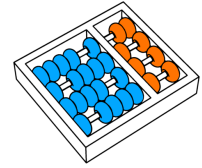


Joana Esther Gonzales Malaverri

“Supporting data quality assessment in eScience: a
provenance based approach”

*“Apoio à avaliação da qualidade de dados em
eScience: uma abordagem baseada em proveniência”*

CAMPINAS
2013



University of Campinas
Institute of Computing

*Universidade Estadual de Campinas
Instituto de Computação*

Joana Esther Gonzales Malaverri

**“Supporting data quality assessment in eScience: a
provenance based approach”**

Supervisor: Profa. Dra. Claudia Maria Bauzer Medeiros
Orientador(a):

***“Apoio à avaliação da qualidade de dados em
eScience: uma abordagem baseada em
proveniência”***

PhD Thesis presented to the Post Graduate Program of the Institute of Computing of the University of Campinas to obtain a PhD degree in Computer Science.

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Doutora em Ciência da Computação.

THIS VOLUME CORRESPONDS TO THE FINAL VERSION OF THE THESIS DEFENDED BY JOANA ESTHER GONZALES MALAVERRI, UNDER THE SUPERVISION OF PROFA. DRA. CLAUDIA MARIA BAUZER MEDEIROS.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA POR JOANA ESTHER GONZALES MALAVERRI, SOB ORIENTAÇÃO DE PROFA. DRA. CLAUDIA MARIA BAUZER MEDEIROS.

Supervisor's signature / *Assinatura do Orientador(a)*

CAMPINAS

2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

G589a Gonzales Malaverri, Joana Esther, 1981-
Apoio à avaliação da qualidade de dados em eScience uma abordagem baseada em proveniência / Joana Esther Gonzales Malaverri. – Campinas, SP : [s.n.], 2013.

Orientador: Claudia Maria Bauzer Medeiros.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Sistemas de informação gerencial - Controle de qualidade. 2. Banco de dados. 3. Metadados. 4. Framework (Programa de computador). 5. Recuperação da informação. I. Medeiros, Claudia Maria Bauzer, 1954-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em inglês: Supporting data quality assessment in eScience a provenance based approach

Palavras-chave em inglês:

Management information systems - Quality control

Databases

Metadata

Framework (Computer program)

Information retrieval

Área de concentração: Ciência da Computação

Titulação: Doutora em Ciência da Computação

Banca examinadora:

Claudia Maria Bauzer Medeiros [Orientador]

Juliano Lopes de Oliveira

José de Jesús Pérez-Alcazar

André Santanchè

Eliane Martins

Data de defesa: 06-05-2013

Programa de Pós-Graduação: Ciência da Computação

TERMO DE APROVAÇÃO

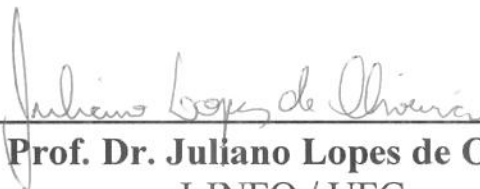
Tese Defendida e Aprovada em 06 de Maio de 2013, pela Banca
examinadora composta pelos Professores Doutores:



Prof. Dr. José de Jesús Pérez Alcázar
EACH / USP



Prof^a. Dr^a. Eliane Martins
IC / UNICAMP



Prof. Dr. Juliano Lopes de Oliveira
I-INFO / UFG



Prof. Dr. André Santanchè
IC / UNICAMP



Prof^a. Dr^a. Claudia Maria Bauzer Medeiros
IC / UNICAMP

Supporting data quality assessment in eScience: a provenance based approach

Joana Esther Gonzales Malaverri

May 06, 2013

Examiner Board/*Banca Examinadora*:

- Profa. Dra. Claudia Maria Bauzer Medeiros (Supervisor/*Orientadora*)
- Prof. Dr. André Santanchè
Institute of Computing - UNICAMP
- Profa. Dra. Eliane Martins
Institute of Computing - UNICAMP
- Prof. Dr. José de Jesús Pérez-Alcazar
EACH - USP
- Prof. Dr. Juliano Lopes de Oliveira
INF - UFG

Abstract

Data quality is a recurrent concern in all scientific domains. Experiments analyze and manipulate several kinds of datasets, and generate data to be (re)used by other experiments. The basis for obtaining good scientific results is highly associated with the degree of quality of such datasets. However, data involved with the experiments are manipulated by a wide range of users, with distinct research interests, using their own vocabularies, work methodologies, models, and sampling needs. Given this scenario, a challenge in computer science is to come up with solutions that help scientists to assess the quality of their data. Different efforts have been proposed addressing the estimation of quality. Some of these efforts outline that data provenance attributes should be used to evaluate quality. However, most of these initiatives address the evaluation of a specific quality attribute, frequently focusing on atomic data values, thereby reducing the applicability of these approaches. Taking this scenario into account, there is a need for new solutions that scientists can adopt to assess how good their data are. In this PhD research, we present an approach to attack this problem based on the notion of data provenance. Unlike other similar approaches, our proposal combines quality attributes specified within a context by specialists and metadata on the provenance of a data set. The main contributions of this work are: (i) the specification of a framework that takes advantage of data provenance to derive quality information; (ii) a methodology associated with this framework that outlines the procedures to support the assessment of quality; (iii) the proposal of two different provenance models to capture provenance information, for fixed and extensible scenarios; and (iv) validation of items (i) through (iii), with their discussion via case studies in agriculture and biodiversity.

Resumo

Qualidade dos dados é um problema recorrente em todos os domínios da ciência. Os experimentos analisam e manipulam uma grande quantidade de conjuntos de dados gerando novos dados para serem (re)utilizados por outros experimentos. A base para a obtenção de bons resultados científicos está fortemente associada ao grau de qualidade de tais dados. No entanto, os dados utilizados nos experimentos são manipulados por uma diversa variedade de usuários, os quais visam interesses diferentes de pesquisa, utilizando seus próprios vocabulários, metodologias de trabalho, modelos, e necessidades de amostragem. Considerando este cenário, um desafio em ciência da computação é oferecer soluções que auxiliem aos cientistas na avaliação da qualidade dos seus dados. Diferentes esforços têm sido propostos abordando a avaliação de qualidade. Alguns trabalhos salientam que os atributos de proveniência dos dados poderiam ser utilizados para avaliar qualidade. No entanto, a maioria destas iniciativas aborda a avaliação de um atributo de qualidade específico, frequentemente focando em valores atômicos de dados. Isto reduz a aplicabilidade destas abordagens. Apesar destes esforços, há uma necessidade de novas soluções que os cientistas possam adotar para avaliar o quão bons seus dados são. Nesta pesquisa de doutorado, apresentamos uma abordagem para lidar com este problema, a qual explora a noção de proveniência de dados. Ao contrário de outras abordagens, nossa proposta combina os atributos de qualidade especificados dentro de um contexto pelos especialistas e os metadados que descrevem a proveniência de um conjunto de dados. As principais contribuições deste trabalho são: (i) a especificação de um framework que aproveita a proveniência dos dados para obter informação de qualidade, (ii) uma metodologia associada a este framework que descreve os procedimentos para apoiar a avaliação da qualidade, (iii) a proposta de dois modelos diferentes de proveniência que possibilitem a captura das informações de proveniência, para cenários fixos e extensíveis, e (iv) a validação dos itens (i) a (iii), com suas discussões via estudos de caso em agricultura e biodiversidade.

Acknowledgements

- Thank you my loving Jesus for the gifts you gave me, to be my comfort in times of loneliness, but over all for your unconditional love and mercy for me.
- I would like to express my sincere thanks and gratitude to professor Claudia Bauzer Medeiros, my advisor along these years of research. Professor Claudia, thanks for your patient guidance and the opportunity that you gave me to become a member of LIS.
- I would like to thank to my beloved mother, granny, sisters and family. Thank you for sharing with your love, patience and prayers this journey with me. Thanks for the unconditional support that each one of you have provided me over the years.
- I would like to thank my friends and colleagues of LIS. Bruno, Ivelize, Ivo, Matheus, Daniel, Alessandra, João, Eduardo, Celso, Carla and Alan, all of you have allowed that these years of research become an unforgettable experience. I also thank professors André Santanchè, Rubens Lamparelli, Jansle Rocha, and also the Phd. student Rafael Moraes for all the suggestions, which were essential for the development of this research.
- I would like to thank to all my brazilian friends. My particular affection to Alana, Tathiana and Carlinha and all my IC's friends for your friendship and support over these years. I also thank my peruvian friends for your fellowship and support despite the distance.
- I would like to express my deepest thanks and appreciation to all my friends from the Pantokrator Catholic Community for your lovely and patient souls with someone infinitely impatient and full of defects.
- Finally, I would like to thank the financial support from Brazilian agencies: CNPq (grant 142337/2010-2), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project and PRONEX-FAPESP), INCT in Web Science(CNPq 557.128/2009-9) and CAPES, as well as individual grants from CNPq.

*“Without love, deeds, even the most
brilliant, count as nothing”*
St. Therese of Lisieux

Contents

Abstract	ix
Resumo	xi
Acknowledgements	xiii
Epigraph	xv
1 Introduction	1
2 A Provenance Approach to Assess the Quality of Geospatial Data	6
2.1 Introduction	6
2.2 Model overview	7
2.3 Quality elements	8
2.4 Conclusions	9
3 Data Quality in Agriculture Applications	10
3.1 Introduction	10
3.2 Data for agriculture applications	11
3.3 Dimensions of data quality	11
3.4 Data Quality Measurement	13
3.4.1 Manual approaches	13
3.4.2 Automatic approaches	15
3.5 Data Quality in Applications in Agriculture	16
3.6 Summing up	18
4 A Provenance-based Approach to Evaluate Data Quality in eScience	21
4.1 Introduction	21
4.2 Provenance and Metadata standards	22
4.3 Description of the framework	23

4.3.1	Architecture of the Framework	23
4.3.2	The Data Provenance Repository	25
4.3.3	Basic methodology	27
4.4	Case study	28
4.4.1	Problem Overview	29
4.4.2	Instantiating our provenance model	30
4.4.3	Applying the methodology	32
4.5	Conclusion	34
5	Estimating the quality of data using provenance: a case study in eScience	36
5.1	Introduction	36
5.2	Background of the solution	38
5.3	ProvenBiO	39
5.4	Case study: using ProvenBiO to derive data quality	40
5.4.1	Motivating Scenario	40
5.4.2	ProvenBiO ontology: a running example	42
5.4.3	Capturing Provenance Information	43
5.4.4	Querying Data Provenance to Derive Quality	44
5.5	Related Work	45
5.6	Conclusion and ongoing work	48
6	Conclusions	49
6.1	Contributions	49
6.2	Extensions	50
	Bibliography	52

List of Tables

3.1	The 15 dimensions framework [85]	12
3.2	The PSP/IQ model [43]	12
3.3	The classification of [61]	14
3.4	Summary of quality dimensions covered by automatic approaches	16
3.5	Classification of quality dimensions	19
3.6	Main data quality dimensions studied for the related work	20
3.7	Main data quality dimensions in agriculture applications	20
4.1	Examples of metrics for the completeness dimension	28
4.2	Artifact relation	30
4.3	Artifact Characteristic relation	31
4.4	Agent relation	31
4.5	Agent Performed Process relation	31
4.6	Generic Process relation	32
4.7	Process relation	32
4.8	Process Used Artifact relation	32
4.9	Domain Criteria relation	33
4.10	Example of result of the query	33

List of Figures

1.1	The Architecture of the Framework	4
1.2	Elements that compose the Provenance Manager	5
2.1	Our provenance model	8
4.1	The Architecture of the Framework	25
4.2	Our provenance model	26
4.3	Workflow of the activities to identify defective pixels for sugarcane areas, based on [55]	29
5.1	Example of a portion of our ProvenBiO ontology and the corresponding SPARQL query	40
5.2	Basic flow concerning processing animal sound recordings – FNJV, inspired on [19]	41
5.3	Workflow of the data cleansing activity, based on [19]	42
5.4	Example of RDF triples of ProvenBiO	43
5.5	Elements that compose the Provenance Manager	44
5.6	SPARQL query corresponding to Item 5	46
5.7	Screen copy of our query prototype. The code for Q2 is partially shown in the window.	47

Chapter 1

Introduction

eScience concerns joint research in computer science with scientists of other domains for the development of models, tools and techniques that support these scientists to develop their own research faster, better or in a different way. This also fosters new results in Computer Science, as witnessed by this text, but the emphasis is to contribute to other areas. Such efforts have brought new research challenges, with focus on the management of knowledge on a global scale. Common challenges that researchers in eScience need to cope with are related to the management of data heterogeneity issues, the supporting of sharing of data and ensuring the quality of the findings produced by scientific studies. This thesis concentrates on this last challenge.

Scientists are aware that the better the data they use are, the better results they can obtain in their investigations. However, the question is how to know whether the data are good enough? In particular, one has to consider that data are represented in different formats and scales, submitted to different transformation processes, and come from a variety of sources. Besides that, data quality needs to be understood within a context. Researchers must be able to show their results in this context, so that such results can be understood and reproduced by other scientists in a reliable manner.

Work related to data quality distinguishes different quality attributes, frequently relevant to business or geospatial data [43, 17, 33]. Quality attributes like accuracy and completeness are often quantitatively measured, while others like reliability are more often qualitatively measured. Qualitative and quantitative approaches depend on how quality attributes are captured and assessed. A variety of approaches ranging from mathematical formula to machine learning techniques are suggested to assign a score to each quality attribute. Another research trend is to analyze the quality of data by tracking their history [25, 68, 35]. Though interesting, most of these strategies are developed to focus on a specific quality attribute, thus reducing the applicability of the solutions.

Given these issues, this thesis has the following goals: (G1) the definition of the

data quality dimensions more interesting to the scientific domain; (G2) the management of data provenance aiming at the assessment of quality in a specific domain; and (G3) the enrichment of data provenance to provide a greater amount of information to help scientists in the assessment of quality. For this thesis, data provenance corresponds to the origins and the transformation processes applied to a dataset until it is (re)used in some experiment. Though provenance attributes can be considered as being part of attributes that contribute to quality, such attributes are treated apart in quality assessment (e.g., as historical information). In other words, related work either considers provenance to assess quality (which we call provenance-based) or disregards it, considering other attributes (a trend we call attributed based). Under this perspective, our work can be considered as provenance based.

Aiming at understanding how provenance can be used to assess quality, we started by surveying models and standards for provenance, to identify strategies to represent and manage provenance. At the same time, we conducted a broad survey of quality requirements for distinct scientific domains. As a result of the first study, we designed two conceptual provenance models, each of which with distinct characteristics, to support the assessment of data quality. The analysis of different scientific domains enabled us to specify a methodology to assess data quality based on provenance. However, quality is user and domain dependent. Thus, using the agriculture domain as a scenario, we studied some quality attributes/dimensions that researchers can take into consideration when developing their applications. This exercise gave us a better insight on procedures and methodologies for provenance-based quality assessment.

From these studies, we specified an extensible framework that can be used by scientists to assess the quality of the data produced by their experiments. Provenance, in our framework, follows efforts started in the context of scientific workflows systems such as Kepler [39], Taverna [79] and Vistrail [82]. Besides other functionalities, these workflow systems have elements that allow to record provenance of tasks performed at each step in experiments. The difference between ours and these other approaches is that we do not require experiment execution by workflow engines. Moreover, we take a step forward by showing that, by correlating quality dimensions and domain provenance, scientists can obtain information that can be used to evaluate the quality of their datasets.

Taking this scenario into account, the main contributions of this thesis in the context of the goals are:

- Investigation of the characteristics that highlight data quality issues in eScience. We identified and discussed different data quality dimensions that are common to a variety of scientific domains. In particular, we found that some of the most common dimensions that predominate in eScience are accuracy, completeness, timeliness, consistency, accessibility and relevancy. Furthermore, these dimensions were studied

in the context of applications in agriculture. Here we found that to better assess the quality of the dimensions, it is necessary to characterize them using sub-dimensions regarding the activities and the intended use of data.

- Specification of a framework that combines a provenance model to keep track of data provenance with a methodology that addresses the utilization of provenance to help scientists on the assessment of the datasets that are used and produced in their studies. Our framework highlights the characterization of the scientific processes and the capture and storage of the provenance information.
- Identification and specialization of generic provenance models considering data quality issues. One concern of this investigation was to study strategies that enable the capture, representation and storage of provenance information. As a result, we proposed two different models to capture provenance. The first, discussed in chapter 4, is based on OPM (Open Provenance Model) [56], while the second, discussed in chapter 5, is a semantic model based on PROV-O (PROV Ontology) [84]. OPM has the advantage of encompassing elements that can be easily adapted in specific domains by using standards. However, the rule defined by OPM highlights that the state of an artifact cannot be modified after its creation. PROV-O, on the other hand, is ontology-based and therefore dynamic and extensible.
- Discussion and application of the above findings in two case studies, in agriculture and biodiversity, thereby showing how scientists can take advantage of our proposal to assess data quality in distinct contexts. We showed how the stages that encompass our methodology are linked to our framework and how both of them may be used considering the peculiarities of each domain.

