata catalogs, web-based indexes) for federated search and discovery. Often, metadata is exchanged with external organizations through structured files such as XML and JSON formats. Such dissemination would contribute to the preservation of metadata as well as to ease the discovery and access of curated metadata.

C.7 METADATA CURATION COORDINATION DIMENSION

The Metadata Curation Coordination Dimension consists of those processes required to promote, engage, monitor and control the contributors efforts toward achieving common and recognized metadata curation goals. According to (KRAUT; STREETER, 1995), coordination can be defined as "the integration or linking together of different parts of an organization to accomplish a collective set of tasks". In Louvre, This means that different contributors agree on a common definition of metadata curation, share information, and work together to achieve their goals. They must have a common view of what metadata should be curated, how metadata curation should be organized, and how it should fit with other activities already in place or undergoing parallel in Data Ecosystems. To curate metadata efficiently, they must not only monitoring and controlling the work being done but also share detailed information about the progress of curation activities. Furthermore, the contributors must coordinate their work so that it gets done and fits together.

Process: Metadata Curation Monitoring

Purpose: Recommends activities for monitoring metadata curation work. Monitoring allows results, processes, and experiences to be documented and used as a basis to steer decision-making and learning processes. Monitoring is checking progress against plans, as well as it reveals mistakes and offers paths for learning and improvement. Monitoring process needs to be planned and carried out on a regular basis throughout metadata curation initiative. Such regular monitoring keeps contributors up to date on the curation progress as well as verify the adherence to the curation plan.

Input:

- Metadata Curation Strategy
- Metadata Curation Policies, Procedures and Standards
- Metadata Requirements
- Contributors Register
- Metadata Curation Schedules
- Metadata Curation Performance Data

Outcomes:

- Monitoring indicators and their related procedures are defined for monitoring the ongoing curation activities against the curation plan.
- Indicators measurements are collected and documented.
- Nonconformities related to metadata curation strategy, policies, procedures, standards, and/or architecture are identified.
- The CurationMaster, key stakeholders and other contributors are notified about non-conformities and progress against curation plans.

Work Products:

- Curation Monitoring Plan: Defines monitoring indicators and the means used to measure them. It describes how contributors plan to meet to monitoring the curation activities.
- Monitoring Report: Documents monitoring indicators values (i.e., results) and identified non-conformities.
- Non-Conformities Notifications: All contributors involved in the curation of metadata are notified of problems detected.

Continued on next page

Table 47 – continued from previous page

Process: Metadata Curation Monitoring

Activities:

- Defining monitoring indicators: Indicators are metrics that are important for any project, particularly for monitoring and evaluation purposes. Through the indicators, contributors can pre-determine how effectiveness will be evaluated in a precise and clear manner.
- Defining monitoring indicators measurement methods: Once the indicators have been defined, a practical issue that needs to be addressed is how indicators will be measured and monitored. To evaluate the metadata curation progress and monitoring the plan, clearly defined procedures must be used consistently.
- Measuring and monitoring indicators: Measure and monitor indicators to evaluate the compliance with defined metadata curation plan. It is also important to establish appropriate resources to measure indicators.
- Monitoring conformance with metadata curation strategy, policies, procedures, standards, and architecture: Part of the curation work is to monitor and ensure conformance with available rules and regulations.
- Reporting status: Status reporting is one element of the curation controlling process. Its purpose is to keep the CurationMaster, key stakeholders and other contributors formally communicated and informed about curation work status.

Process: Recruiting and Engagement Management

Purpose: Recommends activities for recruiting contributors and promoting engagement. A collaborative metadata curation initiative relies on a constant stream of curation activity, with multiple contributors contributing in various ways, at various stages. A pool of volunteer contributors will perform metadata curation work. Without contributors to occupy necessary roles, the metadata curation initiative would cease to function. Therefore, it is necessary to motivate, engage and retain new contributors to promote a sustainable curation initiative. High engagement means that participants care about their work, feel like they're part of a community, are brought into the greater vision, and bring their unique strengths to their work (HASSELL, 2018). Recruiting of contributors should be seen as a continuous process that reaches the widest possible range of participants.

Input:

- Metadata Curation Goals
- Metadata Requirements
- Metadata Curation Strategy
- Data Ecosystem Environmental Factors
- Data Ecosystem Actors Registry

Outcomes:

- Actions to communicate, educate and promote the importance and value of metadata curation are planned and executed.
- A clear mission statement that spells out the metadata curation initiative purpose.
- A group of motivated contributors is recruited to collaborate with metadata curation initiative.
- Contributors are recognized as a valuable resource for the metadata curation initiative.
- Relationships and partnerships among contributors are reinforced and valued, following a more humane and collaborative approach.

Work Products:

- Recruiting and Engagement Plan: Provides guidance on how new contributor should be recruited, staffed and managed.

Activities:

- Developing and promoting metadata curation awareness: Promoting metadata curation awareness means more than ensuring that actors in the Data Ecosystem are aware of the existence of metadata curation issues.
- Creating a clear and compelling cause: Contributors must be provided with a convincing reason to be a part of the metadata curation initiative. The more compelling the mission, the easier it is to incentive actors to contribute to the metadata curation initiative.
- Assigning stewardship: Curation tasks need to be carried out by someone. Thus, it is crucial identifying who will curate the metadata and assigning to them their proper roles.
- Developing partnerships: It is important to know how to develop, maintain, sustain and manage partnerships, which are a key part of any strategic approach to engagement. Strengthening the partnerships can be one of the ways to motivate actors to contribute.
- Understanding and reducing barriers to contribution and engagement: It is important to ensure low practical barriers to participation in the work of metadata curation. The emphasis is on facilitating the engagement of new contributors.

