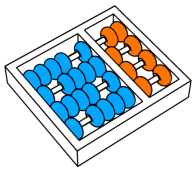
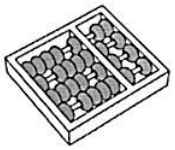


Renato Beserra Sousa

"Quality Flow: a collaborative quality-aware platform for experiments in eScience"

"*Quality Flow: uma plataforma colaborativa orientada a qualidade para experimentos em eScience*"

CAMPINAS
2015



University of Campinas
Institute of Computing

Universidade Estadual de Campinas
Instituto de Computação

Renato Beserra Sousa

"Quality Flow: a collaborative quality-aware platform for experiments in eScience"

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

"Quality Flow: uma plataforma colaborativa orientada a qualidade para experimentos em eScience"

MSc Dissertation presented to the Post Graduate Program of the Institute of Computing of the University of Campinas to obtain a Mestre degree in Computer Science.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Mestre em Ciência da Computação.

THIS VOLUME CORRESPONDS TO THE VERSION OF THE DISSERTATION SUBMITTED TO EXAMINING BOARD BY RENATO BESERRA SOUSA, UNDER THE SUPERVISION OF PROFA. DRA. CLAUDIA MARIA BAUZER MEDEIROS.

ESTE EXEMPLAR CORRESPONDE À VERSÃO DA DISSERTAÇÃO APRESENTADA À BANCA EXAMINADORA POR RENATO BESERRA SOUSA, SOB ORIENTAÇÃO DE PROFA. DRA. CLAUDIA MARIA BAUZER MEDEIROS.

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

CAMPINAS
2015

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

So85q

Sousa, Renato Beserra, 1988-
Quality Flow : a collaborative quality-aware platform for experiments in eScience / Renato Beserra Sousa. – Campinas, SP : [s.n.], 2015.

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

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

Informações para Biblioteca Digital

Título em outro idioma: Quality Flow : uma plataforma colaborativa orientada a qualidade para experimentos em eScience
Palavras-chave em inglês:
Management information systems - Quality control
Workflow
Workflow management systems
Framework (Computer program)
Databases - Development
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Claudia Maria Bauzer Medeiros [Orientador]
Cecília Mary Fischer Rubira
Leonardo Montecchi
Data de defesa: 12-06-2015
Programa de Pós-Graduação: Ciência da Computação

TERMO DE APROVAÇÃO

Defesa de Dissertação de Mestrado em Ciência da Computação, apresentada pelo(a) Mestrando(a) **Renato Beserra Sousa**, aprovado(a) em **12 de junho de 2015**, pela Banca examinadora composta pelos Professores(as) Doutores(as):

Prof(a). Dr(a). Leonardo Montecchi
Titular

Prof(a). Dr(a). Cecília Mary Fischer Rubira
Titular

Prof(a). Dr(a). Claudia Maria Bauzer Medeiros
Presidente

Institute of Computing /*Instituto de Computação*
University of Campinas /*Universidade Estadual de Campinas*

Quality Flow: a collaborative quality-aware platform for experiments in eScience

Renato Beserra Sousa

June 12, 2015

Examiner Board/*Banca Examinadora*:

- Profa. Dra. Claudia Maria Bauzer Medeiros (Supervisor/*Orientadora*)
- Profa. Dra. Cecília Mary Fischer Rubira
  Institute of Computing - UNICAMP
- Prof. Dr. Leonardo Montecchi
  University of Firenze
- Prof. Dr. André Santanchè
  Institute of Computing - UNICAMP (Substitute/*Suplente*)
- Dra. Carla Geovana do Nascimento Macario
  Embrapa - CNPTIA (Substitute/*Suplente*)

Abstract

Many scientific research procedures rely upon the analysis of data obtained from heterogeneous sources. The validity of the research results depends, among others, on the quality of data. Data quality is a topic that has pervaded computer science research for decades. Though there are many proposals for data quality assessment, there are still open problems such as mechanisms to support flexible quality assessment and ways to derive data quality. The goal of this dissertation is to work on these issues. The main contribution of this dissertation is the proposal of QualityFlow: a quality-aware collaborative platform for experiments in eScience. The following contributions were accomplished: to support the creation of quality-aware scientific workflows, allowing the addition of quality attributes to workflows, while at the same time letting distinct users define their specific quality metrics for the same workflow; to allow users to keep track of different quality assessments for a given process, thereby providing insights into the actual value of data and workflow; and to allow scientists to customize data quality dimensions and quality metrics collaboratively. QualityFlow was developed as a web prototype, and executed in two experiments - one based upon a real problem and the other on a sample workflow.