This thesis is organized as a collection of papers, as follows:

Chapter 2 is the paper *A Provenance Approach to Assess the Quality of Geospatial Data*, published in the Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC) in 2012 [53]. This chapter identifies some challenges and requirements that need to be considered to model provenance. Our study of the Open Provenance Model (OPM) [56] identified characteristics that can be used to accommodate provenance information related to geospatial data. From this study, we obtained a simple provenance model that specialists can combine with geospatial data quality attributes, in order to assess the quality of datasets. This model is a first view of the problem that was subsequently refined in Chapter 4.

Chapter 3 is the paper *Data Quality in Agriculture Applications*, published in the Proceedings of the XIII Brazilian Symposium on GeoInformatics (Geoinfo) in 2012 [50], and which received the 3rd prize in the conference. This chapter surveyed data quality

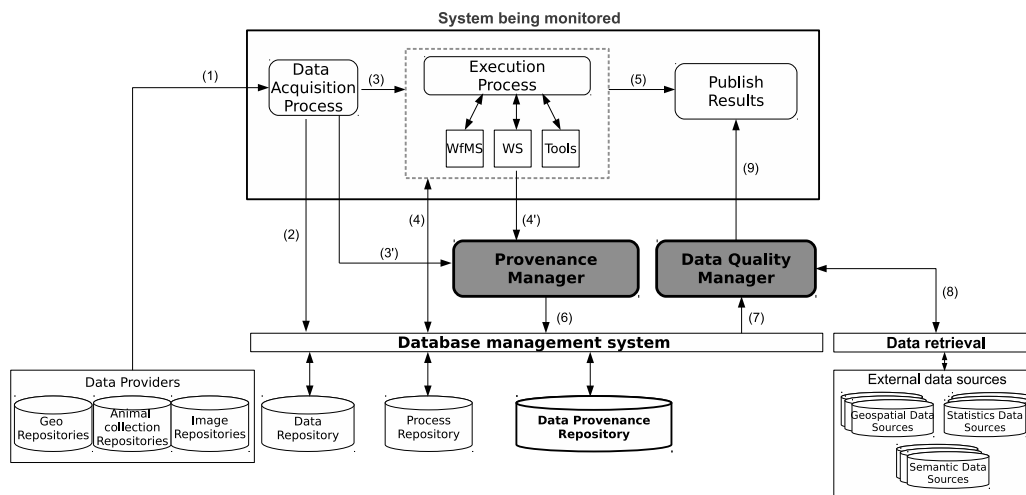


Figure 1.1: The Architecture of the Framework

efforts in agriculture and geospatial science. In order to help researchers to develop better applications, we investigated the different dimensions of quality focusing on the approaches that are used to evaluate them. The chapter shows that in scientific domains, such as agriculture, some of the most common dimensions are accuracy, completeness, timeliness, consistency, relevancy and accessibility. These dimensions can cover a variety of sub-dimensions, in order to better describe the quality characteristic of a data set. This proposal was validated by domain experts.

Chapter 4 is the paper *A Provenance-based Approach to Evaluate Data Quality in eScience*, that has been submitted to the Journal on Metadata, Semantics and Ontology (IJMSO) [51]. This chapter presents the elements that compose our framework to support the assessment of quality. We provide a database schema for data provenance that relies on OPM, and propose a methodology to evaluate the quality of a digital artifact based on its provenance. This methodology relies on user expertise to define and tune dimensions, and to analyze quality. Figure 1.1, repeated from that chapter, gives an overview of the general framework. The highest level represents the application that is being monitored, which encompasses three steps: data acquisition, transformation processes and publishing of the results. In the intermediate level we have the Provenance Manager that is in charge of identifying, capturing and storing the provenance information. Moreover, the Data Quality Manager allows to query the information stored in the Data Provenance Repository based on requests performed by the specialists. When necessary, this module looks for other information in order to complement the information taken from the provenance repository. For the development of the Data Provenance Repository, at the lower level, we extended a generic model known as OPM.

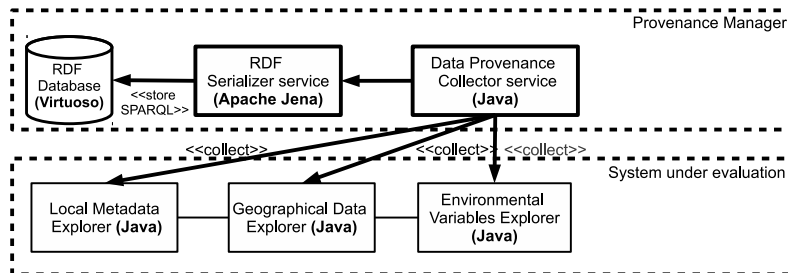


Figure 1.2: Elements that compose the Provenance Manager

Chapter 5 is the paper *Estimating the quality of data using provenance: a case study in eScience*, that has been accepted for publication in the Proceedings of the 19th Americas Conference on Information Systems (AMCIS) [52]. This chapter presents a semantic model to preserve data provenance, ProvenBiO, which extends the PROV-O semantic provenance model so that it can be used in biodiversity applications. We show how domain-specific provenance can improve the process of assessment of quality, and implement a (database) query-based approach to elicit provenance and quality information. We also present some technologies and implementation issues that we adopted to validate our solution. Figure 1.2, extracted from Chapter 5, shows an instantiation of the Provenance Manager of Figure 1.1.

Chapter 6 contains conclusions and some directions for future work.

Besides these papers, this research also produced the following paper: *Handling Provenance in Biodiversity*, published on the Workshop on Challenges in eScience (CIS) [49].

Chapter 2

A Provenance Approach to Assess the Quality of Geospatial Data

2.1 Introduction

We use geospatial data everyday and everywhere. Regardless of the application domain, data collected are manipulated by a wide range of users, with distinct research interests, using their own vocabularies, work methodologies, models, and sampling needs. In particular there is a huge effort to improve the means and methodologies to capture, process and disseminate geospatial data. This information, when adequately described and documented, would help end-users to assess the trustworthiness of an analysis process or a report, and understand the activities associated with in studies involving a given data source [10].

The tracking of historical information concerning a data set is also known as *data provenance*. In the scientific community, *data provenance* has become a basis to determine authorship, data quality, and to allow the reproducibility of findings [77]. In real life situations, provenance information of geospatial data is used to decide pre-processing procedures, storage policies and even data cleaning strategies – with direct impact on data analysis and synthesis policies.

Trust and quality go hand-in-hand. Taking this into account, our work describe a geospatial data provenance model to help to determine whether (and how much) users can trust data sources and data providers, and to assess data quality. Our solution takes advantage of features provided by the Open Provenance Model (OPM) [56] and FGDC geographic metadata standards [30].

2.2 Model overview

The basic premise of our work is that, given its importance, geographic information needs to have elements which allow to know whether the data are reliable, so that it can be consumed. Our second premise is that, once data provenance can be used to estimate data quality, we can use provenance as a means to assess trustworthiness. For instance, if the data to consider is a map, we need to face qualitative (e.g., mapping methodologies) and quantitative (e.g, resolution) factors. Furthermore, we need to know the level of reliability of the entities involved in the data collection (e.g., providers) and analysis activities used to produce the map.

Our research considers the *trustworthiness of source* and *temporality* dimensions of data quality of [68]. *Trustworthiness of sources* (who) refers to the degree of confidence of who created or made available the data. *Temporality of data* (when) includes valid and transaction time. Besides *who* and *when*, we also need to capture the location where a event has happened, i.e *where*.

Figure 2.1 illustrates the main elements of our provenance data model using the entity relationship notation. The part in bold comes from OPM, the rest was added by us. The basic pieces of the model are *Artifact*, *Process* and *Agent*. While the Artifact entity concerns geospatial data products, the Process entity deals with the processes that generated an Artifact. Finally, the Agent entity is in charge of executing processes or providing artifacts. In our model, trust criteria are associated to an Artifact and an Agent and have normalized values ranging from 0 to 1.

Examples of artifacts in this work are a remote sensing image or the level of erosion derived from analysis of this image. An Artifact can be provided by an Agent, for example, an official institution like NASA or Brazil’s National Geographic Institute (IBGE), or may be the result the execution of a process. A Process is controlled by an Agent and it also might trigger subprocesses.

Our model considers that at a specific time a process can have several inputs, but can only generate one outcome. In the geospatial domain, in some cases, the trustworthiness and quality of a source decay with age. Therefore, Valid time concerns an Artifact and Transaction time concerns a Process. *URL Address* links an Artifact to its location in a database or directory file. We assume that data related to geographic coordinates or another kind of spatial features are stored in spatial repositories provided by an Agent. *Measure criteria* about data quality have been taken from the FGDC metadata standard [30] and linked to an Artifact. An Agent uses and applies some methodologies according to the domain where it works. The grade of trust (*Trust Grade*) of an Agent depends on issues such as: is it an official source, the reputation of this provider, is it an academic research group. This scenario shows that assigning a confidence value to an agent can be

very subjective.

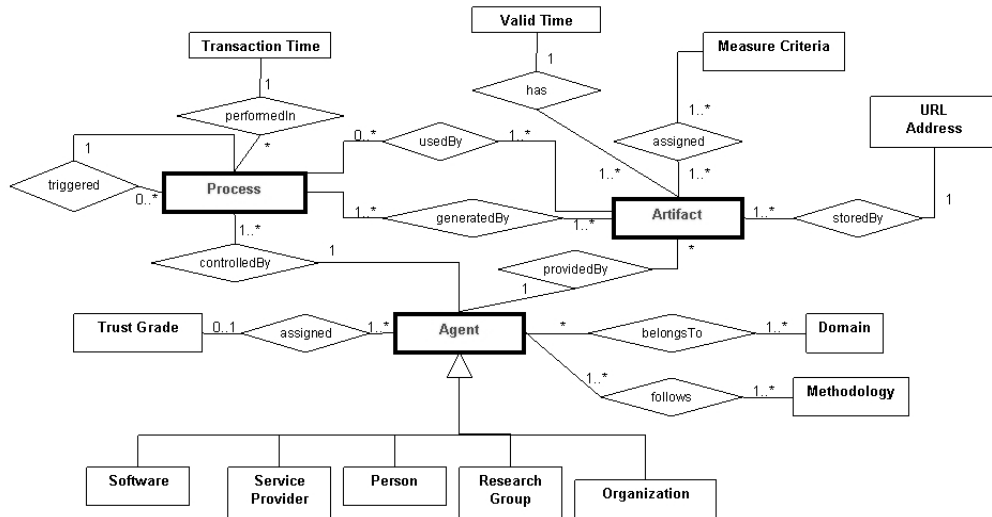


Figure 2.1: Our provenance model

2.3 Quality elements

FGDC [30] provides a set of terms to document digital geospatial data, with several meta-data criteria. We selected the most relevant criteria, taking into account our experience in agricultural planning and monitoring based on processing remote sensing sources (satellite images). These parts are:

- *Positional accuracy*: refers to the accuracy of the positions of spatial objects.
- *Logical consistency*: indicates the fidelity of relationships in the data set and tests used.
- *Completeness*: is information about omissions, selection criteria, generalization, definitions used, and other rules used to derive a data set.
- *Attribute accuracy*: indicates how thoroughly and correctly the features in the data set are described.

Though these are the basic metadata elements that we selected, we can add other elements (e.g., coverage, horizontal accuracy) that complement them. Each of these criteria must be assigned quantifiers, i.e. a value obtained from computing the quality of the

attributes related to the Artifact. However, tuning these quantifiers is not a trivial work and depends on the usage for which the Artifact is intended. As a first step, we begin by assigning trust values ranging from 1 to 0 to the Agent. This means that the higher the trust value is, the most reliable an Agent is.

We are conducting case studies in agriculture to validate our model, storing quality information in database tables. This database is created from the ER diagram in figure 2.1. In such examples, input data concerns satellite images, crop information and others. Outputs include maps and reports, produced after several manual and automatic processing steps. All these are taken into consideration in provenance and quality evaluation.

2.4 Conclusions

Geospatial data are a basis for decision making activities that affect our daily lives. The trustworthiness of these data (and recommendations based on analyses thereof) is becoming increasingly important. This is complicated by the fact that the processing of geospatial data is essentially a cooperative, distributed, effort, which hampers determining its reliability. Most efforts to improve this situation concentrate on establishing documentation about data capture, methodologies, curation standards and quality metadata.

This paper presented a novel approach based on data provenance for alleviating this problem. Our provenance model takes advantage of features provided by the Open Provenance Model, which are being used by the scientific community to instantiate their solutions. The model integrates concepts from the FGDC metadata standard needed for assessment of data quality.

Chapter 3

Data Quality in Agriculture Applications

3.1 Introduction

Agriculture is an important activity for economic growth. In 2011, agricultural activities contributed approximately with 22% of Brazil's Gross National Product [11]. Thus there are major benefits in ensuring the quality of data used by experts and decision makers to support activities such as yield forecast, monitoring and planning methods. The investigation of ways to measure and enhance the quality of data in GIS and remote sensing is not new [16, 54, 45, 17]. The same applies to data managed in, for instance, Information Manufacturing systems [2]; Database systems [87], Web systems [35]; or Data Mining systems [7]. All of these fields are involved in and influence agriculture applications.

Despite these efforts, data quality issues are not often taken into account when different kinds of databases or information systems are modeled. Data produced and reported by these systems is used without considering the defects or errors that data contain [12, 33]. Thus, the information obtained from these data is error prone, and decisions made by experts becomes inaccurate.

There are many challenges in ongoing data quality such as: modeling and management, quality control and assurance, analysis, storage and presentation [12]. The approach used to tackle each one of these issues depends on the application scenario and the level of data quality required for the intended use [81]. Thus, understanding what attributes of quality need to be evaluated in a specific context is a key factor.

This paper presents a brief review from the literature related to issues about data quality with special consideration to data managed in agriculture. The goal is to provide a conceptual background to become the basis for development of applications in agriculture.

3.2 Data for agriculture applications

Data in agriculture applications can be thematic/textual or geospatial, from primary to secondary sources, raw or derived. Thus, rather than just analyzing issues concerning the quality of geospatial data, this paper considers quality in all kinds of data, and provides guidelines to be applied for agriculture applications.

Research related to data quality in agriculture considers several issues. There are papers that concentrate on agricultural statistics data (e.g., production and consumption of crops) like [18] and [41]. The efforts that have been made to study the quality of geospatial data [30, 37, 17, 33] are also taken advantage of in the agriculture domain. However, there are other kinds of data that need to be considered such as files containing sensor-produced data, crop characteristics and soil information, human management procedures, among others [27].

This general scenario shows that agricultural activities encompass different kinds and sets of data from a variety of heterogeneous sources. In particular, the most common kinds of data are regular data and geospatial data. Regular data can be textual or numeric and can be stored on spreadsheets or text files (e.g., crop descriptions from official sources). Geospatial data correspond to georeferenced data sources and can include both raster and vector files, for example, satellite images using GeoTIFF format or a road network on shapefiles. Geospatial data may also come in data streams [1] - packets of continuous data records - that can be obtained from aboard satellites, ground sensors or weather stations (e.g., temperature readings). All these data need different levels of access and manipulation and thus pose several challenges about data quality.

3.3 Dimensions of data quality

Data quality has various definitions and is a very subjective term [12]. A broad and consensual definition for data quality is “fitness for use” [16]. Following this general concept, [85] extended this definition as *data that are fit for use by data consumers*, i.e. those who use the data. Redman [69] complements the data quality concept by claiming that data are fit to be used if they are free of defects, accessible, accurate, timely, complete, consistent with other sources, relevant, comprehensive, provide a proper level of detail, and easy to read and interpret. Quality is context-based: often data that can be considered suitable for one scenario might not be appropriate for another [2].

Data quality is seen as a multi-dimensional concept [85, 2, 7]. Quality dimensions can be considered as attributes that allow to represent a particular characteristic of quality [85]. In particular, accuracy, completeness, timeliness and consistency have been extensively cited in the literature as some of the most important quality dimensions to

information consumers [85, 63]. Correctness, reliability and usability are interesting in areas like simulation modeling process, as discussed in [74].