Process: Communication and Feedback Management

Purpose: Recommends Recommends activities for planning and maintain an effective communication flow between contributors as well as ensuring the ultimate disposition of metadata curation information. Contributors spend part of their time communicating with other contributors, whether they are or not working the same metadata curation tasks. Effective communication creates a bridge between diverse contributors who may have different cultural references, different skill, perspectives and expectations. All of them impact or have an influence upon the metadata curation execution or outcome. In particular, this process should promote feedback that describes a circular process where the output of an activity, a process or a system is returned as input in order to regulate or influence a further output. It helps companies and institutions to capture user/customers concerns and analyze the data for future developments. In a metadata curation environment, feedback allows contributors to become indirectly involved in the curation work. Hence, collecting and analyzing feedback is an essential step towards improving metadata and their curation tasks.

Input:

- Metadata Curation Strategy
- Metadata Curation Procedures
- Metadata Curation Platform Architecture
- Metadata Curation Tools and Services

Outcomes:

- Communication strategy and channels are established.
- Communication to be managed is identified.
- Feedback and other metadata curation information are collected and documented.

Work Products:

- Communications Management Plan: Describes how project communications will be planned, structured, monitored, and controlled.
- Feedback and Communication Data: Feedback and other metadata curation information are documented and published in a shared area/repository to be reused by other coordination and management activities of metadata curation initiative.

Activities:

- Promoting open communication: Open communication is the ability of anyone, on equal conditions are encouraged to share their thoughts and concerns, both positive and negative, without the worry of retaliation when the feedback is bad. In Data Ecosystems, open communication may contribute to promote trust among contributors, and another natural result is the engagement of contributors.
- Defining and establishing communication strategy: A communication strategy defines how to organize and spread information, and ensure control of metadata curation work, while allowing contributors to communicate their progress when appropriate.
- Defining and establishing communication channels: Define communication channels to enable communication strategy. Examples of communications channels include websites, social media platforms, blogs, and newsletter.
- Developing and establishing feedback gathering mechanisms: Develop a range of methods to gather feedback. The feedback gathering mechanisms including, but not limited to, a contact form, forums, ratings and reviews surveys, or a comment box .
- Monitoring communication activities: Monitor communication activities in order to make sure that communications strategy is yielding the expected results, and to determine to what extent communication efforts are effective.
- Seeking contributors' discussions: Moreover, one-on-one discussions and focus groups are important to provide the opportunity to dive deeper into understanding the needs of contributors and can help them improve metadata curation initiative and its assets.

Process: Metadata Curation Controlling

Purpose: Recommends activities for organizing, orchestrating, and leading the metadata curation work. As mentioned before, the metadata curation work will be performed by a set of contributors with assigned roles and responsibilities. Contributors may have varied skill sets, may be assigned full or part-time, and may enter or leave the initiative at any moment. This process also focuses on continuous communication with contributors to understand their needs and expectations, addressing issues as they occur, managing conflicting interests and fostering appropriate engagement in curation activities.

Input:

- Metadata Requirements
- Metadata Curation Strategy
- Metadata Curation Procedures
- Contributors Registry
- Communications Management Plan
- Recruiting and Engagement Plan

Outcomes:

- Metadata needs are established and documented.
- Metadata needs are prioritized and organized into Metadata Backlog.
- Curation teams are formed.
- Curation Backlog, which contains and organizes a set of metadata curation work, is created.
- Metadata curation work progress data is gathered.
- Metadata curation work progress is tracked and monitored.

Work Products:

- Metadata Backlog: Organizes the list of all metadata that should be curated which, at a given moment, are known to be necessary and important to the Data Ecosystem.
- Curation Backlog: Organizes the list of technical metadata curation tasks which the team maintains and which, at a given moment, are known to be necessary to curate a set of metadata.
- Metadata Curation Work Progress Data: Information formally or informally provided by the project team related to the progress of metadata curation work.

Activities:

- Creating the Metadata Backlog: The Metadata Backlog serves to connect the contributors and orchestrate curation work. All metadata that should be curated will represent a work item into a Metadata Backlog.
- Forming curation teams: Contributors should work in collaboration with each other to curating metadata into Data Ecosystems. Contributors have varying areas of expertise, with their diverse set of skills, the contributors should be able to curate metadata as a group.
- Grooming of Curation Backlog: The grooming of Curation Backlog includes the definition of all the tasks related to the curation of selected metadata. These tasks must be aligned with the curation planning.
- Coordinating metadata curation tasks: Coordination of activities is one of the ongoing goals throughout the curation effort.
- Track curation progress: Monitoring the status of metadata curation work to keep track curation progress and provides the means to recognize deviation from the plan and take corrective and preventive actions and thus minimize risk.

C.8 METADATA CURATION PLATFORM ADMINISTRATION DIMENSION

The Metadata Curation Platform Administration Dimension consists of those processes for the design, implementation/deployment, and maintenance of the platform responsible for support the metadata curation work. Such a platform should enable standardization and integration of metadata curation tasks. It should support the metadata needs of the Data Ecosystem.

Process: Metadata Curation Platform Design

Purpose: Recommends activities for the strategic and technical design and planning of metadata curation platform architecture. Metadata curation platform architecture is an integrated set of high-level structures that govern and define how metadata is used, stored, managed and integrated within a Data Ecosystem. Each structure comprises software elements (*i.e.*, tools, systems or applications used to curate metadata), relations among them, and properties of both elements and relations. It functions as a blueprint for the metadata curation platform, laying out the structures to be developed and maintained by PlatformManager. Moreover, such architecture provides an understanding of creating and managing the flow of metadata and how it is processed across curation systems and applications.
Proper implementation of a platform for supporting metadata curation environment is not just about the tools to archive metadata. Rather, it is about creating a strategy to plan, design, and construct a platform capable of addressing curation requirements and allowing supporting the metadata curation strategy. Thus, this process also recommends activities to planning and designing an integrated and holistic platform to support the metadata curation work.