Resumo

Muitos procedimentos de pesquisa científica dependem da análise de dados obtidos de fontes de dados heterogêneas. A validade dos resultados de pesquisa depende, entre outros, da qualidade dos dados - um tópico recorrente na pesquisa em computação há décadas. Embora existam muitas propostas para a avaliação da qualidade de dados, ainda há problemas em aberto, como mecanismos flexíveis para a avaliação de qualidade e maneiras para derivar a qualidade dos dados. O objetivo desta dissertação é trabalhar nesses problemas. A principal contribuição da dissertação é a criação do QualityFlow: uma plataforma colaborativa para avaliação de qualidade para experimentos em eScience. As principais contribuições são: suportar à criação de workflows científicos com parâmetros de qualidade, permitindo a adição de atributos de qualidade a workflows, permitindo ao mesmo tempo que usuários disintos definam métricas de qualidade específicas para o mesmo workflow; permitir aos usuários manter o histórico de diferentes avaliações de qualidade para um mesmo processo, provendo assim melhor compreensão do real valor dos dados e workflows; e permitir aos cientistas customizar dimensões de qualidade de dados e métricas de qualidade colaborativamente. O QualityFlow foi desenvolvido como um protótipo web, e executado para dois experimentos – um baseado em dados reais e o outro em um workflow de exemplo.

Acknowledgements

I would like to thank my advisor, professor Claudia Bauzer Medeiros, not only for her great work but also for her ability to handle my concerns. This gratitude extends to all members of LIS, for their great feedbacks and companionship.

A special thank you is reserved for my family: for my parents, who have always supported my studies above their own needs; for my sister that (almost) always hears my complains; for my brother who always shows how little my problems are; and for my fiancée, who is always by my side.

I would like to thank all my friends - in particular Thiago and his family, who welcomed me in a time of trouble.

I would also like to thank my employer Samsung, for the time given to conclude this dissertation. Finally, this work was directly financed by CAPES (01P-02171-2012, 01P-4528-2013 and 01P-3501-2014). Complementary funding was obtained from individual projects from CNPq, and by projects NAVSCALES (FAPESP 2011/52070-7), the Center for Computational Engineering and Sciences (FAPESP CEPID 2013/08293-7), CNPq-FAPESP INCT in eScience (FAPESP 2011/50761-2), and the INCT in Web Science.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The term eScience can be defined as joint research in Computer Science and other domains, to let scientists from these domains conduct their research faster, better or in a different way, while at the same time advancing the state of the art in Computer Science. eScience is about global collaboration in key areas of science and the next generation of infrastructure that will enable it [19]. Data is a central part of eScience research: data analysis, processing, sharing and visualization provide the basis for eScience studies.

Data is the medium for experiments that lead to new discoveries. The validity of research results relies upon, among others, the quality of the data used in that research. Therefore, mechanisms that improve assessment of data quality are needed to endorse discoveries provided by data analysis.

Figure 1.1 shows a schematic example of an eScience collaborative work, that illustrates the motivation behind this dissertation. Research Group C wants to use the weather station data for a study, but needs it to be preprocessed - e.g., by groups A and B. Data produced by these other research groups meets the requirements of group C. In this scenario, data quality information will help group C decide which data product (from A or B) is better for the study. Therefore, in order to proceed, group C needs to assess the quality of both data products. There is extensive work on quality assessment - e.g., [5, 8, 25, 28]. However, there are still open problems such as mechanisms to support quality assessment and ways to derive data quality, both of which are the emphasis of our contribution. In particular, our approach is based on deriving quality using provenance information.

Scientific collaborations require the management and analysis of data sources provided by many groups, each with their own models, processes and methods. Therefore, some kind of mechanism is needed to coordinate the execution of different processes on data, performed by geographically distributed scientists.