Wang and Strong [85] classified fifteen dimensions of quality grouped in four main categories - see Table 3.1. Dimensions accuracy, believability, objectivity and reputation are distinguished as *intrinsic data quality*. Timeliness and completeness are examples of *contextual data quality*. Interpretability and consistency describe features related to the format of the data and are classified as *representational data quality*. Accessibility and security are labeled as *accessibility data quality*, highlighting the importance of the role of information systems that manage and provide access to information.

Table 3.1: The 15 dimensions framework [85]

Category	Dimensions
Intrinsic DQ	Believability
	Accuracy
	Objectivity
	Reputation
Contextual DQ	Value-added
	Relevancy
	Timeliness
	Completeness
	Appropriate amount of data
Representational DQ	Interpretability
	Ease of understanding
	Representational consistency
	Concise representation
Accessibility DQ	Accessibility
	Access security

Table 3.2: The PSP/IQ model [43]

	Conforms to Specifications	Meets or exceeds Consumer expectations
Product Quality	Free-of-error	Appropriate amount relevancy
	Concise representation	Understandability
	Completeness	Interpretability
	Consistent representation	Objectivity
Service Quality	Timeliness	Believability
	Security	Accessibility
		Ease of operation
		Reputation

The model of [43], Product Service Performance Information Quality (PSP/IQ), consolidates Wang and Strong's framework. Their goal is to represent information quality aspects that are relevant when decisions for improvement of information quality need to be made. Table 3.2 presents the PSP/IQ model showing that information quality can be assessed from the viewpoint of product or service and in terms of the conformance of data to the specifications and consumer expectations.

According to [61] three main factors influence the quality of information: the user's perception, the information itself, and the process to retrieve the information. Based on these factors, the authors classify information quality criteria in 3 classes: *Subject-criteria*, *Object-criteria* and *Process-criteria*. Subject-criteria are those that can be determined by users' personal views, experience, and backgrounds. Object-criteria are specified through the analysis of information. Process-criteria are related to query processing. Table 3.3 shows their list of quality criteria grouped by classes, together with suggested assessment methods for each quality criterion.

USAID [81] provides practical advices and suggestions on issues related to performance monitoring and evaluation. It highlights five quality dimensions: validity, reliability, precision, integrity, and timeliness.

In summary, the concept of quality encompasses different definitions and its dimensions (or attributes) can be generic or specific and this depends on the application domain.

3.4 Data Quality Measurement

A significant amount of work addresses the measurement of the quality of data and information. The distinction between data and information is always tenuous. Although there is a tendency to use information as data that has been processed and interpreted to be used in a specific context - e.g., economics, biology, healthcare - data and information are often used as synonymous [67]. According to [59], information quality measurement is the process of assigning numerical values, i.e. scores, to data quality dimensions. Related work differentiate between manual and automatic measurement of data quality. Manual approaches are based on the experience and users' point of view, i.e. a subjective assessment. Automatic approaches apply different techniques (e.g., mathematical and statistical models) in order to compute the quality of data. There follows an overview of work that investigates these topics.

3.4.1 Manual approaches

Lee et al. [43] measure information quality based on 4 core criteria to classify information: soundness, dependability, usefulness, and usability. Each class includes different quality

Table 3.3: The classification of [61]

Class	Quality Criterion	Assessment Method
Subject Criteria	Believability	User experience
	Concise representation	User sampling
	Interpretability	User sampling
	Relevancy Continuous	User assessment
	Reputation	User experience
	Understandability	User sampling
	Value-Added	Continuous user assessment
Object Criteria	Completeness	Parsing, sampling
	Customer	Support Parsing, contract
	Documentation	Parsing
	Objectivity	Expert input
	Price	Contract
	Reliability	Continuous assessment
	Security	Parsing
	Timeliness	Parsing
Verifiability	Expert input	
Process Criteria	Accuracy	Sampling, cleansing techniques
	Amount of data	Continuous assessment
	Availability	Continuous assessment
	Consistent representation	Parsing
	Latency	Continuous assessment
	Response time	Continuous assessment

dimensions. For instance, soundness encompasses: free-of-error, concise and consistent representation and completeness. The authors apply a survey questionnaire to the users to obtain scores for each criterion ranging from 0 to 1. The interpretation of the quality measure is made using gap analysis techniques. Bobrowski et al. [8] suggest a methodology also based on questionnaires to measure data quality in organizations. Quality criteria are classified as direct or indirect. Direct criteria are computed applying software metrics techniques and these are used to derive the indirect criteria.

While [43] and [8] rely on questionnaires and users' perspective to obtain quality criteria scores, the methodology of [66] uses control matrices for data quality measurement. The columns in the matrix are used to list data quality problems. Rows are used to record quality checks and corrective processes. Each cell measures the effectiveness of the quality check at reducing the level of quality problems. Similarly to [43] and [8], this methodology also requires users' inputs to identify how well the quality check performs its function.

Volunteered geographic information (VGI) is a mechanism for the acquisition and compilation of geographic data in which members of the general public contribute with geo-referenced facts about the Earth's surface to specialist websites where the facts are processed and stored into databases. Goodchild and Li [33] outline three alternative solu-

tions to measure the accuracy of VGI – crowd-sourcing, social, and geographic approaches.

The crowd-sourcing approach reflects the ability of a group of people to validate and correct the errors that an individual might make. The social approach is supported by a hierarchy of a trusted group that plays the role of moderators to assure the quality of the contributions. This approach may be aided by reputation systems as a means to evaluate authors' reliability. The geographic approach is based on rules that allow to know whether a supposed geographic fact is true or false at a given area.

3.4.2 Automatic approaches

Examples of work that use automatic approaches to measure data quality include [2] and [88]. Ballou et al. [2] present an approach for measuring and calculating relevant quality attributes of products. Xie and Burstein [88] describe an attribute-based approach to measure the quality of online information resources. The authors use learning techniques to obtain values of quality attributes of resources based on previous value judgments encoded in resource metadata descriptions.

In order to evaluate the impact of data quality in the outcomes of classification - a general kind of analysis in data mining - [7] compute metrics for accuracy, completeness, consistency and timeliness. Shankaranarayanan and Cai [75] present a decision-support framework for evaluating completeness. Parsian [63] provides a sampling methodology to estimate the effects of data accuracy and completeness on relational aggregate functions (*count*, *sum*, *average*, *max*, and *min*). Madnick and Zhu [47] present an approach based on knowledge representation to improve the consistency dimension of data quality.

Although not always an explicit issue, some authors present the possibility to derive quality dimensions using historic information of data, also known as provenance. For instance, the computing of timeliness in [2] is partially based on the time when a data item was obtained. Examples of work that have a direct association between quality and data provenance are [68], [20] and [35]. Prat and Madnick [68] propose to compute the believability of a data value based on the provenance of this value. The computation of believability has been structured into three complex building blocks: metrics for measuring the believability of data sources, metrics for measuring the believability from process execution and global assessment of data believability. However, the authors only measure the believability of numeric data values, reducing the applicability of the proposal.

Dai et al. [20] present an approach to determine the trustworthiness of data integrity based on source providers and intermediate agents. Hartig and Zhao [35] present a method for evaluating the timeliness of data on the Web and also provide a solution to deal with missing provenance information by associating certainty values with calculated timeliness values. Table 3.4 shows a summary with the quality dimensions studied in automatic

approaches together with the application domain where the dimensions are considered.

Table 3.4: Summary of quality dimensions covered by automatic approaches

Work	Quality Dimension studied	Data managed by
[Ballou et al. 1998]	Accuracy and timeliness	Information Manufacturing System
[Shankaranarayanan and Cai 2006]	Completeness	Decision support system
[Parssian 2006]	Accuracy and completeness	Databases
[Madnick and Zhu 2006]	Consistency	Databases
[Prat and Madnick 2008]	Believability	Databases
[Dai et al. 2008]	Trustworthiness	Databases (data integrity)
[Hartig and Zhao 2009]	Timeliness	Web
[Xie and Burstein 2011]	Reputation	Web (Health Information Portals)
[Blake and Mangiameli 2011]	Accuracy, completeness, consistency and timeliness.	Databases

3.5 Data Quality in Applications in Agriculture

Considering the impact that agriculture has on the world economy, there is a real need to ensure that the data produced and used in this field have a good level of quality. Efforts to enhance the reliability of agricultural data encompass, for example, methodologies for collection and analysis of data, development of novel database systems and software applications.

Since prevention is better than correction, data collection and compilation are some of the first quality issues that need to be considered in the generation of data that are fit for use [12]. For instance, non-reporting data, incomplete coverage of data, imprecise concepts and standard definitions are common problems faced during the collection and compilation of data on land use [28].

Statistical techniques and applications are being used to produce agricultural statistics such as crop yield production, seeding rate, percentage of planted and harvested areas, among others. One example is the CountrySTAT framework [18]. This is a web-based system developed by the Food and Agriculture Organization of the United Nations [29]. It integrates statistical information for food and agriculture coming from different sources. The CountrySTAT is organized into a set of six dimensions of data quality that are: relevance and completeness, timeliness, accessibility and clarity, comparability, coherence, and subjectiveness.

Other example is the Data Quality Assessment Framework (DQAF) [36] that is being used as an international methodology for assessing data quality related to the governance

of statistical systems, statistical processes, and statistical products. It is organized around a set of prerequisites and five dimensions of data quality that are: assurance of integrity, methodological soundness, accuracy and reliability, serviceability, and accessibility.

Based on both the CountrySTAT and the DQAF frameworks, [41] proposed the Agricultural Data Quality Assessment Framework (ADQAF) aiming at the integration of global and national perspectives to measure the quality of agricultural data. It encompasses quantifiable (e.g., accuracy and completeness) and subjective (e.g., relevance and clarity) quality dimensions.

Because of the relevance that land data plays in agriculture (e.g., for crop monitoring or planning for sustainable development), it is necessary to consider data quality issues in the development of agricultural land-use databases. According to [28] the value of land-use databases is influenced by their accuracy, coverage, timeliness, and structure. The importance to maintain suitable geo-referenced data is also recognized.

Since agriculture applications rely heavily on geospatial data, one must consider geospatial metadata standards such as [37] and [30], which have been developed aiming at the documentation and exchange of geospatial data among applications and institutions that use these kind of data. ISO 19115 [37] defines a data quality class to evaluate the quality of a geospatial data set. Besides the description of data sources and processes, this class encompasses positional, thematic and temporal accuracy, completeness, and logical consistency. The FGDC metadata standard includes a data quality section allowing a general assessment of the quality of the data set. The main elements of this section are attribute accuracy, logical consistency report, completeness report, positional accuracy, lineage and cloud cover.

Congalton and Green [17] highlight the need to incorporate positional and thematic accuracy when the quality of geospatial data sets like maps are evaluated. Positional accuracy measures how closely a map fits its true reference location on the ground. Thematic accuracy measures whether the category labeled on a map at a particular time corresponds to the true category labeled on the ground at that time. According to [33] accuracy dimension is also an important attribute in the determination of quality of VGI. This approach is acquiring importance in all domains where non-curated data are used, including agriculture. Beyond accuracy, precision is also an important quality attribute that needs to be considered. Chapman [12] distinguishes statistical and numerical precision. The first one reflects the closeness to obtain the same outcomes by repeated observations and/or measurements. The last one reflects the number of significant digits with which data is recorded. It can lead to false precision values - e.g., when databases store and publish data with a higher precision than the actual value.

Completeness in the context of geospatial data encompasses temporal and spatial coverage [37, 30]. Coverage reflects the spatial or temporal features for geospatial data.

For instance, [3] use the spatial coverage dimension to determine whether a dataset covers (fully or partially) an area of interest.

Remote sensing is another major source of data for agriculture applications, in particular satellite or radar images. Image producers, such as NASA or INPE, directly or indirectly provide quality information together with images - e.g., dates (and thus timeliness), or coordinates (and thus spatial coverage). FGDC's cloud cover is an example of metadata field for images. Methodologies to measure quality of an image set combine manual and automatic processes (e.g., see [55] concerning the cleaning of invalid pixels from a time series of satellite images, to analyze sugar cane yield). Information concerning the sensors aboard satellites is also used to derive quality information. Analogously, information concerning ground sensors is also taken into account.

3.6 Summing up

We distinguish two groups of quality dimensions: qualitative and quantitative - see Table 3.5. We use the dimensions identified by [85], since these authors are the most referenced in the literature.

Qualitative dimensions are those that need direct user interaction and their measurement is based on the experience and background of the measurer. This measurement can be supported by statistical or mathematical models [67]. On the other hand, quantitative dimensions can be measured using a combination of computing techniques - e.g., machine learning, data mining - and mathematical and/or statistical models [48]. For instance, simple ratios are obtained measuring the percentage of data items which meet with specific rules [7]. Parsing techniques consider how the information are structured in a database, in a document, etc [61]. There are dimensions such as believability and accuracy that can be evaluated combining manual and automatic approaches. Choosing the best strategy for measuring the quality of data depends on the application domain and the dimensions of interest for that domain.

Table 3.6 shows the most common quality dimensions investigated by research reviewed in the previous sections. We observe that the most frequent quality dimensions studied in the literature are accuracy, timeliness and completeness, followed by consistency and relevancy. Beyond these dimensions, accessibility is also of interest to agriculture field. This set of dimensions can become the basis to evaluate the quality of data in agricultural applications.

As we have seen, agricultural applications cover a wide variety of data. How to measure and enhance the quality of these data becomes a critical factor. It is important to adopt strategies and rules that allow to maintain the quality of data starting from the collection, consolidation, and storage to the manipulation and presentation of data. Common errors

Table 3.5: Classification of quality dimensions

Dimensions of quality	Qualitative	Quantitative	Type of approach	Example of approach
Believability	x	x	Manual	user feedback
			Automatic	mathematical models
Objectivity	x		Manual	user feedback
Reputation	x		Manual	user experience
Value-added	x		Manual	user feedback
Relevancy	x		Manual	questionnaires
Interpretability	x		Manual	user experience
Ease of understanding	x		Manual	user feedback
Concise representation	x		Manual	user feedback
Accuracy	x	x	Manual	crowd-sourcing
			Automatic	cleansing techniques
Timeliness		x	Automatic	mathematical models
Completeness	x	x	Manual	control matrices
			Automatic	parsing
Consistent representation		x	Automatic	parsing
Access security		x	Automatic	mathematical models
Accessibility		x	Automatic	mathematical models
Appropriate amount of data		x	Automatic	mathematical models

that need to be tackled are related to missing data, duplicate data, outdated data, false precision, inconsistency between datums and projections, violation of an organization's business rules and government policies, among others.

Table 3.7 summarizes the main quality dimensions considered in agriculture, according to our survey. The table shows the dimensions that predominate in the literature and the context where they can be applied. It also shows that some dimensions include other quality attributes to encompass different data types - e.g., completeness for geospatial context is described in terms of spatial and temporal coverage. We point out that most dimensions are common to any kind of application. However, like several other domains, agriculture studies require analysis from multiple spatial scales and include both natural factors (e.g., soil or rainfall) and human factors (e.g., soil management practices). Moreover, such studies need data of a variety of types and devices. One of the problems is that researchers (and often practitioners) concentrate on just a few aspects of the problem.

For instance, those who work on remote sensing aspects seldom consider ground-based sensors; those who perform crop analysis are mainly concerned with biochemical aspects. However, all these researchers store and publish their data. Correlating such data becomes a problem not only because of heterogeneity issues, but also because there is no unified concern with quality issues and the quality of data is seldom made explicit when data are

Table 3.6: Main data quality dimensions studied for the related work

Quality Dimension (QD)	Papers that studied these QD
Believability	[Prat and Madnick 2008]
Reputation	[Xie and Burstein 2011]
Reliability/Trustworthiness	[Dai et al. 2008], [Bobrowski et al. 1999] and [U.S. Agency for International Development 2009]
Relevancy	[CountrySTAT 2012], [Kyeyago et al. 2010], [FAO 1997] and [Bobrowski et al. 1999]
Ease of understanding	[Kyeyago et al. 2010]
Accuracy	[Ballou et al. 1998], [FGDC 1998], [ISO 19115 2003], [Parssian 2006], [Blake and Mangiameli 2011], [Kyeyago et al. 2010], [FAO 1997], [Bobrowski et al. 1999] and [Congalton and Green 2009].
Timeliness	[Ballou et al. 1998], [Hartig and Zhao 2009], [U.S. Agency for International Development 2009], [Blake and Mangiameli 2011], [CountrySTAT 2012], [FAO 1997] and [Bobrowski et al. 1999]
Completeness	[FGDC 1998], [ISO 19115 2003], [Shankaranarayanan and Cai 2006], [Parssian 2006], [CountrySTAT 2012], [Kyeyago et al. 2010], [Bobrowski et al. 1999] and [Barbosa and Casanova 2011].
Consistency	[FGDC 1998], [ISO 19115 2003], [Madnick and Zhu 2006], [Blake and Mangiameli 2011] and [Bobrowski et al. 1999]
Accessibility	[CountrySTAT 2012] and [Kyeyago et al. 2010]

published. This paper is a step towards trying to minimize this problem, by pointing out aspects that should be considered in the global view. As mentioned before, these issues are not unique to agriculture applications and can be found in, for instance, biodiversity or climate studies.