Input:
- Metadata Requirements
- Metadata Curation Strategy
- Metadata Curation Procedures
- Metadata Curation Policies
- Metadata Curation Standards

Outcomes:
- Identified metadata curation concerns are addressed by the Metadata Curation Platform architecture.
- Metadata Curation Platform architecture candidate models are developed.
- A set of technologies alternatives for Metadata Curation Platform are identified and related to architecture elements.

Work Products:
- Metadata Curation Platform Architecture: Defines the structure and behavior of Metadata Curation Platform. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the envisioned platform.
- Suitable Technologies Report: Documents the set of technologies alternatives for Metadata Curation Platform are identified and related to architecture elements.

Continued on next page

Table 51 – continued from previous page

Process: Metadata Curation Platform Design

Activities:

- Initialize the definition of the platform architecture: Build an understanding of the environment/context of use for which the platform is needed in order to establish platform functions. Therefore, establish a platform architecture roadmap and strategy that should include methods, modeling techniques, tools, need for any enabling curation platform.
- Developing candidate architectures models: Using relevant modeling techniques, and in conjunction with the curation needs and Requirement Engineering process, determine architectural entities, which address the different types of curation requirements.
- Identifying and analyzing available technologies: Selecting appropriate metadata related technology is an important metadata curation responsibility. It is necessary to identify and analyze such technology to meet curation needs, including total cost, reliability, and other aspects.
- Relating platform architecture to available technologies: Define the technologies (e.g., systems, tools, repositories) that reflect the platform architecture.
- Managing the designed architecture: Establish and maintain the rationale for all selections among alternatives and decision for the architecture. This includes concordance, completeness, and changes due to the environment or context changes, technological, implementation, and operational experiences.

Process: Metadata Curation Platform Deployment

Purpose: Recommends activities for establishing the metadata curation platform (*i.e.*, infrastructure and services) to support metadata curation work. This process defines or develops, deploys and maintains the facilities, tools, and communications and information technology assets needed for metadata curation work with respect to the architecture designed in the Metadata Curation Platform Design process.

Input:

- Metadata Curation Platform Architecture
- Suitable Technologies Report

Outcomes:

- Metadata curation technologies are installed and supported.
- Metadata Curation Platform is implemented to support metadata curation work.

Work Products:

- Metadata Curation Platform: Consists of facilities, tools, and communications and information technology assets needed for metadata curation work.

Activities:

- Establish the Metadata Curation Platform: Identify, obtain and provide technologies (resources, systems, tools and services) that are needed to implement/deploy the Metadata Curation Platform.
- Maintain the Metadata Curation Platform. Evaluate the degree to which Metadata Curation Platform satisfies curation needs. Moreover, identify and provide improvements or changes to the platform as the curation requirements change.

Process: Metadata Curation Platform Maintenance

Purpose: Recommends activities for the day-to-day technical supervision of the metadata curation platform. The metadata curation platform must to meet necessary conditions throughout the entire metadata lifecycle. Such management does not have one objective. There are many, including performance, efficiency, security, and privacy. The management of metadata curation environment comprises a number of proactive techniques including performance monitoring and tuning, storage and capacity planning, backup and recovery.

Input:

- Metadata Curation Requirements
- Metadata Curation Strategy
- Metadata Curation Procedures
- Metadata Curation Policies
- Metadata Curation Platform Architecture
- Metadata Curation Platform

Outcomes:

- Risks are evaluated and analyzed.
- Risk treatment options are identified, prioritized, and selected.
- Mechanisms are defined and implemented for ensuring metadata security and platform reliability.
- Metadata curation platform is monitored and optimized.

Work Products:

- Metadata security mechanisms: Defines a set of policies, standards, controls, and procedures for ensuring metadata security and platform reliability.
- Risk Assessment Report: Describes the identified risks in as much detail as is reasonable.

Activities:

- Performing a risk assessment: Understand, manage, control and mitigate risk to the metadata curation assets. The risk assessment includes, but not limited to, identify and prioritize assets, identify threats and vulnerabilities, analyze and establish controls to minimize or eliminate the probability of a threat.
- Developing and establishing metadata security mechanisms: Establish and maintain policies, controls, and procedures for metadata security. It includes creating plans to ensure that metadata curation platform can recover and continue as failures, defects or serious incidents occur.
- Monitoring and optimizing platform: To ensure ongoing access to metadata, it is important to monitor the condition of the platform continually. In particular, it is recommended routinely perform test retrievals or restorations of stored metadata. This activity also aims to optimize metadata curation platform performance both proactively and reactively, by monitoring performance and by responding to problems quickly and competently.

APPENDIX D – SURVEY QUESTIONNAIRE

Louvre Framework Evaluation Survey

* Required

Research Information

My name is Marcelo Iury de Sousa Oliveira. I am a doctoral student in Computer Science at the Center of Informatics (CIn) of the Federal University of Pernambuco (UFPE). I am supervised by the professor Bernadette Farias Lóscio. I would like to thank you for collaborating to my work by answering this survey. Your feedback is extremely valuable to the completion of my doctoral research.

The following questionnaire evaluates a framework to guide the curation of metadata into Data Ecosystems. In particular, the proposed framework is composed of a set of dimensions, processes, activities, practices and outcomes that serve as a reference for the development, implementation and maintenance of metadata curation initiatives in Data Ecosystems.