Workflow systems have been adopted as a part of the infrastructure that allows the

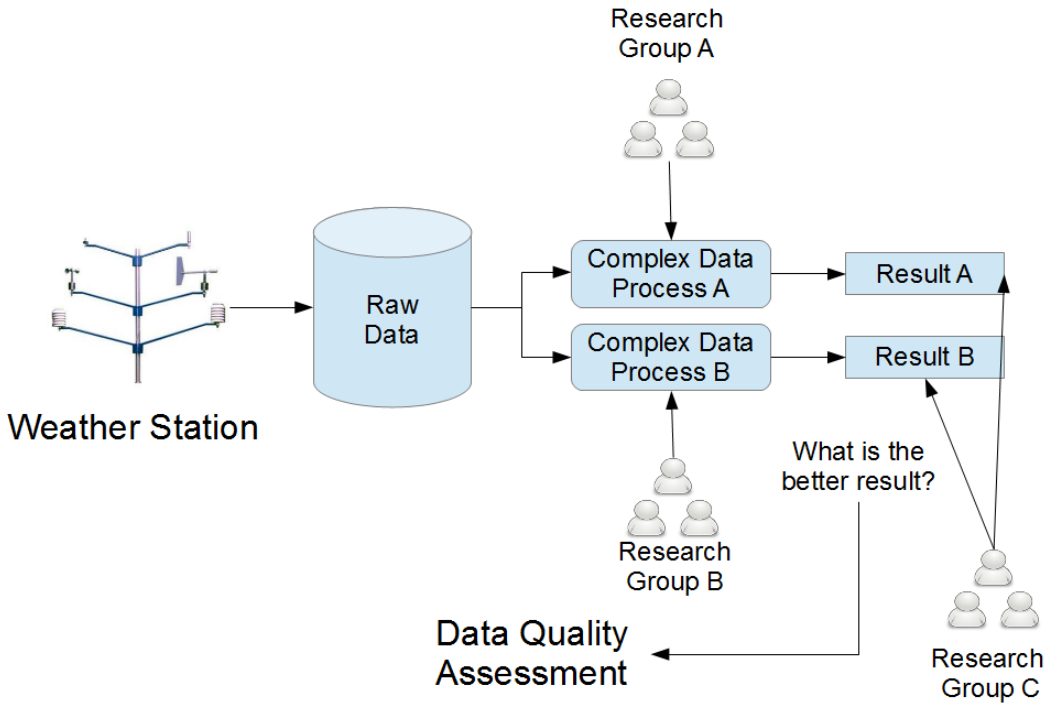


Figure 1.1: Example of an eScience Typical Scenario

coordination of multi-step, distributed data processing, involving many scientific groups. The term *workflow* is used to describe a set of tasks and procedures organized to achieve a goal. "Scientific workflows describe the scientific process from experimental design, data capture, integration, processing, and analysis that leads to scientific discovery", as defined by Lacroix et al. [24]. Workflow systems like Kepler [6] and Taverna [21] [20] organize the execution of experiments that rely on data analysis on many scientific domains. The entire process of Figure 1.1, as well as Complex Data Processes A and B may be seen as workflows, which are executed using some workflow management system (WFMS). We will keep this interpretation throughout this text.

Given this motivating scenario, let us return to the data quality problem. Quality of data is ideally derived from its 'fitness for use'. According to Wang et al. [38], one may define a set of metadata quality attributes to measure quality. These attributes can be quantitative or qualitative. Their meaning and weight might change according to the domain or researcher point of view. Related work also has shown that the concept of data provenance [40] [22] [9] is deeply connected with data quality assessment. *Data provenance* is the history of a data element - how, when, where and by whom it was created. In this work, we consider the hypothesis that data quality of a element can be

derived from its provenance.

Returning to the example in Figure 1.1, research Group C can assess the quality of results A and B using data quality information that can be provided in many different ways: quality metadata from data providers or external sources, quality information derived from data provenance, custom metrics defined by external sources and so on.

Summing up, our main problem is to provide means to assess data quality, given that current mechanisms are not able to materialize the concept of fitness for use. Considering this problem, our main goals are: to propose a data quality assessment mechanism, that supports flexible and multifaceted data quality analysis and that is able to generate quality information from data provenance.

Given this context, the main contribution of this dissertation is the proposal of QualityFlow: a quality-aware collaborative platform to manage quality assessment for eScience applications. QualityFlow is based on Malaverri's work [28], a provenance-based approach for data-quality assessment. Her thesis [28] pointed out several open issues, some of which are covered by this work. Particularly, our specific contributions are:

- to support the creation of quality-aware scientific workflows, allowing users to add quality information to workflow specifications;
- to allow scientists to customize data quality dimensions and metrics collaboratively, so that the result of running a given workflow can have distinct assessments, depending on the user;
- to derive data quality information using a combination of provenance records and attributes defined by scientists;
- to support these contributions via the implementation of QualityFlow platform.

The chapters that follow present related work, the designed architecture, the implementation and conducted experiments with real scientific data and sum up the research. Part of this work has been published as follows: Renato Beserra Sousa, Daniel Cintra Cugler, Joana Esther Gonzales Malaverri, and Claudia Bauzer Medeiros. A provenance-based approach to manage long term preservation of scientific data. In Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on, pages 162–133. IEEE, 2014 [36].