Table 3.7: Main data quality dimensions in agriculture applications

Quality dimensions	Context	Example of kinds of data
Accuracy:	Relational databases, statistical information and data files	table, tuple, attribute, query, yield information, production of crops, growth rate, XML files, spreadsheets documents, etc.
Positional and Thematic accuracy	Geospatial datasets	geographic coordinates, VGI, satellite images, maps, aerial photography, etc.
Completeness:	Relational databases, statistical information and data files	schema, column, attribute, population census, land data, rates of harvested areas, farm production, CVS text files, spreadsheets, etc.
Spatial and Temporal coverage	Geospatial datasets	cartographic materials, geographic coordinates, etc.
Timeliness	Information Manufacturing systems	age and shelf life of products, delivery time of products, etc.
	(Geographic) Information/Web systems and statistical information	access, creation or delivery time of data items, age of a data item, sensor data streams, population census, harvest dates, etc.
Consistency	(Geospatial) Databases	tables, data, maps, time series, reports and charts, etc.
Relevancy	Information systems, databases and statistical information	text and spreadsheets documents, census, historical weather datasets, trade information, etc.

Chapter 4

A Provenance-based Approach to Evaluate Data Quality in eScience

4.1 Introduction

One of the concerns in eScience research is the design and development of novel solutions to support global, collaborative and multidisciplinary work. One challenge that pervades all scientific domains is to ensure the quality of findings produced by scientific studies. Indeed, data with good quality are vital for scientific research. But how to measure quality?

Related work covers topics that range from data quality standards to a variety of data quality models and assessment methodologies - e.g., [85, 30, 68, 35, 70, 13, 44]. Nevertheless, in order to select the best standards and models to use and to evaluate quality, we need to understand the domain requirements and the intended use of the data. Data quality is a multi-dimensional concept and people that participate in the evaluation of quality must define what dimensions will be addressed, and this again depends on the application domain. Dimensions of quality (e.g., accuracy, completeness, reliability) may be considered as attributes that allow to represent a particular characteristic of quality [85].

This paper presents an approach to help attack the quality challenge, offering a solution to handle data quality issues. Our approach is based on the use of *data provenance*, i.e. the history of the origins and transformation processes applied to a given data product. Despite the vast research on data provenance in, e.g., databases [83, 10, 15], scientific workflow systems [79, 82] and the semantic Web [25, 34], there has been relatively little investigation to bridge the gap between data quality and provenance.

The main contributions of this research include: (i) the specification of a framework to track data provenance and use this information to derive quality information; (ii) a

model for data provenance based on the Open Provenance Model (OPM) [56]; and (iii) a methodology to help evaluate the quality of some digital artifacts based on its provenance.

Model and provenance information are translated into metadata that are stored in a database and associated with the corresponding data products. Model and methodology are validated against a real case study in agriculture, in which images used to monitor biomass for a given crop are analyzed against specific quality dimensions. Our choice of dimensions for this case study is based on a survey of quality criteria in eScience applications, directed towards agricultural problems, which has been published in [50].

The rest of this paper is organized as follows. Section 4.2 describes some issues related to metadata standards and data provenance. Section 4.3 presents our framework, describing details of the data provenance model and the methodology. Section 4.4 describes a case study in agriculture. Section 4.5 describes conclusions and future work.

4.2 Provenance and Metadata standards

Data provenance can be used for several purposes, such as to estimate data quality; to support the audit trail of data, by tracking steps involved in the processing of data; to repeat data derivation processes; to establish data ownership and liability and data discovery. Some characteristics associated to provenance are: (i) approaches to collect it; (ii) approaches to represent it; and (iii) strategies to store and means to disseminate provenance [77]. However, to be effectively used, provenance needs to be digitally discoverable, accessible, comprehensible, and provide necessary context information to reproduce data analysis results [57].

This, in turn has prompted research in the capture and storage of provenance. Provenance may be entered manually by experts (e.g., when documenting the provenance of data also provided by third parties). Often provenance is captured from executions logs (e.g., transaction history in database systems [31]). In eScience, workflow management systems (WfMS) allow traceability of process execution. Moreover, they also enable reproducibility of experiments, thereby enabling monitoring and provenance checking. Many workflow management systems provide traceability (and thus provenance) information at different levels of granularity [4, 79, 82]. For this reason, our framework relies on workflow to help extract provenance information.

There remains the issue of representing and storing provenance. Such information is often seen as a kind of metadata used to describe the derivation history of a data product. It represents the *who*, *what*, *when*, *where*, *why* and *how* associated with a resource. There are several metadata standards that were designed to be applied to specific domains and that include information that can be used to describe provenance. An example is the spatial metadata standard provided by the Federal Geographic Data Committee [30],

which provides a set of elements that help to obtain provenance (e.g., identification, spatial data organization, attribute information). Though provenance attributes can be considered as being part of attributes that contribute to quality, such attributes are treated apart in quality assessment (e.g., as historical information). In other words, related work either considers provenance to assess quality (which we call provenance-based) or disregards it, considering other attributes (a trend we call attributed based). Under this perspective, our work can be considered as provenance based.

While metadata standards can be used to record provenance facts, there remains the problem of modeling provenance. Our such approach is the Open Provenance Model (OPM) [56]. It is composed by three basic entities: agent, artifact and process. OPM is domain independence, since it does not standardize metadata about processes or data products. Other models include the work of Hartig and Zhao [35] and the PROV Ontology (PROV-O) that is a candidate to become a standard for W3C [84]. The work of Hartig and Zhao [35] concerns provenance of data from the Web, aiming to support the assessment of data quality such as timeliness in this context. PROV-O is an ontology based on OWL2 that specifies a data model to express provenance records generated in different systems and under different contexts. These and other initiatives like [71] aims at construction of a common provenance model to enhance interoperability.

As will be seen, we have chosen OPM as a basis for our work because it has been adopted and adapted in many contexts (e.g., [82, 78, 62, 80]). It supports our needs to collect and store provenance metadata related to the scientific processes performed by researchers. This information is delivered to the scientists to help evaluate the quality of results produced by them.

4.3 The Framework: using provenance information to support the evaluation of the quality of data

4.3.1 Architecture of the Framework

Basically, when data are processed by an eScience system, three stages can be distinguished [64]: (i) *data acquisition*, (ii) *analysis* and (iii) *publication of results*. *Data acquisition* is the process that allows collecting data associated with a study, for example, data related to land surface characterization. The *analysis* involves the processing of the data acquired; this step can be supported by a combination of software tools, reference books, Web sources, information from experts, among others. Finally, the *publication of results* delivers the results obtained - e.g., publishing files in the Web.

Figure 4.1 shows our architecture to assist in evaluating the quality of a data product delivered at stage (iii), based on data provenance at stages (i) and (ii). The system being

monitored for provenance information is inside the dark box on the top of the figure. Each box denotes different data access and manipulation levels. One important characteristic is that all data processing activities (from input to output) may be expressed through and handled either by a scientific workflow or by specialists who execute a set of independent processes and tools.

Our framework combines concepts of a database-centered model with work on traceability that was implemented for food supply chains [40]. The latter is centered on specifying a given supply chain using workflows, which invoke web services, i.e. the activities of the chain. Next, the services are executed following the workflow specification, monitoring all events in the chain and storing related metadata in a database (e.g., what process was used to transform a product, when this transformation occurred and which agent was responsible for the transformation). The workflow can be automatically executed by a WfMS or steered by scientists. We point out that this gives us control over which events we need to monitor for provenance extraction. When not directly using a WfMS, the overall effect for our purposes is the same - i.e., monitoring data transformation processes to extract provenance information of a result.

In more detail, in Figure 4.1 raw data can be acquired from a variety of data sources such as files (e.g., spreadsheets, images), databases, service providers. The Data Acquisition Process is responsible for the stage (i). Data can be provided using some kind of data acquisition software, which works as a mediator to data sources, or be directly inserted by scientists. In the Execution Process (second stage), activities are invoked to execute specific tasks based on the users' needs. These activities have been previously specified by experts and are stored in the Process Repository. Publish Results (third stage) invokes distinct software modules that publish the results generated.

The Provenance Manager identifies who performed each activity, what processes were executed and when; and inputs/outputs of these processes, storing the corresponding metadata in the Data Provenance Repository. The database management system is used as an intermediate layer between the repositories and the upper layers. The Data Quality Manager encapsulates specific processes to assist in the evaluation of data quality dimensions. Execution flow within the framework is as follows. Raw data are processed (1) and stored in the Data Repository (2). These data are used by processes (3) that are retrieved from the Process Repository (4). At all these steps, the Provenance Manager (3') and (4') extracts information from data and processes, storing such information as metadata in the Data Provenance Repository (6). The results generated are then published (5) by specific processes. Finally, based on requests performed by the specialists the Data Quality Manager is invoked (7), in order to retrieve the information stored in the Data Provenance Repository.

The Data Quality Manager can also look for (8) information from external data sources

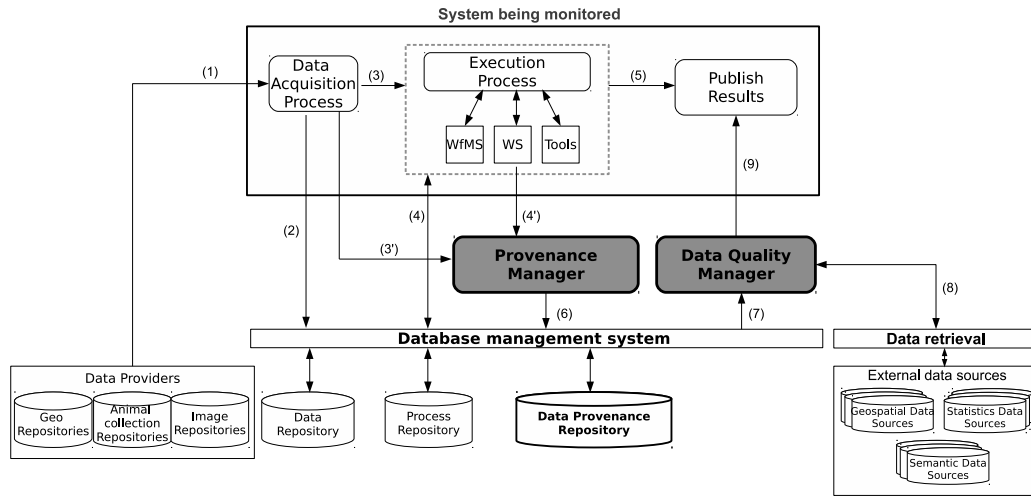


Figure 4.1: The Architecture of the Framework

to complement the information obtained from the provenance repository. Examples of such external data sources are geospatial or statistical data sources from government official web portals. At the end, all these information are delivered to be consumed by the users (9). In the figure, the Data Quality Manager only handles quality of final products, but this can be extended, so that the Manager monitors quality at each processing stage. In this paper, we describe how the information collected by our Provenance Manager and stored in the Data Provenance Repository may assist in the evaluation of quality.

4.3.2 The Data Provenance Repository

This section describes the model to characterize and store provenance metadata when a data product is generated within a scientific experiment. Our ultimate goal is to obtain, using this model, the information that allows evaluation of different data quality dimensions. Our data provenance model is an instantiation of the features provided by OPM [56]. Not only does OPM explicitly define rules and relationships among provenance elements, i.e. artifact, process and agent - it is also being implemented (and specialized) by different scientific workflow systems such as [79] and [82].

Figure 4.2 illustrates the main elements of our provenance model. *Artifact*, *Process* and *Agent* are the main concepts taken from OPM. The Artifact entity corresponds to different kinds of data products that can be produced in a variety of scientific studies. For instance, consider that the main activity is to perform environmental monitoring in a given region. Examples of artifacts can be satellite images covering distinct geographic areas. These images may be used to derive other artifacts, such as maps, to monitor pollution. The Process entity represents activities performed resulting in new artifacts,

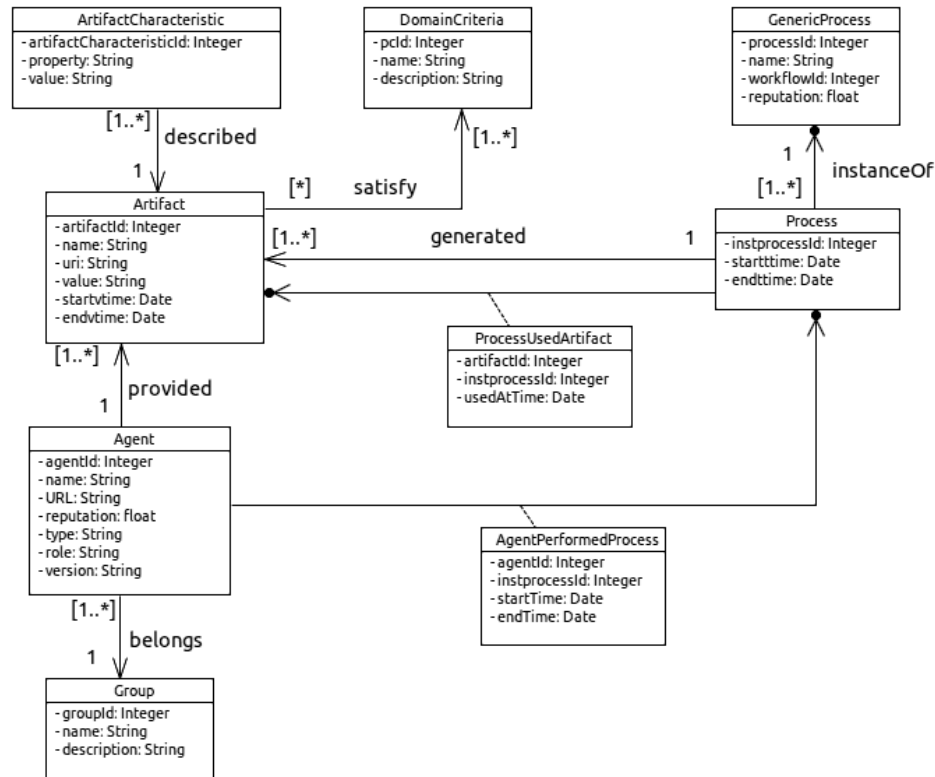


Figure 4.2: Our provenance model

and can be composed by subprocesses. The Agent entity represents entities controlling the processes.

Artifact Characteristic, *GenericProcess*, *Domain Criteria* and *Group* are new elements introduced in our model to support the evaluation of the quality of data in eScience applications. The Artifact Characteristic entity contains the description of each instance of an artifact. Instances of Domain Criteria entity are specified based on the expertise of the users or from rules that define the conditions for an artifact to be accepted. For instance, suppose the artifact in question is a data set containing readings from a temperature sensor device in a given region. Then, an applicable domain criterium might be the acceptable range of temperature values for that sensor. Outliers would indicate sensor failure, or drastic weather change.

The Generic Process entity represents an abstraction of instances of Processes which were executed at a specific time. We adopt this strategy to describe the different executions of a process in an application system. According to OPM specification, artifacts and agents are considered stable elements over time. For this reason, it was not necessary to represent in a generic way the artifact and agent entities. In this case, the

instances of artifacts and agents are represented in the associations *ProcessUsedArtifact* and *AgentPerformedProcess* respectively.

Our model considers that processes are executed by agents to generate artifacts. Agents can be persons or software tools that, in turn, belong to a Group. Since agents have a direct participation on the generation of an artifact, we are interested in their reputation. According to [38] reputation is a kind of collective measure of trustworthiness based on the referrals or ratings from members in a group. In our model, the reputation of human agents is based on the ratings score assigned by a group evaluator (e.g., official institution or scientific society), whereas reputation about non human agents can be obtained directly from reputation systems, or alternatively, also assigned by a human evaluator. These reputation scores are stored as normalized values ranging from 0 to 1.

4.3.3 Basic methodology

In order to evaluate the quality of data produced in a scientific experiment or activity, we basically follow three main steps: (1) choice of the quality dimensions; (2) extraction of the information that is necessary to measure the chosen dimensions; and (3) computation of the scores for each dimension. Then, scientists can assess the quality of their results by contrasting the scores with their predefined criteria.