The questionnaire takes, on average, 1 hour to be answered completely. Your answers are anonymous, i.e., you do not need to provide your personal information, other than your email, if you want to receive the results of this survey.

We kindly ask you to answer the questionnaire COMPLETELY, otherwise we will have to discard it, since incomplete questionnaires will not be considered valid for the survey.

If you have any questions or doubts, please contact me by email (miso@cin.ufpe.br or marcelo.iury@ufrpe.br)

Information about the Respondent

The following questions serve merely to characterize the profile of the respondents. The respondents will not be identified or individually related within the work.

1. **What is your current position/function?** *

2. **Do you develop your activities in which context?** *
   Mark only one oval.
   - ◯ Academy
   - ◯ Professional/Industry
   - ◯ Both

3. **What is your educational background?** *
   Mark only one oval.
   - ◯ Computer Science
   - ◯ Information Science
   - ◯ Administration
   - ◯ Statistics
   - ◯ Other:

4. **Indicate your educational degree (options from the Brazilian educational system)** *
   Mark only one oval.
   - ◯ Elementary school
   - ◯ High school
   - ◯ Technical/vocational training
   - ◯ Some college credit, no degree
   - ◯ Bachelor's degree
   - ◯ Master's degree
   - ◯ Doctorate degree
   - ◯ Post-doctorate degree

5. **How would you rate your knowledge on the Data Ecosystem field?** *
   Mark only one oval.

	1	2	3	4	5	
Poor	◯	◯	◯	◯	◯	Excelent

6. **Have you participated directly or indirectly in a Data Ecosystem?** *
Mark only one oval.

- ◯ Yes
- ◯ No

7. **If you selected yes, how many years have you been participating in Data Ecosystem initiatives?** *
Mark only one oval.

- ◯ Less than 1 year
- ◯ 1 - 5 years
- ◯ 6 - 10 years
- ◯ More than 10 years

8. **How would you rate your knowledge regarding metadata curation or metadata management?** *
Mark only one oval.

	1	2	3	4	5	
Poor	◯	◯	◯	◯	◯	Excelent

9. **Do you have direct or indirectly curated metadata related to Data Ecosystems?** *
Mark only one oval.

- ◯ Yes
- ◯ No

10. **If you selected yes, how would you rate what extent your metadata curation practice was structured and organized?** *
Mark only one oval.

	1	2	3	4	5	
Ad-hoc	◯	◯	◯	◯	◯	Well-Organized

11. **How would you rate the importance of metadata curation for Data Ecosystems?** *
Mark only one oval.

	1	2	3	4	5	
Not Important	◯	◯	◯	◯	◯	Very Important

Louvre Framework Evaluation

The main goal of this survey is to evaluate the Louvre framework feasibility, completeness and adequacy based on community opinion. This survey has specific questions to evaluate the major elements of Louvre. In particular, the agile practices, dimensions and processes.

To answer this questionnaire, you should only consider one of the following criteria:
- If you are currently directly related to some metadata curation initiative, consider this background for answering the evaluation questions;
- Or, if you are not directly related to any metadata curation initiative, answer the evaluation questions based on your Data Ecosystem experience.

Roles Evaluation

12. **Can you indicate to what extent the following statements hold for the roles?** *
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The set of roles is correctly described	◯	◯	◯	◯	◯
The set of roles is complete	◯	◯	◯	◯	◯
The set of roles is appropriate	◯	◯	◯	◯	◯

13. **Please describe here any other suggestions for the Actors and Roles:**

Agile Practices Evaluation

14. **Can you indicate to what extent the following statements hold for the agile practices?** *
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The set of agile practices is correctly described					
The set of agile practices is complete					
The set of agile practices is appropriate					
The set of agile practices is feasible to be implemented					

15. **Please describe here any other suggestions for the agile practices:**

Metadata Curation Planning Dimmension Evaluation

16. **Is the Metadata Curation Planning Dimmension described correctly?** *
Mark only one oval.

- No, some processes in the dimension needs to be moved
- No, some processes in the process the dimension needs to be excluded
- No, some processes in the process the dimension needs to be updated
- No, some processes in the process the dimension needs to be included
- Yes, the set of processes are enough for the dimension purpose.

17. **Please describe here any other suggestions for Metadata Curation Planning Dimmension:**

Please indicate to what extent the following statements hold for each process specification

18. **The process is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Requirements Engineering					
Metadata Curation Planning					

19. **The process is enough for its purpose**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Requirements Engineering					
Metadata Curation Planning					

20. **The set of activities is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Requirements Engineering					
Metadata Curation Planning					

21. **The set of activities is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Requirements Engineering					
Metadata Curation Planning					

22. **The implementation of the set of activities is feasible**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Requirements Engineering					
Metadata Curation Planning					

23. **The set of outcomes is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Requirements Engineering					
Metadata Curation Planning					

24. **The set of outcomes is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Requirements Engineering					
Metadata Curation Planning					

25. **The set of outcomes is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Requirements Engineering					
Metadata Curation Planning					

26. **The set of outcomes is capable of being achieve**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Requirements Engineering					
Metadata Curation Planning					

Metadata Curation Monitoring and Controlling Dimension Evaluation

27. **Is the Metadata Curation Monitoring and Controlling Dimmension described correctly?** *
Mark only one oval.

- No, some processes in the dimension needs to be moved
- No, some processes in the process the dimension needs to be excluded
- No, some processes in the process the dimension needs to be updated
- No, some processes in the process the dimension needs to be included
- Yes, the set of processes are enough for the dimension purpose.