Chapter 2

Background

QualityFlow is a workflow-based environment to support collaborative scientific experiments in which scientists need to assess the quality of data and processes, and where quality assessment strongly relies on provenance. This chapter discusses related work in this context. It starts by analyzing some related work regarding data quality and workflows. It also briefly discusses data quality models, data cleaning architectures, data quality assessment approaches and research on data provenance.

2.1 Data quality

Data quality is a multidimensional concept. According to [38], a *quality dimension* can be defined as a set of data quality attributes - e.g. completeness, precision, believability, accuracy - that allow to represent a particular characteristic of quality. Data quality assessment teams must decide which dimensions will be considered, depending on the application domain. In this work, we consider the following representation: a data quality dimension is a quality attribute, composed of a name, a value and a description. For example, accuracy is a quality dimension that appears in most studies on data quality. So, our representation of the dimension 'accuracy' is: {Name: Accuracy; Value: 0.7; Description: Numerical value between 0 and 1 that represents how close the data is to the correct value. }. We have included the description field in order to mitigate the impact of misinterpretation of the meaning of a dimension.

There are many studies related to data quality, which range from analysis of quality dimensions to systems proposed to help users in the assessment or improvement of data quality. This dissertation is focused on the latter subject. Therefore, this section presents some work related to quality-aware systems.

Lima's dissertation [26] is an example of such a system, and is based on the fact that when working with data integration [7] [29], we should consider the quality of data

measurements. That work also discusses system-related quality issues. Examples of those concepts include: (a) unused data do not remain correct for long; (b) data quality in an information system is a function of its use, i.e. it is not an intrinsic data property of the dataset; (c) data quality problems tend to get worse as the system gets older; and (d) data quality rules apply to both data and metadata. That work presents an architecture which integrates data from pluviometric sensors. This architecture has a specific module for data quality assessment. That module receives integrated data as input and considers the quality dimensions of completeness, absence of error, timeliness and amount of data for the data quality assessment. The module provides output metrics such as daily averages, estimated missing data and monthly totals. It does not directly help users in asssessing the quality of the output.

Gamble and Goble [18] and Malaverri [28] follow the same line of work, first reviewing work on quality dimensions and data provenance, then proposing systems for data quality assessment. Both concern eScience studies. Gamble and Goble focused on giving users support to assess data quality, regarding scientific data available on the web. Their work emphasizes the separation between the dimensions of trust, quality and utility, defining them as entities for the computation of data quality. It defines a mechanism to combine the assessment of the different dimensions by using provenance information. That work uses decision networks to assess data quality. In the network they typically model the decision of accepting or rejecting the data, according to quality and trust dimensions, and also considering a data utility variable. The output of the decision network is a utility index that can be used for scoring and ranking data.

Malaverri's thesis, which is extended in this dissertation, conducted a data quality analysis [38], distinguishing between quantitative and qualitative dimensions. She proposed a framework that uses data provenance to semi-automatically obtain information to be used for data quality assessment. The framework supports a methodology to assess the quality of data produced by scientific experiments. She proposed an architecture and two models for data provenance, one an extension to the Open Provenance Model [30] and the other based on the PROV [37] ontology. Section 2.3 presents more details on her work, since it is used as a basis in this dissertation.

Other studies concerning data quality analysis include Alencar [14], Na'im et al. [31] and Lemos [25]. In [14], Alencar analyzes the use of metadata to represent some quality dimensions in geographic information systems. Alencar also defined criteria for choosing data sources, debated about problems on data preprocessing and presented quality attributes used for data collection and conversion. Examples are positional accuracy, coverage, completeness, timeliness, attribute accuracy, realiability and provenance. Na'im et al. [31] describe a mechanism for assessment of data quality during workflow execution. There is a component that provides a real-time monitor for data quality. The user can

define some thresholds for data quality, so that the monitor shows the level of data quality based on intermediate results.

In a more recent work, Lemos' research group proposed an approach, called Qbox-Foundation [15], whose goal is to smooth the definition of appropriate metrics for quality measurement methods, according to the specific uses of an organization. The author proposed a new quality metamodel, which considers data quality assessment tools defined by users or provided by a third party. The research also provided a service oriented infrastructure and the definition of a multidimensional platform for quality analysis, that were validated with the development of a prototype. The input is based on the definition of quality goals. A quality goal is decomposed in a quality factor to which is applied a set quality metrics. A set of services are responsible for the enforcing of these quality metrics. Q-box provides visualization of quality measurements, values assigned to quality metrics, independently from the visualization of data.