- Step 1: in order to choose the quality dimensions it is necessary to understand the application domain where data are created. Domain scientists participate actively in this step, not only to determine the dimensions but also to specify in what level of detail these dimensions should be evaluated. Here, we can follow practical guidelines described in [85, 61].
- Step 2: this stage comprises queries to the Data Provenance Repository to obtain the information needed to evaluate the quality dimension(s). When necessary, additional data can be obtained from external sources. For instance, if one wants to know the accuracy of the temperature measurements made by sensors deployed in a region, it may be necessary to analyze the readings made by the sensors of other institutions deployed in the same region and at the same period of time.
- Step 3: in this stage the quality for each predefined dimension is evaluated using metrics chosen by scientists and data from the previous step. Examples of useful metrics and techniques to assess some quality dimensions such as completeness, accuracy, timeliness and believability are described in [67, 73, 68, 17].

However, the use of a metric depends on the kind of artifact under evaluation. For instance, the evaluation of completeness when data are stored in a database considers

Table 4.1: Examples of metrics for the completeness dimension

Type of artifact	Metrics		Reference
Databases	$1 - \left(\frac{\# \text{ of incomplete data elements}}{\text{Total \# of data elements}} \right)$		Pipino et al. (2002)
Information sources	Completeness of source: $C(S)$	$C(S) = c(S), d(S)$	Naumann (2002)
	coverage: $c(S)$	$\frac{\# \text{ of entities represented in the source}}{\text{Total \# of entities in the real world}}$	
	density of attributes: $d(A)$	$d(A) = \frac{\# \text{ of non-null values}}{\text{Total \# of values}}$	
	density of sources: $d(S)$	$d(S) = \text{Avg}(d(A_i))$	
Geospatial data	Spatial coverage	$\text{Overlapping} \left(\frac{\text{extents of the dataset}}{\text{target area}} \right)$	Barbosa and Casanova (2011)
	Temporal coverage	$1 - \left(\frac{\# \text{ of dataset with incomplete period of time}}{\text{Total \# of datasets}} \right)$	Chapman (2005)

the schema completeness, the column completeness and the population completeness [67]. Schema completeness reflects the rate to which relations and their attributes are not missing from the schema. Column completeness represents the degree of the missing values in a column of a table. Population completeness measures missing values related to values in a reference dataset.

The completeness of information sources is assessed based on their density and coverage [60]. In this context, coverage measures the percentage of real world entities represented in the source. Density is defined considering attributes and source. The first measures the ratio of non-null values to all values stored in a data source. The second reflects the average density of all attributes stored in the data source.

In the context of geospatial applications, data completeness may be assessed considering its temporal or spatial coverage, or both [12, 30]. For instance, in order to measure the spatial coverage of a coffee crop map, the region represented in this dataset is overlapped with the target area to compute the ratio of the target area covered by the map [3]. On the other hand, temporal coverage reflects the time scope relevant to a dataset. It can be expressed in terms of an interval of time described using dates. Table 4.1 shows a summary encompassing the metrics for completeness of databases, information sources and geospatial datasets.

4.4 Case study

Our case study involves e-Agriculture, being based on [55]; it concerns analyzing the quality of image files produced to perform crop monitoring. Section 4.4.1 presents the

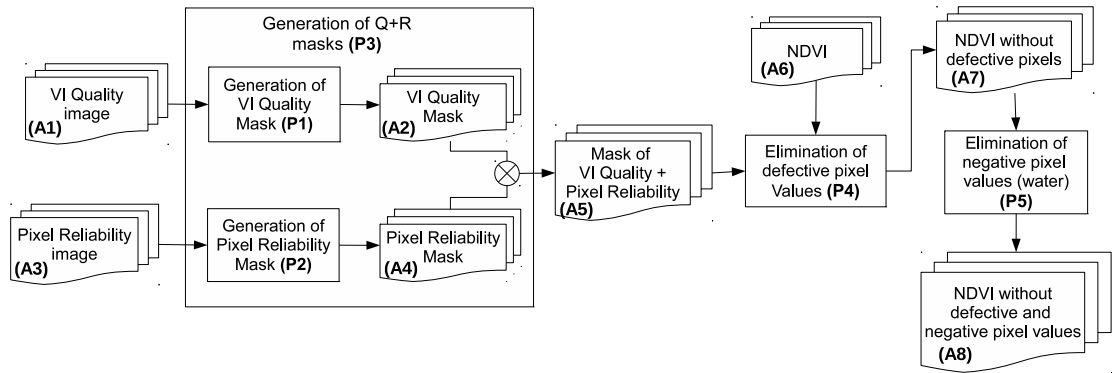


Figure 4.3: Workflow of the activities to identify defective pixels for sugarcane areas, based on [55]

main characteristics of the study. Section 4.4.2 describes the use of our provenance model. Finally, Section 4.4.3 shows how our methodology deals with the evaluation of data quality dimensions related to the case study.

4.4.1 Problem Overview

NDVI (Normalized Difference Vegetation Index) images [23], whose pixels contain NDVI values, are being used as a means to monitor the biomass conditions for vegetation in different domains (e.g. agriculture, environmental planning, biodiversity) [46]. However, to increase the utility of these kinds of images, some issues need to be solved – in particular related to the presence of clouds and noise in the images, which degrades the quality of images and hampers analysis [9, 55].

Taking this into account, Moraes and Rocha [55] propose a method to reduce the presence of clouds or poor quality in NDVI images. This solution was applied to sugarcane areas in São Paulo state (responsible for 60 percent of sugarcane production in Brazil). The method is based on the identification and separation of defective pixels, where sugarcane masks were produced using pre-established criteria. Masking is a technique that is used to select the desired areas of an image based on filtering criteria [32].

Figure 4.3 illustrates the workflow of the main activities executed by [55]. In the figure, input/output artifacts identified with the letter “A” are series of images (e.g., A1 and A4) and the activities that process and produce artifacts are represented by letter “P” (e.g., P1). The entire process was repeated for a series of images produced every 16 days for the period 08/28/2008 through 11/01/2009. NASA [58] was the provider of all input images (A1, A2 and A3 in the figure).

In the first step, sugarcane masks were generated using the VI Quality and Pixel Reliability images and considering specific criteria, such as vegetation index utility (for

Table 4.2: Artifact relation

ArtifactId	Name	URI	startVTime	endVTime	FK_Ins Process	FK_Agent
A1	VI Quality image Period1	wist.echo.nasa.gov/api/	28/08/08	12/09/08		AG3
A2	VI Quality mask Period1	/mask.PR/2008/VIQ.20086523	28/08/08	12/09/08	IP1	
A3	Pixel Reliability image Period1	wist.echo.nasa.gov/api/	28/08/08	12/09/08		AG3
A4	Pixel Reliability mask Period1	/mask.PR/2008/PR.20084523	28/08/08	12/09/08	IP2	
A5	Mask Q+R Period1	/mask_QR/2008/NDV.20081111.tif	28/08/08	12/09/08	IP3	

VI Quality) and values between 0 and 1 (for Pixel Reliability). These criteria are given by the image provider, in this case [58]. Once the masks for VI Quality and Pixel Reliability were constructed, they were applied to the input NDVI images to separate the defective pixels.

The last step consists in eliminating, from each corrected NDVI image, pixels whose values indicate presence of water (i.e. pixel with negative values). The final result is a set of images that have better quality than the original set of input images. Figure 4.3 corresponds to a workflow that takes sets of images (VI Quality, Pixel Reliability and NDVI images) as input and produces a set of NDVI images without defective and negative pixel values as the outcome.

4.4.2 Instantiating our provenance model

Let us now apply our provenance model to this case study, in order to help experts evaluate the quality of the images produced by the output (A8). We start by identifying artifacts, processes and agents. Tables 4.2 through 4.8 show provenance metadata stored during the execution of the workflow of Figure 4.3. These tables contain only a subset of the actual data for illustration purposes.

Table 4.2 concerns Artifacts with part of their attributes. Attribute *ArtifactId* identifies an artifact. The artifact’s name is recorded by the attribute *Name*. References to the artifact’s physical location are identified by an *URI*. For instance, Artifacts A1 and A2 were provided by external agent AG3 (Table 4.4), EOSDIS, while artifacts A2, A4 and A5 were produced by processes within the workflow.

The Artifact relation is complemented by the Artifact Characteristic relation, shown in Table 4.3, that defines the properties or metadata of an artifact instance. Each property is specified as a set of three attributes: *ArtifactCharId*, *Property* and *Value*. For instance, row 4 in Table 4.2 shows that the “Pixel Reliability mask Period1” (Artifact Id A4)

Table 4.3: Artifact Characteristic relation

ArtifactCharId	Property	Value	FK_ArtifactId
AC11	spatial resolution	250 m	A1
AC12	EPSG Projection	4326	A1
AC21	MODIS VI user's guide (MOD13)	VI QA Science Data Sets	A2
AC22	VI Usefulness	0000	A2
AC41	MODIS VI user's guide (MOD13)	MOD13Q1/A1 Pixel Reliability	A4
AC42	Pixel Reliability	0	A4

artifact is valid from 28/08/2008 until 12/09/2008. It is related to process IP2 in Table 4.7 (see further). In addition, Table 4.3 shows that this mask was generated following the user's guide for MOD13Q1 MODIS products, using pixel reliability with value 0, and so on.

Table 4.4: Agent relation

AgentId	Name	URI	Reputation	Type	Role	Version
AG1	Rafael A. Moraes	lattes.cnpq.br /id18	0.8	P	Responsible	
AG2	MODIS Reprojection Tool	lpdaac.usgs.gov /lpdaac/	0.9	S	Software tool	unknown
AG3	EOSDIS	www.echo.nasa.gov/	1	S	Image provider	4.1
AG4	ENVI	main.inforest.gr/	1	S	Software tool	4.5
AG5	ESRI ArcMap	www.esri.com/software /arcgis	0.9	S	Software tool	9.3

Table 4.4 records the descriptions of agents (human AG1 or software AG2). Agents and the processes they executed are stored in Table 4.5. This table shows that, for instance, agents AG1 and AG4 were in charge of the execution of process IP1.

Table 4.5: Agent Performed Process relation

AgPerProId	AgentId	instprocessId	startTime	endTime
AP1	AG1	IP1	10/11/2011 09:30:08	10/11/2011 10:30:10
AP2	AG4	IP1	10/11/2011 09:30:10	10/11/2011 10:30:05
AP3	AG5	IP3	10/11/2011 01:00:00	10/11/2011 01:30:05

Attributes of GenericProcess such as *ProcessId*, *Name* and *WorkflowId* are shown in Table 4.6. Processes are stored as workflows that encapsulate the sequence of activities needed to produce a specific artifact. These processes can be executed several times and produce a new artifact at each execution. Every time a process is executed, a new instance thereof is generated. These process instances are recorded in the entity Process as illustrated in Table 4.7.

Table 4.6: Generic Process relation

ProcessId	Name	WorkflowId	Reputation
P1	Generation of VI Quality mask	W1	0.9
P2	Generation of Pixel Reliability mask	W2	0.9
P3	Generation of Q+R mask	W3	0.9
P4	Elimination of defective pixels values	W4	0.9
P5	Elimination of negative pixels values (water)	W5	1.0

Table 4.7: Process relation

InstProcessId	startTime	endTime	FK_ProcessId
IP1	10/11/2011 09:30:12	10/11/2011 10:30:00	P1
IP2	10/11/2011 11:00:00	10/11/2011 12:00:00	P2
IP3	10/11/2011 01:00:00	10/11/2011 01:30:00	P3

For example, process P3 (that generates the Q+R mask) was executed only once, to generate the Mask Q+R Period1 artifact (A5). Table 4.8 represents the input parameters that are required to produce a new artifact. It shows, for example, that artifacts A2 and A4 were used as input to process IP3 (used to generate the Mask Q+R Period1 image). In the example, Table 4.7 and 4.8 show information generated for a single run of the workflow of Figure 4.3.

Table 4.8: Process Used Artifact relation

UsedId	ArtifactID	InstProcessId
U1	A1	IP1
U2	A3	IP2
U3	A2	IP3
U5	A4	IP3

Domain criteria that an artifact must meet are recorded in Table 4.9. According to our domain experts, the presence of defective pixels greater than 5% in a set of 28 images (corresponding to the sugarcane cycle for the 2008 - 2009 cropping season) reduces the quality of the maps, thus limiting their use in agricultural monitoring activities. This is expressed by the domain criterion DC1.

4.4.3 Applying the methodology

Let us now consider the workflow of Figure 4.3 whose final artifact is a set of images A8. Consider the specific execution of this workflow, related by [55]. Imagine that some agricultural scientist wants to use A8 in her research, and wants to evaluate its quality. The scientist starts by defining the quality dimensions (s)he prioritizes (step 1 of the

Table 4.9: Domain Criteria relation

PCId	Name	Description
DC1	Defective pixel	dp <= 0.5
DC2	Kappa index	0.6 <= k <= 1
DC3	Spatial coverage	0.8 <= sc <= 1

methodology). In this case, the dimensions include: (1) spatial coverage of the images; (2) reputation of the image providers; (3) number of correct pixels (and thus indirectly accuracy); (4) comparison with a known reliable set of images (say, from previous experiments).

Next there are two issues to be considered (step 2): (i) extraction of the information chain associated to the production of A8 and (ii) searching for other data sets concerning the same region and period of A8, and of the same nature (i.e., set of images) to analyze how well they agree with A8. The information associated with the production of A8 is extracted from the Data Provenance repository using the query module of the Data Quality Manager.

The corresponding query, represented in SQL, appears below. The code retrieves processing steps that produced A8. This can be composed with other queries that will backtrack through the workflow to extract the entire production chain.

```

1 select artf.name, artf.instprocessId, pro.processId, pro.name
2 from artifact artf, generic_process pro, process iop
3 WHERE artf.artifactId = A8 and artf.instprocessId = iop.instprocessId
4 and iop.processId = pro.processId

```

The result of this query is partially presented in Table 4.10. It shows that the name of artifact A8 is “NDVI without defective and negative pixel values” and that it was produced by the instance “IP5” of process “P5” (“Elimination of negative pixel values”). Step 3 concerns using all metadata collected in the previous step to evaluate the predefined

Table 4.10: Example of result of the query

Artifact Name	Process Id	Process name	Process instance Id
NDVI without defective and negative pixel values	P5	Elimination of negative negative pixel values	IP5

dimension. The number of correct pixel, in this specific case 99.3%, is provided as part of A8 metadata (recorded within Table 4.3). In this case, this information is available because it comes the data provider. The reputation of P5 (in this case 1.0) is obtained from Table 4.6. Other reputation values of intermediate processing steps are obtained the same way. The scientist can then evaluate the overall reputation of A8 using whichever means

(s)he decides, which may also include the reputation of the provider of the original data sources (in this case, EOSDIS). In order to assess the spatial coverage of A8, its images are matched against a user-provided bounding box. Finally, the scientist will check the values obtained from these three evaluations (pixel, reputation, spatial coverage) with those of the previous experiments.

The overall result of these evaluations in step 3 provides quantitative quality information. We also point out that provenance can also be used to investigate elements in the production process that may be responsible for the degradation of a final product. For instance, suppose that A8.1, A8.2...A8.n are n sets of images, each of which produced by a different execution of the workflow of Figure 4.3. Then if the reputation of A8.1...A8.n is consistently below an acceptable level, the workflow has to be reengineered to produce more reliable data.

4.5 Conclusion

This research described and discussed a semi-automatic approach to assist scientists in assessing the quality of the data produced in an experiment. Our approach is composed by a provenance model that takes advantage of features provided by OPM [56] and a methodology that outlines the basis for evaluating data quality. Our solution is applied to a case study for agriculture. Though our example is for this domain, our extension to OPM covers a broad range of e-Science application domains.

First, it allows generic domain criteria to be recorded, for each Artifact handled in an experiment. Second, it supports storing domain-specific Artifact properties in the ArtifactCharacteristic table; this is structured in a RDF-like organization, which allows recording an arbitrary number of provenance-relevant data items. Also, by associating Agent with Group, it supports identification of the research group associated with the execution of a process. Third, it enables storing the properties of the different executions of a Process over time, in order to generated artifacts. All of these characteristics maintain OPM's generality, while at the same time supporting the storage of provenance details on Artifact, Process and Agent.

In a previous survey we pointed out that some interesting quality attributes to eScience domains such as biology and agriculture are completeness, accuracy and timeliness. We believe that information managed by our framework related to the processing steps that generate data can help scientists in the evaluation of the quality of their findings.