28. **Please describe here any other suggestions for Metadata Curation Monitoring and Controlling Dimmension:**

Please indicate to what extent the following statements hold for each process specification

29. **The process is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

30. **The process is enough for its purpose**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

31. **The set of activities is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

32. **The set of activities is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

33. **The set of activities is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

34. **The implementation of the set of activities is feasible**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

35. **The set of outcomes is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

36. **The set of outcomes is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

37. **The set of outcomes is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

38. **The set of outcomes is capable of being achieved**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Monitoring					
Recruiting and Engagement Management					
Communication and Feedback Management					
Metadata Curation Coordination					

Metadata Curation Platform Administration Dimension Evaluation

39. **Is the Metadata Curation Platform Administration Dimmension described correctly?** *
Mark only one oval.

- No, some processes in the dimension needs to be moved
- No, some processes in the process the dimension needs to be excluded
- No, some processes in the process the dimension needs to be updated
- No, some processes in the process the dimension needs to be included
- Yes, the set of processes are enough for the dimension purpose.

40. **Please describe here any other suggestions for Metadata Curation Platform Administration Dimmension:**

Please indicate to what extent the following statements hold for each process specification

41. **The process is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

42. **The process is enough for its purpose**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

43. **The set of activities is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

44. **The set of activities is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

45. **The set of activities is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

46. **The implementation of the set of activities is feasible**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

47. **The set of outcomes is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

48. **The set of outcomes is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

49. **The set of outcomes is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

50. **The set of outcomes is capable of being achieved**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Curation Platform Design					
Metadata Curation Platform Implementation					
Metadata Curation Platform Maintenance					

Metadata Acquisition Dimension Evaluation

51. **Is the Metadata Acquisition Dimension described correctly?** *
Mark only one oval.

- No, some processes in the dimension needs to be moved
- No, some processes in the process the dimension needs to be excluded
- No, some processes in the process the dimension needs to be updated
- No, some processes in the process the dimension needs to be included
- Yes, the set of processes are enough for the dimension purpose.

52. **Please describe here any other suggestions for Metadata Acquisition Dimmension:**

Please indicate to what extent the following statements hold for each process specification

53. **The process is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

54. **The process is enough for its purpose**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

55. **The set of activities is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

56. **The set of activities is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

57. **The set of activities is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

58. **The implementation of the set of activities is feasible**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

59. **The set of outcomes is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

60. **The set of outcomes is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

61. **The set of outcomes is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

62. **The set of outcomes is capable of being achieved**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Creation Management					
Metadata Harvesting Management					
Metadata Model Management					
Metadata Appraisal and Selection Management					

Metadata Quality Management Dimension Evaluation

63. **Is the Metadata Quality Management Dimension described correctly?** *
Mark only one oval.

- No, some processes in the dimension needs to be moved
- No, some processes in the process the dimension needs to be excluded
- No, some processes in the process the dimension needs to be updated
- No, some processes in the process the dimension needs to be included
- Yes, the set of processes are enough for the dimension purpose.

64. **Please describe here any other suggestions for the Metadata Quality Management Dimmension:**

Please indicate to what extent the following statements hold for each process specification

65. **The process is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

66. **The process is enough for its purpose**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

67. **The set of activities is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

68. **The set of activities is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

69. **The set of activities is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

70. **The implementation of the set of activities is feasible**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

71. **The set of outcomes is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

72. **The set of outcomes is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

73. **The set of outcomes is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

74. **The set of outcomes is capable of being achieved**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Quality Control					
Metadata Quality Improvement					

Metadata Preservation and Dissemination Dimension

75. **Is the Metadata Preservation and Dissemination Dimension described correctly?** *
Mark only one oval.

- No, some processes in the dimension needs to be moved
- No, some processes in the process the dimension needs to be excluded
- No, some processes in the process the dimension needs to be updated
- No, some processes in the process the dimension needs to be included
- Yes, the set of processes are enough for the dimension purpose.

76. **Please describe here any other suggestions for the Metadata Preservation and Dissemination Dimension:**

Please indicate to what extent the following statements hold for each process specification

77. **The process is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

78. **The process is enough for its purpose**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

79. **The set of activities is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

80. **The set of activities is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

81. **The set of activities is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

82. **The implementation of the set of activities is feasible**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

83. **The set of outcomes is correctly described**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

84. **The set of outcomes is complete**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

85. **The set of outcomes is appropriate**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

86. **The set of outcomes is capable of being achieved**
Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Metadata Ingest Management					
Metadata Versioning Management					
Metadata Integration Management					
Metadata Access Management					

Louvre Evaluation General Questions

87. **To what degree the Louvre is adaptable?**
Mark only one oval.

	1	2	3	4	5	
Completely not						Completely

88. **To what degree the Louvre is flexible?**
Mark only one oval.

	1	2	3	4	5	
Completely not						Completely

89. **To what degree a Data Ecosystem will obtain benefits of metadata curation?**
Mark only one oval.

	1	2	3	4	5	
Completely not	◯	◯	◯	◯	◯	Completely

90. **To what degree a Data Ecosystem will obtain benefits using Louvre Framework?**
Mark only one oval.

	1	2	3	4	5	
Completely not	◯	◯	◯	◯	◯	Completely

APPENDIX E – FOCUS GROUP QUESTIONNAIRE

Formulário de Avaliação do Louvre - Grupo Focal

Meu nome é Marcelo Iury de Sousa Oliveira. Sou doutorando em Ciência da Computação no Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE). Eu sou orientando da professora Bernadette Farias Lóscio. Gostaria de agradecer por colaborar no meu trabalho participando deste grupo focal. Seu feedback é extremamente valioso para a conclusão da minha pesquisa de doutorado.