Another quality-oriented framework appears in Al Balushi et al. [5], who discusses a framework and a tool to support requirements engineering activities considering quality ontologies and knowledge techniques based on the ISO/IEC 9126 quality model. They propose the ElicitO framework that is composed of: domain ontologies; process guidelines to support requirement engineering activities; specification of relationships between domain requirements and quality concepts; inter-relationships among quality attributes; and reasoning techniques for requirements engineering activities. ElicitO uses quality ontologies to support elicitation and prioritization of quality requirements. The ontology implements the quality attributes and metrics described by the ISO standard and the framework and tool provide a knowledge repository to the requirements analysts to uniformly deal with quality in the requirement engineering activities.

Another system-based quality implementation using workflows is proposed by Reiter et al. [35], who present an architecture of a web-service policy-based language to describe data quality requirements and capabilities in the context of simulation workflows. They extend the architecture of conventional WfMSs to create the "Quality of Data driven Simulation Workflow Environment Architecture", and define a "WS-Policy-based language for Quality of Data".

Studies such as those of Al Balushi et al. [5] and Reiter et al. [35] differ from ours in the approach and context but have the same goal of providing a quality-aware environment for data-centric activities. The work of Lemos, like ours, lets users define their quality dimensions, but is less customizable. These papers illustrate the heterogeneity of problems and solutions regarding the assessment of data quality.

Data cleaning is considered by many authors as a necessary step to assure data quality. In this dissertation we will take advantage of data cleaning as one of many data processes that generate data quality information.

The work of Rahm and Hai Do [34] is noteworthy in this context because they made a survey of different data cleaning approaches. They define the following steps for data cleaning: data analysis, definition of transformation workflow and mapping rules, verification, transformation and backflow of cleaned data.

Other studies on data cleaning appear in Cugler et al. [13] and Chapman [10]. Cugler studies the problem of improving the quality of metadata of biological observation databases, particularly on those related to observations of living beings, which are often used as a starting point for biodiversity analyses. The focus of the work is the curation of observation metadata, involving data cleaning procedures. Chapman also performed a study in biology about cleaning species records, to check nomenclature and taxonomic errors. The author discusses error prevention, nomenclature error handling and the appearance of errors due to the integration of different databases.

Throughout the text, we will mention the following dimensions, using our own definitions in some cases and Wang et al. [33] for others:

- accuraccy - the extent of systematic errors in a measurement;
- coverage - for geographic data, the extent of the target area that is covered in a measurement;
- precision - the extent of random errors in a measurement;
- freshness - the extent of the influence of measurement age for data validity;
- believability - the extent to which data is regarded as true and credible [33];
- completeness - the extent to which data is not missing and is of sufficient breadth and depth for the task at hand [33];
- free-of-error - the extent to which data is correct and reliable [33];
- reputation - the extent to witch data is highly regarded in terms of its source or content [33];
- timeliness - the extent to which the data is sufficiently up-to-date for the task at hand [33];

2.2 Scientific workflows

Ludaescher et al. [27] defined scientific workflows as process networks that are typically used as “data analysis pipelines” or for comparing observed and predicted data. In this

context, a process is a program or a manual procedure that composes a workflow. That can include a wide range of components, e.g., for querying databases, for data transformation and data mining steps, for execution of simulation codes on high performance computers, etc.

eScience research presents a vast literature on the use and development of scientific workflow systems to support, execute and monitor experiments. Barseghian et al. [6] describe extensions to the Kepler workflow management system, to allow the ease of use for data collection, processing and analysis for sensor network data and pre-stored historical data. That work publishes extensions to Kepler and generates workflows using these extensions to meet a variety of needs. They consider two case studies, one related to ecological soil sensors and the other on oceanography. The data collected was published using the open-source platform DataTurbine [16].

Another scientific workflow system is Taverna [39], which facilitates integration of tools and databases for scientific research, particularly by using web services. Taverna gained prominence by allowing the development of workflows to perform different analyses on bioinformatics, such as genetic sequence analysis and genome annotation [21]. Taverna has been used on a variety of studies on life sciences [32] [39] and also on general research, with contributions on other domains. Holl et al. [20], for example, propose an addition of a new optimization step to the workflow lifecycle, implemented as a Taverna extension - a similar approach could be used to add a quality verification step to the workflow lifecycle. Our dissertation will use Taverna, because of its widespread use. Moreover, it has been used in other research the Laboratory of Information Systems (LIS) of Unicamp.