Extensions related to this work include expanding our framework to other application scenarios. Our data provenance model can be extended to encompass additional metadata characteristics considering not only the domain but also the different semantics of specific quality dimensions (e.g., completeness). In order to validate and enhance the interoper-

ability of our model, we are studying other approaches such as the use of ontology models that allow to represent and interchange provenance information in different contexts.

Chapter 5

Estimating the quality of data using provenance: a case study in eScience

5.1 Introduction

Challenges related to the quality of data are common to applications in a variety of domains. Not only can it directly affect decision processes in an organization, but also in a scientific context (e.g., healthcare, environmental sciences, astronomy, etc). With the data deluge generated by groups and organizations around the world, there is a growing demand for new computing solutions to help decision-makers to select the best data that match their needs. The same can be extended to a scientific environment: before scientists can take actions to analyze their findings, they need to know the quality of the data sets they are working on.

Problems to be faced include, for instance, data incompleteness, inconsistency, lack of standardization of formats, inaccurate data, among others. Besides that, data of different nature and the variety of information systems hamper the obtention of good quality data [6]. As will be seen, though our case study is in a specific domain (biodiversity), our proposal is generic enough to be applied and extended to any (computational/organizational) environment that requires cooperative work, and that must rely on integration of heterogeneous data sources. The underlying hypothesis is that there are a set of common characteristics in all such environments - such as the need for collaboration among actors with distinct needs and views of the issue at hand, a wide variety of heterogeneous data sources, and the need to coordinate complex data-driven processes.

Depending on the application domain, each of these problems demands different strategies to solve data quality issues. For instance, in the context of database systems, incompleteness of data might be tackled considering the granularity of its elements, i.e., completeness of value, tuple, attribute and relation [6]. In the context of Web data, the

same problem might be characterized by evolution in time - i.e., the speed at which the data will be completed [65].

Related work has shown data quality to be a problem that has to be attacked under a multidimensional view [85, 7]. Quality dimensions can be considered as attributes that allow to represent a particular characteristic of quality [85]. In particular, accuracy, completeness, timeliness and consistency have been extensively cited in the literature as some of the most important quality dimensions to information consumers [12, 63, 6]. These general dimensions can be considered common to both business and scientific domains.

The tracking of historical information concerning the creation of a dataset is also known as *data provenance* [56]. Provenance is seen as a kind of metadata that gives information about the what, when, where, how, by whom, and why a dataset, object or artifact was created [72]. Taking these characteristics into account, we explore provenance as a strategy to provide information to evaluate the quality of data.

In some domains and applications, provenance information can involve a complex and scalable relationship network between different resources and processes [14, 33, 5]. In this work we take advantage of the RDF/OWL model flexibility [42, 21] and scalability [86] as a means to represent provenance information and its internal relationships, focusing on the biodiversity domain.

Unlike solutions centered on workflow systems such as [82, 39, 79], which aim to provide native support for provenance to reproduce the planning and running of data processing and management steps, our approach can be adopted in different systems to collect domain-specific provenance and use this information to evaluate quality. Although this kind of approach is also investigated in [72] to allow knowledge discovery, we believe that different considerations need to be taken into account when it is used to analyze how good are the data produced by automated processes.

Our solution also addresses two requirements identified by the international provenance challenge¹ proposed in the context of the Open Provenance Model [56]. First, we show the applicability of provenance in the quality context by using it as a key parameter to help determine the quality of data in scientific organizational environment. Second, by making use of ontologies to represent provenance, we allow interoperability among groups, enabling them to share and compare the information produced in their work.

The main contributions of this research include: (i) supporting the assessment of quality of scientific data based on its provenance and (ii) the adoption of a semantic model (PROV-O) to represent provenance. The latter extends our earlier work - in which we use a relational model to store provenance. Here, rather than a relational model, we extend the PROV-O semantic model to a new ontology, to consider domain-specific characteristics. We validate our approach through a case study concerning metadata

¹<http://twiki.ipaw.info/bin/view/Challenge/WebHome>

generated in an information-intensive biodiversity experiment.

5.2 Background of the solution

In our previous work [51] we presented a conceptual framework to support keeping track of data provenance, in a relational model, to assess data quality. Our framework embeds a Provenance Manager service, a provenance database model, a Data Quality Manager service and a methodology to support the evaluation of the quality of data. Our focus in that work is the development of the data provenance repository and the application of the methodology in the estimation of the quality of data and reports produced in agricultural planning.

In that framework, the database that stores provenance information was designed using the Open Provenance Model (OPM) specification [56]. It represents data lineage in terms of **agents** that control **processes** to modify/produce **artifacts**. These elements are associated through five causal relationships within a provenance graph (e.g., an artifact **was generated** by a process). OPM only allows to represent artifacts as *immutable pieces* of state. This means that the state of an artifact cannot be modified after its creation.

Our methodology to support the evaluation of the quality of data in computational processes encompasses three main steps: (i) selection of the quality dimension(s) of interest; (ii) extraction of the information that is necessary to estimate the quality of the target dimensions; and (iii) computation of the score for each dimension. Users might use metrics to estimate the quality score or directly assign the scores based on the provenance information requested. We pointed out that each one of these stages is directly associated with the application domain under study and the activity that will be performed. Users choose the quality dimensions of their interest based on the kind of artifact under study (e.g., a spreadsheet file, a picture, a data statistics graph or a database table).

In this paper we adapt the methodology so that stage 2 should also consider the capture of metadata that will compose the provenance information. The retrieval of this provenance information is part of the activity that can be used to estimate the quality score. Given the fact that OPM does not allow object evolution, and considering that data evolution is a natural state of the world, we decided to change our provenance model. We extended PROV-O, a provenance semantic model, to represent domain-specific provenance. We adopted this approach because of the flexibility that ontologies provide. Provenance metadata are captured during the execution of operations on data. This can be achieved, as shown by previous work, by progressively storing execution traces, as well as information on data state changes – e.g., see [40]

Earlier studies have investigated the support of provenance management based on domain ontologies [72, 90]. The novelty in our work is to support data consumers on the

estimation of quality. We can take advantage of several characteristics using an ontology-driven approach to represent and store provenance information. First, semantic modeling improves both interoperability and scalability of systems, since the schema and data can be more easily aligned with other schemas or instances. Second, adopting strategies like linked data [89], each item of the schema and each data instance may have unique identifiers, thus enabling alignment with data from other sources that have also been modeled using semantics. This enables interoperability - not only within an organization, but across organizations, or groups.

Our work has been developed in the context of management of data and activities performed by scientists in the animal sound collection of the State University of Campinas, UNICAMP, Brazil (from now on called FNJV²). As will be seen, from a high level point of view, these activities are comparable to those executed by people in any information-sensitive organization, to collect, clean and publish their data sets.

5.3 ProvenBiO: A PROV-O-based ontology for provenance information for the biodiversity domain

Any information-rich environment involves a complex and scalable relationship network between different and distributed resources, processes and users. Distinct organizational scenarios adopt distinct tools, vocabularies and methodologies. In order to represent provenance information and its relationships, we take advantage of the expressiveness that RDF/OWL provide, focusing on the specificities of the biodiversity domain.

PROV-O [84] is an ontology based on OWL2 that specifies a data model to express provenance records in different application scenarios. PROV-O is a candidate recommendation in development by the W3C Provenance Working Group. It defines a set of starting point terms which are three core classes: **entity**, **agent** and **activity**. These classes are associated by nine relations such as *wasAttributedTo* and *wasInformedBy*. PROV-O provides additional subclasses and sub-properties that can be used to complement the initial terms and also to add more details among the relations – and thus specialize it to distinct usage domains.

Basically, the datasets that are submitted to a transformation process are instances of the **entity** class and the processes that modify and use the datasets are instances of the **activity** class. The entity responsible for commanding the execution of an activity is modeled as an **agent** class. Agents can also command other agents.

We implemented an instance of PROV-O that we call ProvenBiO – A Provenance Biodiversity Ontology, available at <http://purl.org/provenbio/ontology#>. The goal of

²Fonoteca Neotropical Jacques Viellard, Institute of Biology, UNICAMP

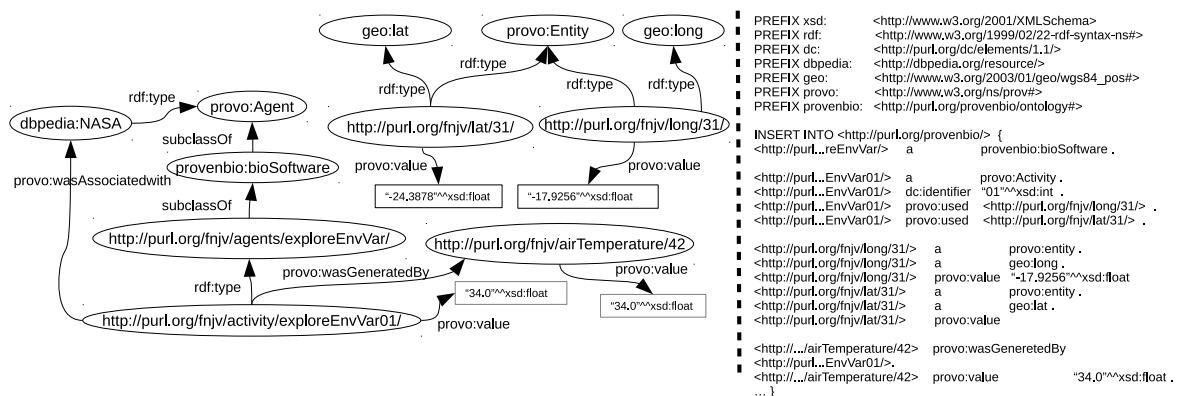


Figure 5.1: Example of a portion of our ProvenBiO ontology and the corresponding SPARQL query

ProvenBiO is to preserve provenance information related to applications in the biodiversity domain and use this information to support the assessment of quality of data used and/or generated by domain experts. ProvenBiO adopts widely used vocabularies and ontologies (e.g., Dublin core [22], Geospecies [24], Darwin Core [26]) aiming at enriching the provenance metadata with terms interesting to the biodiversity context. Figure 5.1 illustrates a portion of a set of procedures and elements modeled in a ProvenBiO graph together with their corresponding RDF triples.

The figure shows, for example, the properties that we adopted from PROV-O, describing the interaction among them (e.g., entity `http://purl.org/fnjev/airtemperature/42` `prov:wasGeneratedBy` the activity `http://purl.org/fnjev/activity/exploreEnvVar01/`). To better distinguish an activity that represents a concept (e.g., `provenbio:bioSoftware`) from an activity that was performed within a system (e.g., `http://purl.org/fnjev/activity/-exploreEnvVariable01/`), it was necessary to add a new class. Thus, we specialize the class `agent` of PROV-O with a new class called `bioSoftware`. The figure also shows some terms such as `geo:lat` from GeoNames and `dc:identifier` from Dublin Core. In other words, PROV-O can be specialized and modified to meet distinct domain requirements.

5.4 Case study: using ProvenBiO to derive data quality

5.4.1 Motivating Scenario

The volume and variety of data types, their storage using different formats and through distributed repositories are common problems that hamper the assessment of the quality

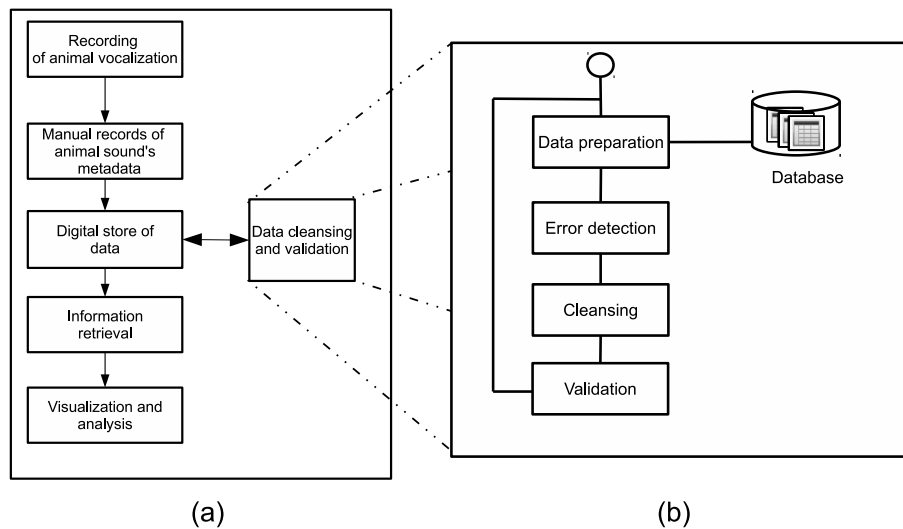


Figure 5.2: Basic flow concerning processing animal sound recordings – FNJV, inspired on [19]

of data in the domain of biological diversity [12]. To illustrate a scenario, we briefly describe some challenges concerning the management of sound recordings faced by FNJV [19].

FNJV maintains the largest collection of animal vocalization recordings in the Neotropics. In order to preserve these recordings, researchers have created a digital repository for them. Metadata are essential to manage recordings, and thus the quality of information provided by metadata has become a crucial issue. Problems found related to such metadata include, for instance, variety of formats, missing data values, abbreviations, misspellings, missing or wrong information about species location. Common data quality problems are related to completeness, accuracy and consistency of data. Our case study investigates the quality of such metadata, in particular after they have been curated and gone through several cleaning processes (and thus, how good are the processes that were run to improve metadata quality).

Figure 5.2 (a) depicts the basic process concerning the management of the recordings. First, biologists record animal vocalizations using distinct devices. Next, they write metadata in their notebooks (e.g., geographic location, scientific name, weather conditions) concerning the sound recorded and recording environment. Subsequently, all the metadata is stored in a database³. A data cleaning process follows this step. Finally, in order to perform scientific analyses, biologists query the database.

Cugler et al. [19] faced a subset of these problems by proposing an approach to fill missing metadata fields and derive such information automatically, from external Web

³using a system developed by [19]

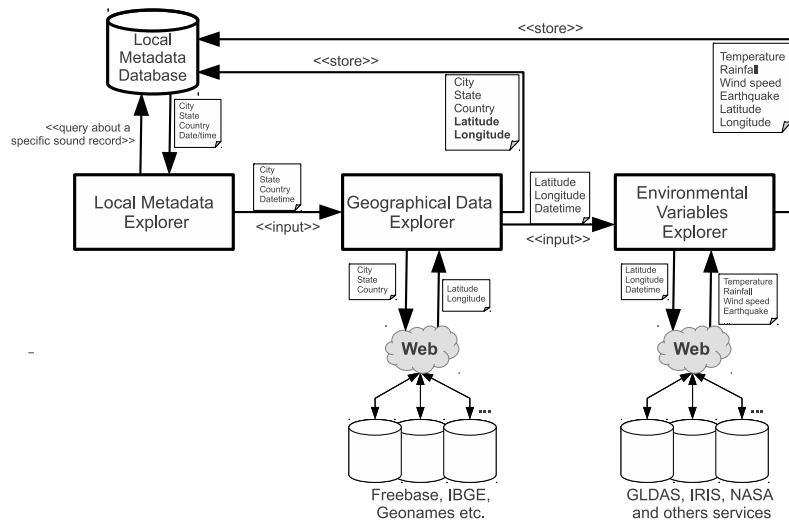


Figure 5.3: Workflow of the data cleansing activity, based on [19]

sources. However, no evaluation about the quality of the original metadata and the derived datasets was performed. Taking this into account, we focus on the evaluation of data quality when the process to clean and fill missing metadata values is executed. Figure 5.2 (b) shows the general steps of the data cleansing process.

Figure 5.3 describes the workflow that is used by [19] to fill missing metadata values. Notice that this is a generic workflow that can be specialized for domain-specific cleansing activities. In the case study, processing starts from the geographic region (usually a location) where the sounds were recorded, from which missing environmental information can be obtained. The location name metadata is used to query the Freebase knowledge base, in order to derive the latitude and longitude of the location informed⁴. Next, the latitude and longitude obtained are combined with stored metadata values “collect time” and “collect date” to be used as input to web services such as NASA’s GLDAS and IRIS. These and other services are used to derive metadata on environmental variables at the time and location of the recording. We inserted probes in this workflow to capture provenance information at each stage of the workflow execution. This information is represented as instances of ProvenBiO.