O questionário a seguir avalia uma estrutura para orientar a curadoria de metadados em ecossistemas de dados. Em particular, a estrutura proposta é composta por um conjunto de dimensões, processos, atividades, práticas e resultados que servem como referência para o desenvolvimento, implementação e manutenção de iniciativas de curadoria de metadados em ecossistemas de dados.

Suas respostas são anônimas, ou seja, você não precisa fornecer suas informações pessoais, além do seu e-mail, se quiser receber os resultados dessa pesquisa.

Informações do Respondente

As informações abaixo servem meramente para caracterização do perfil dos respondentes. Em nenhum momento os respondentes serão identificados ou relacionados individualmente dentro do trabalho.

1. **Qual o seu cargo/função atualmente?**

2. **Você desenvolve suas atividades em qual meio?**
   Mark only one oval.
   - ◯ Acadêmico
   - ◯ Profissional
   - ◯ Ambos

3. **Qual o seu nível de formação (concluído)?**
   Mark only one oval.
   - ◯ Ensino Médio
   - ◯ Ensino Técnico
   - ◯ Graduação
   - ◯ Especialização
   - ◯ Mestrado
   - ◯ Doutorado
   - ◯ Pós-doutorado

4. **Você tem participado direta ou indiretamente de atividades relacionadas a Ecossistemas de Dados?**
   Mark only one oval.
   - ◯ Sim
   - ◯ Não

5. **Em caso de sim, há quantos anos participando de atividades relacionadas a Ecossistemas de Dados?**

6. **Você tem participado direta ou indiretamente de iniciativas/projetos para curadoria/gestão de dados e/ou metadados?**
   Mark only one oval.
   - ◯ Sim
   - ◯ Não

7. **Em caso de sim, há quantos anos participando de iniciativas/projetos para curadoria/gestão de dados e/ou metadados?**

Instruções para Preenchimento do Questionário

O framework de curadoria de metadados para Ecossistemas de Dados denominado Louvre contém 6 dimensões e 19 processos. Este questionário pretende apoiar a avaliação do modelo proposto através da realização de um grupo focal.

Para responder o presente questionário você deverá levar em consideração sua experiência na área de:
- Ecossistemas de Dados
- Curadoria e/ou gestão de dados e metadados

Componente: Agile Practices

8. **O conjunto de práticas ágeis está descrito corretamente? A ideia aqui era identificar se as práticas são viáveis, completas e adequadas para a curadoria de metadados em Ecossistemas de Dados. Também queríamos identificar lacunas, erros e / ou possibilidades de melhoria.**
   Mark only one oval.

   - ◯ Não, uma ou mais práticas ágeis precisam ser excluídas
   - ◯ Não, uma ou mais práticas ágeis precisamser atualizadas
   - ◯ Não, uma ou mais práticas ágeis precisam ser incluídas
   - ◯ Sim, o conjunto de práticas ágeis está descrito corretamente

9. **Quais são as suas sugestões em caso de necessidade de inclusão, exclusão ou atualização?**

Componente: Roles and Actors

10. **O conjunto de papeis está descrito corretamente? A ideia aqui era identificar se os papeis são viáveis, completos e adequados para a curadoria de metadados em Ecossistemas de Dados. Também queríamos identificar lacunas, erros e / ou possibilidades de melhoria.**
    Mark only one oval.

    - ◯ Não, um ou mais papeis precisam ser excluídos
    - ◯ Não, um ou mais papeis precisam ser atualizados
    - ◯ Não, um ou mais papeis precisam ser incluídos
    - ◯ Sim, o conjunto de papéis está descrito corretamente

11. **Quais são as suas sugestões em caso de necessidade de inclusão, exclusão ou atualização?**

Dimensão: Metadata Curation Analysis and Planning

12. **A dimensão está descrita corretamente? A ideia aqui era identificar se os seus processos são viáveis, completos e adequados para a respectiva dimensão. Também queríamos identificar lacunas, erros e / ou possibilidades de melhoria.**
    Mark only one oval.

    - ◯ Não, um ou mais processos precisam ser movidos para outra dimensão
    - ◯ Não, um ou mais processos precisam ser excluídos
    - ◯ Não, um ou mais processos precisam ser atualizados
    - ◯ Não, um ou mais processos precisam ser integrados a outro processo
    - ◯ Sim, a dimensão está descrita corretamente

13. **Quais são as suas sugestões em caso de necessidade de inclusão, exclusão, agrupamento ou atualização?**

Dimensão: Metadata Curation Monitoring and Controlling

14. **A dimensão está descrita corretamente? A ideia aqui era identificar se os seus processos são viáveis, completos e adequados para a respectiva dimensão. Também queríamos identificar lacunas, erros e / ou possibilidades de melhoria.**
Mark only one oval.

- ◯ Não, um ou mais processos precisam ser movidos para outra dimensão
- ◯ Não, um ou mais processos precisam ser excluídos
- ◯ Não, um ou mais processos precisam ser atualizados
- ◯ Não, um ou mais processos precisam ser integrados a outro processo
- ◯ Sim, a dimensão está descrita corretamente

15. **Quais são as suas sugestões em caso de necessidade de inclusão, exclusão, agrupamento ou atualização?**

Metadata Curation Platform Administration

16. **A dimensão está descrita corretamente? A ideia aqui era identificar se os seus processos são viáveis, completos e adequados para a respectiva dimensão. Também queríamos identificar lacunas, erros e / ou possibilidades de melhoria.**
Mark only one oval.