2.3 Malaverri's work

The work of Malaverri [28] provides the basis for this dissertation, and for this reason is analyzed apart. Figure 2.1 shows the architecture proposed by Malaverri, in which the numbers identify the data flow among the components. The central idea is to derive data quality based on data provenance. Provenance is derived from monitoring the workflow that produces the data. In this architecture, raw data are acquired and processed (1) and stored in the Data Repository (2). These data are used by processes (3) that are retrieved from the Process Repository (4). Processes are run using workflows, webservices (WS) and workflow tools. 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.

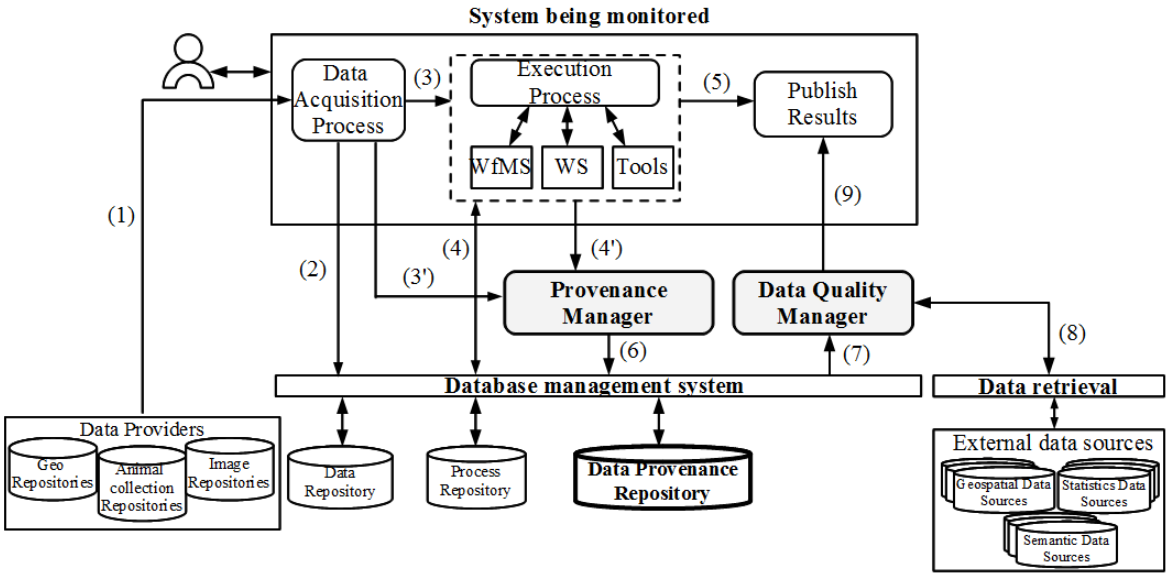


Figure 2.1: Reference architecture proposed by Malaverri [28]

The Data Quality Manager can also look for (8) information from external semantic data sources to complement the data obtained from the Provenance Repository. Semantic data sources are all kinds of sources in which data are stored together with means to attach semantics to them - eg., including a formal description of concepts, terms, and relationships within a given knowledge domain.

Although the Quality Manager is part of Malaverri's proposed framework, its architecture was not specified. Her thesis is centered on the Provenance Manager, and the Quality Manager was indicated as a component of the architecture given its importance for better quality assessment. As will be seen, our work specifies and provides a prototype of the Quality Manager, extends the Provenance Manager, and adds the Quality Adapter modules, thus providing useful functionaliy for filling the gap between these modules.

2.4 Summing up

Data quality is topic with a great variety of research lines. This chapter presented studies that approach this problem with different perspectives, illustrating the most important subtopics related to our research. The most common approach we have observed to deal with data quality is the proposal of models and systems to deal with it for specific use cases. We have followed this concept from the literature, as our approach proposes a data

model and a platform to handle data quality in a workflow environment.

Chapter 3

QualityFlow

Our work is centered on an environment to support eScience experiments, called QualityFlow, that relies on a Workflow Management System (WfMS). This collaborative platform provides tools to augment workflows, processes and data with quality metadata, and allows scientists to tailor quality metrics to their experiments. As a result, scientists may have a qualitative reference in re-using others' work and be able to assess the reliability of the result of their analyses.

3.1 Introduction

As defined in Chapter 1, a workflow is a specification (or model) of a process, which is a set of inter-dependent steps needed to complete a certain task [23]. A scientific workflow is the specification of design, data capture, integration, processing, and analysis in scientific experiments [24].