5.4.2 ProvenBiO ontology: a running example

In this scenario, provenance plays an important role since biologists need to know how the fields were completed, and track the cleaning processes, users and resources in order to consume the data in their investigations. Typical questions that experts may ask are:

⁴Most records date back to the 70’s, a pre-GPS era. If lat/long is available, this step is bypassed

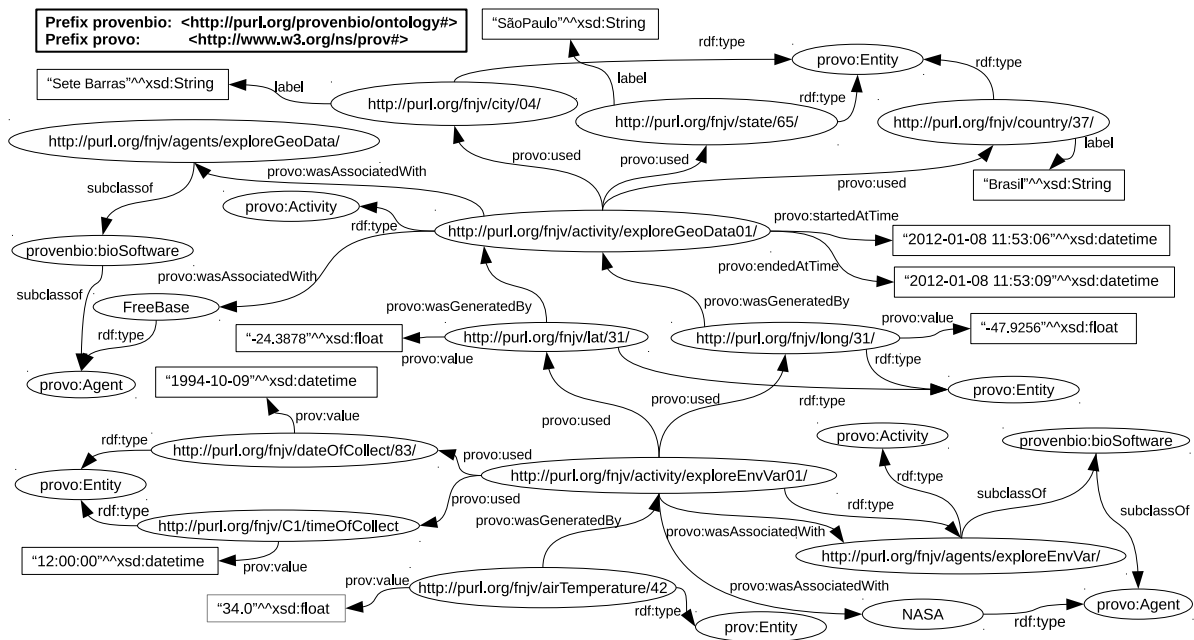


Figure 5.4: Example of RDF triples of ProvenBiO

“were these metadata fields filled by an expert or a novice user?”; “can I rely on the data collected from this specific source?”; “are the derived metadata complete enough?”

Figure 5.4 illustrates a portion of ProvenBiO RDF triples. This ontology is also a result of our previous experience in modeling provenance [51]. In the figure, triples correspond to the provenance information collected in one execution of the prototype shown in 5.3. In the figure, resources and values are nodes and properties are edges. The figure shows, for instance, that there exist OWL classes that represent *Activities* and *bioSoftware* agents. The Activity `http://purl.org/fnjl/activity/exploreGeoData01/` has properties such as `startedAtTime`, `endedAtTime` and `wasAssociatedWith`, which hold the interval when the instance was executed and its associated agents like `FreeBase`. Furthermore, we also have the data produced by this activity – uniquely identified as `http://purl.org/fnjl/lat/31/` and `http://purl.org/fnjl/long/31/` with their respective values.

5.4.3 Capturing Provenance Information

The Provenance Manager is composed by a set of services that we implemented to allow to capture provenance information. Figure 5.5 depicts the elements - the Data Provenance Collector and the RDF Serializer services – that compose the Provenance Manager.

The *Data Provenance Collector service* is in charge of capturing the provenance meta-

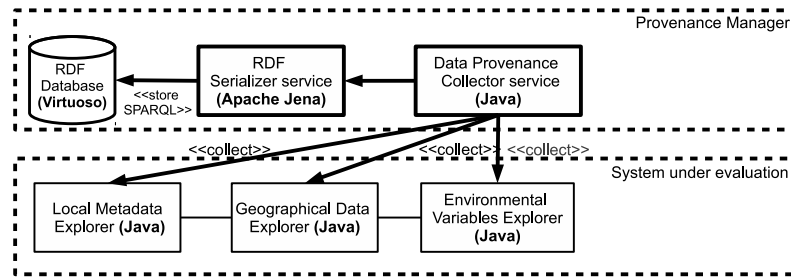


Figure 5.5: Elements that compose the Provenance Manager

data. The goal of this service is the extraction and classification of the metadata that keep track of the activities and entities that participate in the generation of missing values. The figure shows that the Data Provenance Collector service can be plugged in at each processing stage of the data cleaning system being monitored. Information such as people and tools that participated in the generation of a piece of missing data, details of processes, and parameters used by the processes, are collected when the system is executed.

Next, the information collected and classified by the Provenance Collector service is delivered to the *RDF Serializer service*. It takes the provenance information and submits it to a categorization process, where this information is mapped with a corresponding ontology term. We use the set of terms and properties defined in ProvenBiO to represent provenance at this stage. Once all information is instantiated, it is stored into the Provenance Repository in the format of RDF triples. We implemented these services using Java technology. Furthermore, we use the Apache Jena framework to build and write RDF triples which are stored into the Virtuoso database (via SPARQL queries).

5.4.4 Querying Data Provenance to Derive Quality

Let us now regard the workflow of Figure 5.2(b) whose general goal is to perform data cleaning and fill in missing metadata values (using the workflow of Figure 5.3). Consider that an expert wants to know the quality of the datasets that resulted from the data cleansing process, so that (s)he can subsequently use such datasets. Using the strategy described in this work, users can pose queries against the RDF Database (our Provenance Repository), in order to retrieve information to estimate quality. For instance, imagine that the specialists are interested in evaluating the completeness and confidence of the datasets. Examples of queries that we can answer are:

1. *Search for the metadata records that were completed using data sources whose average reputation is higher than X;*

2. *Retrieve metadata records for which the cleansing activity is over, and which started before a date D;*
3. *Find newly completed metadata records for species found in tropical countries;*
4. *Retrieve newly completed metadata records that are related to endangered species;*
5. *Retrieve the identifiers of all databases whose reputation is higher than 0.6 and which were used in the workflow of Figure 5.2 to fill missing metadata for Passeriformes species recordings.*

Queries 1 and 2 are simple to solve in a relational database and are supported by other provenance. However queries 3, 4 and 5 are more complex and may involve further information and relationships that we can only solve using ProvenBiO ontology. Also notice that some queries are specific to the domain of our case study, while others can be considered in the context of generic information handling environments. Figure 5.6 shows the SPARQL query for item 5. The first lines shows that consensual vocabularies and ontologies like Geospecies and Dbpedia-owl were adopted. The second part concerns the query itself.

Once the information is delivered to the specialist, (s)he can apply specific rules to decide whether the data are good enough. We developed a prototype to query the data provenance captured by our Provenance Manager, available at <http://purl.org/provenbio/-?task=do/querynav>. Related work, as discussed in the next section, considers only stored (meta) data. Our approach, on the other hand, allows finding additional information, which is obtained from relationships among stored data and ontologies. Thus, our provenance-based queries can return much more than information restricted to the stored provenance metadata. In other words, the results of these queries are data that can be analyzed by users to evaluate quality according to their criteria. Figure 5.7 presents a screen copy of our query prototype. It shows some basic queries that specialists can perform in order to evaluate the quality of their data.

5.5 Related Work

Data quality is seen as a subjective concept. Frequently data considered good enough for a group of users can be considered bad for others [76, 12]. Thus, the assessment of the quality of data needs to consider the characteristics of a specific context (e.g., e-Business, healthcare, environmental sciences). There are many research initiatives that tackle the assessment of quality by presenting methodologies to measure different data quality dimensions – e.g., [67, 2]. However relatively little work explores and applies the

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geospecies: <http://rdf.geospecies.org/ont/geospecies#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX provo: <http://www.w3.org/ns/prov#>
PREFIX provenbio: <http://purl.org/provenbio/ontology#>

select * from <http://purl.org/provenbio/> {

?databaseAgent a provo:Agent.
?databaseAgent a foaf:Organization.
?databaseAgent a provenbio:publicDataOrganization.
?databaseAgent provenbio:trustScore ?trustScore

?swagent a provenbio:swagent.
?swagent provenbio:isAssociatedWith <http://purl.org/fnjl/agents/nasa>

?activityInstance provenbio:bioSoftware ?swagent
?activityInstance a provenbio:ActivityInstance
?activityInstance provo:startedAtTime ?startDateTime
?activityInstance provo:wasAssociatedWith ?databaseAgent.
?activityInstance provo:used ?instanceInputSpecies.

?instanceInputSpecies a geospecies:SpeciesConcept
?instanceInputSpecies geospecies:inOrder <http://lod.geospecies.org/orders/hNvZJ>
?instanceInputSpecies geospecies:hasLocation ?speciesCountry
?speciesCountry dbpedia-owl:country dbpedia:Brazil

FILTER (?trustScore > 0.6) }

```

Figure 5.6: SPARQL query corresponding to Item 5

information produced when a dataset is generated - i.e. its provenance - as a key piece to evaluate the quality of data.

Simmhan and Plale [76] describe an approach for personalized quality scoring to rank scientific datasets based on a quality profile. Provenance metadata is used to model a quality function based on weights setting on a user's quality profile. Machine learning techniques are used to construct a quality function to produce a quality score. The main idea behind this solution is to predefine quality scores of the input data to map to the quality score for the derived output data. Although our solution can use the expertise of specialists to annotate quality scores of input data (e.g., confidence of a data source), we believe that this kind of approach can be time consuming - the broader the application domain is, the greater the effort to configure a quality profile. Rather than relying on (manual) user-assigned scores, our approach tries to automatically get as much information as possible that is produced when a dataset is generated to be used by the specialists in the quality assessment process. Moreover, we do not compute quality scores. Rather, it is up to the user to derive information (s)he considers useful to obtain

ProvenBiO Project
Capturing and querying provenance data from applications in the biodiversity domain

Home
About
Contact
Examples
Data
Playground
SPARQL
Endpoint
License

Example Queries:

Q1: Search for the metadata records that were completed from sources whose average reputation is higher than "0.6".
 Q2: Which are the metadata records that were completed for species found in tropical countries ?
 Q3: Which were the sources used to derive latitudes related to species that belong to Passeriformes ?

```

PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geospecies: <http://rdf.geospecies.org/ont/geospecies#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX provenbio: <http://purl.org/provenbio/ontology#>

select distinct(?metadataRecord), ?associatedSpecies from <http://purl.org/provenbio/> {