- ◯ Não, um ou mais processos precisam ser movidos para outra dimensão
- ◯ Não, um ou mais processos precisam ser excluídos
- ◯ Não, um ou mais processos precisam ser atualizados
- ◯ Não, um ou mais processos precisam ser integrados a outro processo
- ◯ Sim, a dimensão está descrita corretamente

17. **Quais são as suas sugestões em caso de necessidade de inclusão, exclusão, agrupamento ou atualização?**

Dimensão: Metadata Acquisition

18. **A dimensão está descrita corretamente? A ideia aqui era identificar se os seus processos são viáveis, completos e adequados para a respectiva dimensão. Também queríamos identificar lacunas, erros e / ou possibilidades de melhoria.**
Mark only one oval.

- ◯ Não, um ou mais processos precisam ser movidos para outra dimensão
- ◯ Não, um ou mais processos precisam ser excluídos
- ◯ Não, um ou mais processos precisam ser atualizados
- ◯ Não, um ou mais processos precisam ser integrados a outro processo
- ◯ Sim, a dimensão está descrita corretamente

19. **Quais são as suas sugestões em caso de necessidade de inclusão, exclusão, agrupamento ou atualização?**

Dimensão: Metadata Quality Management

20. **A dimensão está descrita corretamente? A ideia aqui era identificar se os seus processos são viáveis, completos e adequados para a respectiva dimensão. Também queríamos identificar lacunas, erros e / ou possibilidades de melhoria.**
Mark only one oval.

- ◯ Não, um ou mais processos precisam ser movidos para outra dimensão
- ◯ Não, um ou mais processos precisam ser excluídos
- ◯ Não, um ou mais processos precisam ser atualizados
- ◯ Não, um ou mais processos precisam ser integrados a outro processo
- ◯ Sim, a dimensão está descrita corretamente

21. **Quais são as suas sugestões em caso de necessidade de inclusão, exclusão, agrupamento ou atualização?**

Dimensão: Metadata Preservation and Dissemination

22. **A dimensão está descrita corretamente? A ideia aqui era identificar se os seus processos são viáveis, completos e adequados para a respectiva dimensão. Também queríamos identificar lacunas, erros e / ou possibilidades de melhoria.**
Mark only one oval.

- ◯ Não, um ou mais processos precisam ser movidos para outra dimensão
- ◯ Não, um ou mais processos precisam ser excluídos
- ◯ Não, um ou mais processos precisam ser atualizados
- ◯ Não, um ou mais processos precisam ser integrados a outro processo
- ◯ Sim, a dimensão está descrita corretamente

23. **Quais são as suas sugestões em caso de necessidade de inclusão, exclusão, agrupamento ou atualização?**

Conjunto de Dimensões

24. **O conjunto de dimensões está descrito corretamente? A ideia aqui era identificar se as dimensões são viáveis, completas e adequadas para a curadoria de metadados em Ecossistemas de Dados. Também queríamos identificar lacunas, erros e / ou possibilidades de melhoria.**
Mark only one oval.

- ◯ Não, uma ou mais dimensões precisam ser excluídas
- ◯ Não, uma ou mais dimensões precisam ser atualizadas
- ◯ Não, uma ou mais dimensões precisam ser incluídas
- ◯ Não, uma ou mais dimensões precisam ser agrupadas
- ◯ Sim, o conjunto de dimensões está descrito corretamente

25. **Quais são as suas sugestões em caso de necessidade de inclusão, exclusão, agrupamento ou atualização?**

Avaliação Geral

O Louvre é apropriado para orientar e apoiar a curadoria dos metadados?

Numa escala de 1 a 5, na qual 1 significa “Discordo Fortemente” e 5 significa “Concordo Fortemente”, como você avalia as afirmações abaixo:

- O Louvre é adaptável

- O Louvre é flexível

Louvre fornece subsídios para usuários saírem da forma ad-hoc e passar a fazer a curadoria sistemática de metadados?

APPENDIX F – PUBLICATIONS

- Journal Papers

  1. SANTOS, H. D. A.; OLIVEIRA, Marcelo Iury S.; LIMA, G. F. A. B.; SILVA, K. M.; CRUZ, R. V.; LÓSCIO, B. F.; Investigations into data published and consumed on the Web: a systematic mapping study. Journal of the Brazilian Computer Society, v. 24, n. 1, p. 14, 2018.
  2. OLIVEIRA, Marcelo Iury S.; LIMA, Glória de Fátima Barros; LÓSCIO, Bernadette Farias. Investigations into Data Ecosystems: a systematic mapping study. Knowledge and Information Systems, p. 1-42, 2019;

- Conference Papers:

  1. OLIVEIRA, Marcelo Iury S.; LOSCIO, B. F. Louvre: A Framework for Metadata Curation in Data Ecosystem. XV Simpósio Brasileiro de Sistemas de Informação, 2019, Aracajú-SE
  2. SANTOS, H. D. A.; OLIVEIRA, Marcelo Iury S.; LÓSCIO, B. F.; Uma estratégia para o refinamento colaborativo de dados na Web baseada em social coding. Submitted to: Brazilian Symposium on Multimedia and the Web, Salvador-BA, 2018
  3. LIMA, G. F. A. B.; OLIVEIRA, Marcelo Iury S.; LOSCIO, B. F. . Avaliação da Saúde de Ecossistemas de Dados. In Thesis and Dissertation Workshop, Simpósio Brasileiro de Banco de Dados, Rio de Janeiro-RJ, 2018
  4. OLIVEIRA, Marcelo Iury S.; OLIVEIRA, L. E. R. A. ; BATISTA, M. G. R.; LOSCIO, B. F. . Towards a Meta-model for Data Ecosystems. International Conference on Digital Government Research, 2018, Delft.
  5. OLIVEIRA, Marcelo Iury S.; LOSCIO, B. F. . What is a Data Ecosystem?. International Conference on Digital Government Research, 2018, Delft.
  6. OLIVEIRA, L. E. R. A. ; OLIVEIRA, Marcelo Iury S.; SANTOS, W.C.R.; LOSCIO, B. F. . Data on the Web Management System: A Reference Model. International Conference on Digital Government Research, 2018, Delft.
  7. OLIVEIRA, L. E. R. A.; OLIVEIRA, Marcelo Iury S.; LOSCIO, B. F. Um Survey sobre Soluções para Publicação de Dados na Web sob a Perspectiva das Boas Práticas do W3C. In 32º Simpósio Brasileiro de Banco de Dados, Umberlândia-MG, 2017