In our work, we concentrate on workflow processes and the data that flows through the workflow. Each of these processes executes some operation on these input data – e.g., extraction, computation, transformation or fusion. A process may be a local script call, a web service invocation, a database query, a subworkflow, and so on. Multiple outputs of each of these processes become inputs of others.

There are many modeling and implementation proposals to interconnect different processes and data sources to allow their orchestrated execution. As explained previously, we adopt the workflow model because existing WfMSs allow us to think in a higher level of abstraction. Moreover, eScience environments rely heavily on workflows.

We also assume that when the WfMS manages the execution of a workflow, it keeps a trace of the execution. This trace (that provides provenance information) is stored somewhere within the execution environment – e.g., a log – under some provenance data model like OPM [30], for example. From now on, to simplify the text, we refer to this set

of provenance records as TraceLog.

3.2 Description of QualityFlow

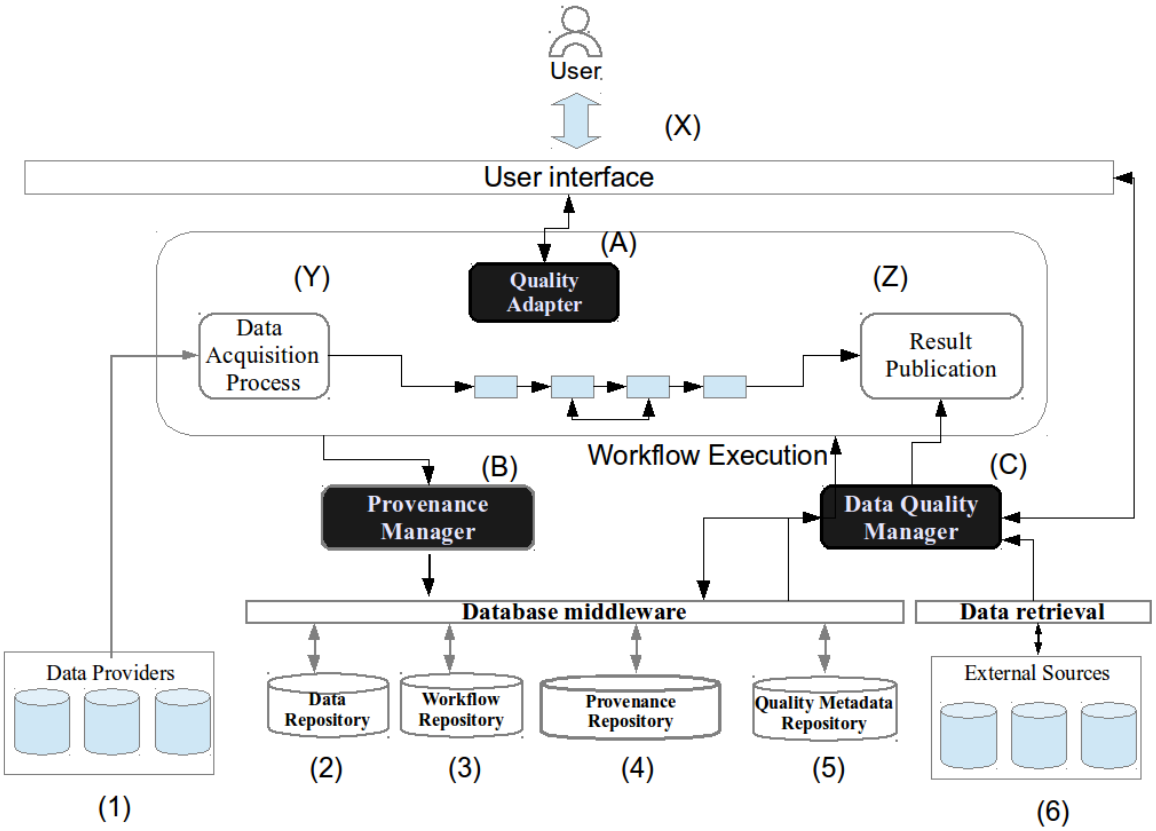


Figure 3.1: Schematic overview of QualityFlow

Figure 3.1 shows the main elements of QualityFlow, where the dark boxes concentrate the main contributions of this dissertation. Users are mainly scientists and experiments are run via workflow execution, where each workflow (or multiple workflows linked together) specifies an experiment. The modules of this execution architecture extend the information provided by the WfMS with additional quality information. In the end, QualityFlow provides scientists with means to evaluate the quality of the final result, using a provenance-based paradigm (see chapter 2). Quality parameters are either defined by experts (implemented as workflow annotations), or obtained from external sources. Quality

computation at the end of the workflow execution is based on parameters and functions provided by expert users. The overall goal is that, at the end, users are able to get provenance information on workflow execution, quality metrics associated to processes and data services, and request evaluation of specific quality metrics based on all this information.