  ?metadataRecord a prov:entity.
  ?metadataRecord prov:wasGeneratedBy ?instancesOfProcess .
  ?instancesOfProcess prov:wasAssociatedWith ?bases .
  ?instancesOfProcess prov:relatedSpecies ?associatedSpecies.
  ?associatedSpecies provenbio:inFoundInTropical provenbio:true.
}

```

[Run Query](#)

Figure 5.7: Screen copy of our query prototype. The code for Q2 is partially shown in the window.

quality information. Though this is an ad hoc process, on the other hand users are free to investigate any quality criterion of their choice.

In order to compute a quality score that can to be used in the evaluation of the quality of data on the Web, [35] describe a solution to annotate provenance metadata (e.g., date of creation) with impact values. The provenance model constructed is directly associated with the timeliness quality dimension. Unlike this work, we do not need to specialize provenance for each dimension of quality. In our case, the specialist can choose the quality dimensions of interest and request for information that can help to assess the dimensions.

Similar to [35], [68] also propose an approach to compute the believability quality dimension based on the provenance of a data value. The computation of believability has been structured into three complex building blocks: metrics for assessing the believability of data sources, metrics for assessing the believability from process execution and global assessment of data believability. Although this is a precise approach to measure believability, the authors only measure the believability of a numeric data value, which limits its applicability.

Notice that one singular characteristic of our work is that we generated an instance of a generic ontology for provenance representation. This ontology allows to collect information related to a specific domain and store data provenance that is used to assess quality in a specific context.

5.6 Conclusion and ongoing work

This paper presented an approach to support specialists in the estimation of quality of datasets based on provenance information, for data-intensive applications. Rather than concentrating our study on standard organizational environments, we analyze environments in which scientific experiments are planned, specified and executed, insofar as they reflect a particular set of procedures and processes to run experiments. In order to provide domain-specific provenance, we generated an ontology instance (ProvenBiO) based on the W3C PROV-O ontology and data model. Besides typical queries focused on provenance from a system point of view (e.g., processes), this solution enables specialists to investigate relationships among elements within a specific domain. Aiming at the expressiveness of ProvenBiO, we aggregated widely adopted vocabularies such as DwC and Geospecies. This enhances interoperability across distinct groups that want to share and reuse data sets in their processes.

In particular, we use the provenance information to allow experts to perform queries aimed at assessing the quality of data. Distinct members/roles in a given group or organization can be interested in different dimensions of quality, depending on the kind of activity that they are performing. For this reason, the automation of the measurement of quality can be a difficult task, especially if we consider that each dimension of quality may cover other sub-dimensions.

Our solution was validated using a case study concerning recordings of animal sound vocalizations. We implemented a set of services that enable to capture and identify provenance metadata when a system is being executed, and a service that allows to query this information. Future work that we want to investigate is related to the propagation of data provenance among the transformation processes through which a dataset is submitted. Another extension is related to the analysis of provenance as a criterion to adapt a workflow to a specific organizational context. Our queries require knowledge of SPARQL. Future work also involves developing an interface with translation mechanisms to transform user requests into SPARQL queries.

We point out that though our work is concerned with scientific processes and data, it is generic enough to be applicable to other organizational contexts. It suffices to adapt the ontology to contemplate concepts and relations of the domain of interest.

Chapter 6

Conclusions

6.1 Contributions

This PhD research addressed some Computer Science issues involving scientific data management - more specifically data provenance and quality. The focus of this work was to investigate how the provenance information generated by scientific activities can help the specialists in the assessment of the quality of the datasets used or produced by their research. We use two scientific scenarios - in agriculture and biodiversity - to perform requirements elicitation and to validate our approach. Our three main goals were:

- G1.** Definition of the data quality dimensions more interesting to the scientific domain;
- G2.** Management of data provenance aiming at the assessment of quality in a specific domain and
- G3.** Enrichment of data provenance to provide a greater amount of information to help scientists in the assessment of quality.

Motivated by the different views of quality, we conducted a literature study to understand what quality means in various scientific domains. In the context of the first goal (G1), we presented a list of data quality dimensions and sub-dimensions useful to environmental science, and concluded that some data quality dimensions, such as accuracy, timeliness and accessibility are frequently studied in the scientific domain. However the evaluation of these dimension must be performed considering the scientific activities, needs of each research groups and intended use of data.

Aiming at achieving the second goal (G2), we proposed an architecture that allows to record data provenance and use this information to derive quality information. The framework relies on a provenance model and a methodology to be used as a basis for evaluating data quality. Our study on provenance targeted on OPM, which we specialized

to develop a database that allows to store provenance information. This model was used in a case study on agriculture – related to images used to monitor biomass for a given crop. The case study illustrates how our methodology is applied and how we can use the information stored to answer queries about quality.

In order to handle interoperability issues related to the third goal (G3), we searched for solutions that help scientists better understand and assess the quality of their datasets, by taking into consideration a wide range of information sources. We were inspired here by work that explored domain-specific provenance. As a result, we proposed a new ontology instance, called ProvenBiO, which is based on PROV-O. ProvenBiO can be used to answer questions related to quality by correlating data provenance generated by applications (e.g., the system described in [19]) and elements that belong to a specific domain. We validated this proposal within a scenario concerning metadata generated in an information-intensive biodiversity experiment, where system and domain-specific provenance are represented as instances of ProvenBiO. Once these instances are generated, we show how to derive information related to the quality of a dataset using queries on ProvenBiO.

Summing up the main contributions of this thesis were:

- Investigation of the characteristics that highlight data quality issues in the context of eScience.
- Identification and specialization of generic provenance models considering data quality issues.
- Specification of a framework that combines a provenance model to keep track of data provenance with a methodology that addresses the utilization of provenance to assess the quality of the datasets.
- Adoption of a semantic model and well known standards to facilitate the retrieval of information that can be useful in the evaluation of quality of the data sets.

6.2 Extensions

There are many possible extensions to this work encompassing theoretical and practical proposals. Examples of some of these extensions are:

- Specification of a set of quality rules to define which are the conditions necessary for a given dataset to be considered with a good level of quality by the specialists. The investigation of a formal language to describe the rules is another possible extension.
- Investigation of approaches that use semantics to enrich provenance information, and highlight new features to be incorporated by our framework. Furthermore,

explore if system and domain-specific provenance information is enough to answer users' quality concerns or whether it is necessary to incorporate other elements to obtain better results. For example, what were the processes that degraded the precision of a map.

- Study of issues concerning quality of legacy data. Challenges such as the original quality of data and the quality after data are exported to a new database are some problems that need to be faced.
- Propagation of the capture of provenance through the different levels of the transformation processes applied to the data (e.g., as in [40]), and to use this in assessing quality. This kind of strategy can be useful to understand how the quality of a specific dataset changes at each processing step.
- Use of actual workflow engines to capture provenance automatically, and combine this information with ontology based models to conduct experimental studies in order to derive quality.
- Design and implementation of a query module that can be used to customize users' queries. This module should be able to encompass a wide range of terms and predicates, in order to retrieve information relevant to data quality assessment.
- Use data provenance as a criterium to adapt the execution of scientific workflows.
- In the model of Chapter 4, only one data output is considered. This can be easily generalized to multiple outputs, being left for future work. In the same model, artifacts can also be linked to transaction time (and not only validation time). Still other extensions to this model are supporting artifact derivation and composition and also to consider aspects of quality of stream data, given their variability over time.

Bibliography

- [1] S. Babu and J. Widom. Continuous queries over data streams. *SIGMOD Rec.*, 30(3):109–120, 2001.
- [2] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi. Modeling Information Manufacturing Systems to Determine Information Product Quality. *Manage. Sci.*, 44:462–484, 1998.
- [3] I. Barbosa and M. A. Casanova. Trust Indicator for Decisions Based on Geospatial Data. In *Proc. XII Brazilian Symposium on GeoInformatics*, pages 49–60, 2011.
- [4] R. S. Barga and L. A. Digiampietri. Automatic Generation of Workflow Provenance. In Luc Moreau and Ian T. Foster, editors, *Proc. of the 2006 Int. Conf. on Provenance and Annotation of Data*, volume 4145, pages 1–9. Springer-Verlag, 2006.
- [5] R. S. Barga and L. A. Digiampietri. Automatic capture and efficient storage of e-Science experiment provenance. *Concurr. Comput. : Pract. Exper.*, 20(5):419–429, 2008.
- [6] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer-Verlag, 2006.
- [7] R. Blake and P. Mangiameli. The Effects and Interactions of Data Quality and Problem Complexity on Classification. *J. Data and Information Quality*, 2:8:1–8:28, 2011.
- [8] M. Bobrowski, M. Marré, and D. Yankelevich. A Homogeneous Framework to Measure Data Quality. In *Proc. IQ*, pages 115–124. MIT, 1999.
- [9] M. E. Brown, J. E. Pinzón, K. Didan, J. T. Morissette, and C. J. Tucker. Evaluation of the consistency of long-term NDVI time series derived from AVHRR, SPOT-vegetation, SeaWiFS, MODIS, and Landsat ETM+ sensors. *IEEE T. Geoscience and Remote Sensing*, 44(7-1):1787–1793, 2006.

- [10] P. Buneman, A. Chapman, and J. Cheney. Provenance management in curated databases. In *Proc. of the 2006 ACM SIGMOD Int. Conf. on Management of data*, pages 539–550, New York, NY, USA, 2006. ACM.
- [11] CEPEA. Center of Advanced Studies in Applied Economics. <http://cepea.esalq.usp.br/pib/>, 2012. Accessed in June 2012.
- [12] A. D. Chapman. Principles of Data Quality. *Global Biodiversity Information Facility, Copenhagen*, 2005.
- [13] Y. Cheah and B. Plale. Provenance Analysis: Towards Quality Provenance. In *Proc. 8th IEEE Int. Conf. on eScience 2012*, 2012.
- [14] P. Chen, B. Plale, and M. S. Aktas. Temporal Representation for Scientific Data Provenance. In *Proc. 8th IEEE Int. Conf. on eScience 2012*, 2012.
- [15] J. Cheney, L. Chiticariu, and W. Tan. Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases*, 1(4):379–474, 2009.
- [16] N. R. Chrisman. The Role of Quality Information in the Long-term Functioning of a Geographic Information System. *Cartographica*, 21(2/3):79–87, 1984.
- [17] R. G. Congalton and K. Green. *Assessing the accuracy of remotely sensed data: principles and practices*. Number 13. CRC Press, Boca Raton, FL, 2 edition, 2009.
- [18] CountrySTAT. Food and Agriculture Organization of the United Nations. www.fao.org/countrystat, 2012. Accessed on March 2012.
- [19] D. C. Cugler, C. B. Medeiros, and F. Toledo. An architecture for retrieval of animal sound recordings based on context variables. *Concurrency and Computation - Practice and Experience*, 2012.
- [20] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu. An Approach to Evaluate Data Trustworthiness Based on Data Provenance. In *Proc. of the 5th VLDB Workshop on Secure Data Management*, pages 82–98, Berlin, Heidelberg, 2008. Springer-Verlag.
- [21] J. Davies, R. Studer, and P. Warren. *Semantic Web Technologies: Trends and Research in Ontology-based Systems*. Wiley, 2006.
- [22] DCMI. The Dublin Core Metadata Initiative. <http://dublincore.org/>, 2010. Accessed in May 2011.
- [23] D.W. Deering. *Rangeland Reflectance Characteristics Measured by Aircraft and Spacecraft Sensors*. PhD thesis, Texas A&M Univ., College Station, 1978. 338 pp.

- [24] P. J. DeVries. Geospecies ontology. <http://bioportal.bioontology.org/ontologies/1247>, 2009. Accessed in January 2013.
- [25] L. Ding, P. Kolari, T. Finin, A. Joshi, Y. Peng, and Y. Yesha. On Homeland Security and the Semantic Web: A Provenance and Trust Aware Inference Framework. In *AAAI Spring Symposium: AI Technologies for Homeland Security*, pages 157–160. AAAI, 2005.
- [26] DwC. Darwin core task group. <http://www.tdwg.org/standards/450/>, 2009. Accessed February 2013.
- [27] eFarms. <http://proj.lis.ic.unicamp.br/efarms/>, 2008. Accessed in June 2012.
- [28] FAO. *Land Quality Indicators and Their Use in Sustainable Agriculture and Rural Development*. FAO Land and Water Bulletin. 1997. Accessed in January 2012.
- [29] FAO. Food and Agriculture Organization of the United Nations. <http://www.fao.org/>, 2012. Accessed on March 2012.
- [30] FGDC. Content Standard for Digital Geospatial Metadata FGDC-STD-001-1998. Technical report, US Geological Survey, 1998.
- [31] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice Hall Press, 2008.
- [32] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 3rd edition edition, 2006.
- [33] M. F. Goodchild and L. Li. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1:110–120, 2012.
- [34] O. Hartig. Provenance information in the web of data. In *Proc. of the 2nd Workshop on Linked Data on the Web (LDOW2009)*, 2009.
- [35] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *Proc. of the Workshop on Semantic Web and Provenance Management at ISWC*, 2009.
- [36] International Monetary Fund. Data Quality Assessment Framework. <http://dsbb.imf.org/>, 2003. Accessed on January 2012.
- [37] ISO 19115. Geographic information – Metadata. <http://www.iso.org/iso/>, 2003. Accessed on January 2012.

- [38] A. Jøsang, R. Ismail, and C. Boyd. A Survey of Trust and Reputation Systems for Online Service Provision. *Decis. Support Syst.*, 43:618–644, 2007.
- [39] Kepler. The Kepler Project. <https://kepler-project.org/>, 2011. Accessed in January 2013.
- [40] A. A. Kondo, C. B. Medeiros, E. Bacarin, and E. R. M. Madeira. Traceability in Food for Supply Chains. In *Proc. 3rd Int. Conf. on Web Information Systems and Technologies (WEBIST)*, pages 121–127. INSTICC, 2007.
- [41] F. O. Kyeyago, E. M. Zake, and S. Mayinza. In the Construction of an International Agricultural Data Quality Assessment Framework (ADQAF). In *The 5th Int. Conf. on Agricultural Statistics (ICAS V)m*, 2010.
- [42] O. Lassila and R. R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>, 1999. Accessed in January 2013.
- [43] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang. AIMQ: a methodology for information quality assessment. *Information & Management*, 40(2):133–146, 2002.
- [44] F. Lemos. *Infrastructure and Algorithms for Information Quality Analysis and Process Discovery*. PhD thesis, Ingénierie des Systèmes d’Information, 2013.
- [45] R. S. Lunetta and J. G. Lyon. *Remote Sensing and GIS Accuracy Assessment*. CRC Press, 2004.
- [46] C. G. N. Macário and C. B. Medeiros. A Framework for Semantic Annotation of Geospatial Data for Agriculture. *Int. J. Metadata, Semantics and Ontology - Special Issue on Agricultural Metadata and Semantics*, 4(1/2):118–132, June 2009.
- [47] S. Madnick and H. Zhu. Improving data quality through effective use of data semantics. *Data Knowl. Eng.*, 59:460–475, 2006.
- [48] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu. Overview and Framework for Data and Information Quality Research. *J. Data and Information Quality*, 1:2:1–2:22, 2009.
- [49] J. E. G. Malaverri and C. B. Medeiros. Handling Provenance in Biodiversity. In *Proc. Workshop on Challenges in eScience (CIS)*, 2010.
- [50] J. E. G. Malaverri and C. B. Medeiros. Data Quality in Agriculture Applications. In *Proc. XIII Brazilian Symposium on GeoInformatics (GeoInfo)*, 2012.

- [51] J. E. G. Malaverri and C. B. Medeiros. A Provenance-based Approach to Evaluate Data Quality in eScience. *Int. J. Metadata, Semantics and Ontology - Special Issue on "Metadata for e-science and e-research"*, 2013. Submitted for publication.
- [52] J. E. G. Malaverri and C. B. Medeiros. Estimating the quality of data using provenance: a case study in eScience. In *Proc. 19th Americas Conf. on Information Systems*, 2013. Accepted for publication.
- [53] J. E. G. Malaverri, C. B. Medeiros, and R. C. Lamparelli. A Provenance Approach to Assess the Quality of Geospatial Data. In *Proc. 27th Annual ACM Symposium on Applied Computing*, pages 2043–2044, 2012.
- [54] C. B. Medeiros and A. C. de Alencar. Data Quality and Interoperability in GIS. In *Proc. of GeoInfo*, 1999. In portuguese.
- [55] R. A. Moraes and J. Rocha. Imagens de coeficiente de qualidade (Quality) e de confiabilidade (Reliability) para seleção de pixels em imagens de NDVI do sensor MODIS para monitoramento da cana-de-açúcar no estado de São Paulo. In *Proc. of Brazilian Remote Sensing Symposium*, 2011.
- [56] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. T. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. G. Stephan, and J. V. den Bussche. The Open Provenance Model core specification (v1.1). *Future Generation Comp. Syst.*, 27(6):743–756, 2011.
- [57] J. D. Myers, J. Futrelle, J. Gaynor, J. Plutchak, P. Bajcsy, J. Kastner, K. Kotwani, J. Sung Lee, L. Marini, Rob Kooper, R. McGrath, T. McLaren, A. Rodriguez, and Y. Liu. Embedding Data within Knowledge Spaces. *CoRR*, 2009.
- [58] NASA. National Aeronautics and Space Administration. <https://wist.echo.nasa.gov/api/>, 2012. Accessed in April 2012.
- [59] F. Naumann. From Databases to Information Systems - Information Quality Makes the Difference. In *Proc. IQ*, 2001.
- [60] F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems*, volume 2261. Springer, 2002.
- [61] F. Naumann and C. Rolker. Assessment Methods for Information Quality Criteria. In *IQ*, pages 148–162. MIT, 2000.
- [62] NCSA. National center for supercomputing applications. <http://leovip217.ncsa.uiuc.edu/>, 2009. Accessed in May 2013.

- [63] A. Parsian. Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. *Decis. Support Syst.*, 42:1494–1502, 2006.
- [64] G. Z. Pastorello Jr. *Managing the lifecycle of sensor data: from production to consumption*. PhD thesis, Institute of Computing, 2008.
- [65] B. Pernici and M. Scannapieco. Data Quality in Web Information Systems. In *Proc. of the 21st Int. Conf. on Conceptual Modeling*, pages 397–413. Springer-Verlag, 2002.
- [66] E. M. Pierce. Assessing data quality with control matrices. *Commun. ACM*, 47:82–86, 2004.
- [67] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data Quality Assessment. *Commun. ACM*, 45:211–218, 2002.
- [68] N. Prat and S. Madnick. Measuring Data Believability: A Provenance Approach. In *Proc. of the 41st Hawaii Int. Conf. on System Sciences*, page 393, 2008.
- [69] T. C. Redman. *Data quality : The Field Guide*. Digital Pr. [u.a.], 2001.
- [70] M. Reiter, U. Breitenbücher, S. Dustdar, D. Karastoyanova, F. Leymann, and Hong-Lin Truong. A Novel Framework for Monitoring and Analyzing Quality of Data in Simulation Workflows. In *Proc. of the 2011 IEEE 7th International Conf. on eScience*, pages 105–112. IEEE, 2011.
- [71] S. S. Sahoo, R. S. Barga, J. Goldstein, and A. Sheth. Provenance algebra and materialized view-based provenance management. Technical report, Microsoft Research, 2008.
- [72] S. S. Sahoo, A. P. Sheth, and C. A. Henson. Semantic Provenance for eScience: Managing the Deluge of Scientific Data. *IEEE Internet Computing*, 12(4):46–54, 2008.
- [73] S. de F. M. Sampaio, C. Dong, and P. Sampaio. Incorporating the Timeliness Quality Dimension in Internet Query Systems. In *WISE Workshops*, volume 3807 of *Lecture Notes in Computer Science*. Springer, 2005.
- [74] H. Scholten and A. J. U. T. Cate. Quality assessment of the simulation modeling process. *Comput. Electron. Agric.*, 22(2-3):199–208, 1999.
- [75] G. Shankaranarayanan and Yu Cai. Supporting data quality management in decision-making. *Decis. Support Syst.*, 42:302–317, 2006.

- [76] Y. Simmhan and B. Plale. Using Provenance for Personalized Quality Ranking of Scientific Datasets. *I. J. Comput. Appl.*, 18(3):180–195, 2011.
- [77] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31–36, 2005.
- [78] Swift. The swift project. <http://www.ci.uchicago.edu/swift/>, 2007. Accessed in May 2013.
- [79] Taverna. The Taverna Project. <http://www.taverna.org.uk/>, 2009. Accessed in January 2013.
- [80] Trident. Project trident: A scientific workflow workbench. <http://research.microsoft.com/en-us/collaboration/tools/trident.aspx>, 2009. Accessed in May 2013.
- [81] U.S. Agency for International Development. TIPS 12: Data Quality Standards. <http://www.usaid.gov/policy/evalweb/documents/TIPS-DataQualityStandards.pdf>, 2009. Accessed in January 2012.
- [82] VisTrails. The VisTrails Project. <http://www.vistrails.org>, 2011. Accessed in January 2013.
- [83] A. Voisard and G. Medeiros, C. B. and Jomier. Database Support for Cooperative Work Documentation. In *Proc. of COOP'2000*, 2000.
- [84] W3C. The PROV Ontology. <http://www.w3.org/TR/prov-o/>, 2012. Accessed in January 2013.
- [85] R. Y. Wang and D. M. Strong. Beyond accuracy : What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34, 1996.
- [86] X. Wang, R. Gorlitsky, and J. S. Almeida. From XML to RDF: how semantic web technologies will change the design of 'omic' standards. *Nat Biotech*, 23(9):1099–1103, 2005.
- [87] J. Widom. Trio: A System for Integrated Management of Data, Accuracy, and Lineage. In *Proc. of the 2nd Biennial Conf. on Innovative Data Systems Research (CIDR)*, 2005.
- [88] J. Xie and F. Burstein. Using machine learning to support resource quality assessment: an adaptive attribute-based approach for health information portals. In *Proc. of the 16th Int. Conf. on Database Systems for Advanced Applications*, 2011.

- [89] S. H. Yeganeh, O. Hassanzadeh, and R. J. Miller. Linking Semistructured Data on the Web. In *Proc. 14th Int. Workshop on the Web and Databases*, 2011.
- [90] J. Zhao, C. Goble, R. Stevens, and D. Turi. Mining Taverna's semantic web of provenance. *Concurr. Comput. : Pract. Exper.*, 20:463–472, April 2008.