8. OLIVEIRA, Marcelo Iury S.; LOSCIO, B. F. Metadata Curation Framework for Supporting Data Ecosystems. In Thesis and Dissertation Workshop, Simpósio Brasileiro de Banco de Dados, Umberlândia-MG, 2017

9. SANTOS, H. D. A.; OLIVEIRA, Marcelo Iury S.; LOSCIO, B. F. Datafeed: Uma Ferramenta para coleta e visualização de feedbacks sobre dados publicados na Web. In Simpósio Brasileiro de Tecnologia da Informação, Jaboatão de Guararapes-PE, 2017

10. OLIVEIRA, Marcelo Iury S.; OLIVEIRA, L. E. R. A. ; LIMA, G. F. A. B. ; LOSCIO, B. F. . Enabling a Unified View of Open Data Catalogs. In: International Conference on Enter-prise Information Systems, 2016, Roma. Proceedings of the 18th International Conference on Enterprise Information Systems. Portugal: SCITEPRESS ? Science and Technology Pub-lications, Lda, 2016. v. 2. p. 230-239.

11. TRINDADE, A. T. ; Oliveira, Marcelo Iury S. ; LOSCIO, B. F. . Data Producer Catalogs for the Web of Things: A Study on NoSQL Solutions. In: ACM Symposium on Applied Com-puting, 2016, Pisa. Proceedings of the 31st ACM Symposium on Applied Computing, 2016. v. 1. p. 980-985.

12. OLIVEIRA, Marcelo Iury S.; OLIVEIRA, H. R. ; OLIVEIRA, L. E. R. A. ; LOSCIO, B. F. . Open Government Data Portals Analysis: The Brazilian Case. In: International Conference on Digital Government Research, 2016, Shanghai. Proceedings of the 17th Annual International Conference on Digital Government Research, 2016.

13. OLIVEIRA, Marcelo Iury S.; OLIVEIRA, L. E. R. A. ; LOSCIO, B. F. . Portal Dados Abertos Brasil: Uma solução para catalogação dos portais de dados abertos brasileiros. In: Conferência Web.br, 2015, São Paulo. Anais da conferência Web.br, 2015.

14. OLIVEIRA, Marcelo Iury S.; GAMA, K. S. ; LOSCIO, B. F. . Waldo: Serviço para Publicação e Descoberta de Produtores de Dados para Middleware de Cidades Inteligentes. In: XI Simpósio Brasileiro de Sistemas de Informação, 2015, Goiania. XI Simpósio Brasileiro de Sistemas de Informação, 2015.

15. OLIVEIRA, Marcelo Iury S.; LOSCIO, B. F. ; GAMA, K. S. . Análise de Desempenho de Catálogo de Produtores de Dados para Internet das Coisas baseado em SensorML e NoSQL. In: XIV Workshop em Desempenho de Sistemas Computacionais e de Comunicação, 2015, Recife. XIV Workshop em Desempenho de Sistemas Computacionais e de Comunicação, 2015.

16. SILVA, E. C. G. F. ; OLIVEIRA, Marcelo Iury S.; OLIVEIRA, E. ; GAMA, K. S. ; LOSCIO, B. F. . Um Survey sobre Plataformas de Mediação de Dados

para Internet das Coisas. In: 42º Seminário Integrado de Software e Hardware, 2015, Recife.

17. PESSOA, D. E. R. ; OLIVEIRA, Marcelo Iury S.; SALGADO, A. C. ; LOSCIO, B. F. . A proposal for RDF data integration benchmarking. In: 2nd International Workshop on Ben-chmarking RDF Systems, 2014, Hangzhou. 2nd International Workshop on Benchmarking RDF Systems, 2014.

- Chapters and Books

1. LOSCIO, B. F. ; BURLE, C. ; Oliveira, Marcelo Iury S. ; CALEGARI, N. . Fundamentos para publicação de dados na web. 1. ed. São Paulo: Comitê Gestor da Internet no Brasil, 2018. v. 1. 64p .
2. LOSCIO, B. F. ; OLIVEIRA, Marcelo Iury S. ; BITTENCOURT, I. I. . Publicação e Consumo de Dados na Web: Conceitos e Desafios. In: Carme S. Hara; Fábio Porto; Eduardo Ogasawar. (Org.). Tópicos em Gerenciamento de Dados e Informações. xvied.Petrópolis: Simpósio Brasileiro de Banco de Dados, 2015, v. , p. 39-
3. OLIVEIRA, Marcelo Iury S.; OLIVEIRA, L. E. R. A.; LIMA, G. F. A. B.; LÓSCIO, B. F. A platform for supporting open data ecosystems. In Enterprise Information Systems 291, pp. 313–337, 2017, DOI:10.1007/978-3-319-62386-3

- Awards

1. Best Student Paper Award on 18th International Conference on Enterprise Information Systems (ICEIS).