The database middleware (and its repositories) stores references to workflow descriptions, executions, (meta)data, its processes and naturally the quality annotations added by users. It also persists all related metadata like workflow author, metadata author, execution timestamps etc. QualityFlow assumes there are two kinds of users:

1. Regular users – scientists who are involved in executing the workflow and analyzing results. They can assess the quality of these results by using quality dimensions, quality metrics and quality evaluation functions. They do not directly enter such factors into the system, but they can choose them from menus, and provide values to quality parameters.

2. Advanced users (quality designers) – scientists and/or computing experts (e.g., programmers, workflow designers) who can enter new quality parameters and evaluation functions. Such functions define how to calculate quantitative quality dimensions like accuracy, efficiency and precision or how to relate and summarize different qualitative quality dimensions – like timeliness, utility and so on. Such dimensions can be processed while the workflow is being executed, or, at the end, by user request. The results of such computations are stored in the Quality Metadata repository.

Data sources can be internal or external - respectively (1) and (6) in the figure. In both cases, they may have associated quality metadata. Quality metadata from internal sources is stored in the Quality Metadata repository - (5) in the figure. Quality metadata extracted from external sources may also be stored in this repository. Since the quality assessment process is user-driven in the current platform, means to extract and store external quality metadata automatically are a possible extension of this work as explained in Chapter 5.

The database middleware stores internal system metadata. This middleware is responsible for dealing with the following data repositories and tables:

- Data Repository (2): holds information about datasets used by workflows (i.e., worfklow inputs and outputs);

- Workflow Repository (3): stores workflow description files;

- Provenance Repository (4): stores TraceLogs files.

- Quality Metadata Repository (5): stores quality metadata on workflows, processes and data. This repository stores information about the provenance of the quality - e.g., who specified the dimension, and when. Thus, a given dataset may have distinct quality characteristics, depending on its use.

The architecture modules are divided in:

(i) Modules for workflow execution: The WfMS (omitted in the figure for readability) receives data from the Data Acquisition (Y) process, executes the workflow, and outputs results via the Result Publication (Z) module. Results may be published as a reference to a database, a web report,a file, etc.

(ii) Quality assessment modules- responsible for managing, extracting and processing quality information:

- Quality Adapter, (A) in the figure - allows advanced users to add quality dimensions to a workflow, a process or a TraceLog (which is stored in (4)). The Adapter does not change the workflow model.
- Provenance Manager (B) - tracks workflow execution and stores TraceLog information.
- Quality Manager (C) - allows users to define quality metrics and assess quality of published data.

The users interact with the system via the User Interface, (X) in the figure, which, as will be seen later on, was implemented as a web tool. Our system runs in the server-side of this tool. The WfMS is the execution environment for the workflows.

In an eScience scenario, different kinds of data sources are expected: environmental data, biological observation (spatio-temporal) databases, image repositories and so on. This list is not exhaustive, just serves as an illustration.

External data sources used by QualityFlow in quality assessment - (6) in the figure - are represented separately because they contain publicly available data, like semantic data sources, which can be accessed by the scientists but cannot be modified directly – e.g., myexperiment.org reputation index. The main role of these data sources is to provide additional data to validate data quality dimensions and metrics generated within the system.

3.3 Quality Processing

The Provenance Manager processes TraceLogs that have been augmented with quality information to store useful provenance information that can be retrieved by the Quality

Manager. This provenance information consists of metadata from both original data providers and also from intermediary processes, describing therefore the history of data. Provenance is a particular type of quality metadata and it is also stored in the quality metadata repository.

The Data Quality Manager is responsible for handling data quality metrics and assessing data quality, based on expert requirements. The definition of quality metrics may be as simple as defining equations on numeric values or as complex as set of inference rules. This module generates quality information from: (a) the provenance information stored by the Provenance Manager, (b) the quality metadata added to workflows by the Quality Adapter and (c) external data sources. The Quality Manager gets this information from, respectively, the Provenance Repository and the Quality Repository; quality is compiled according to metrics defined by scientists. In short, the Quality Manager uses all available information to derive a useful report for the user.

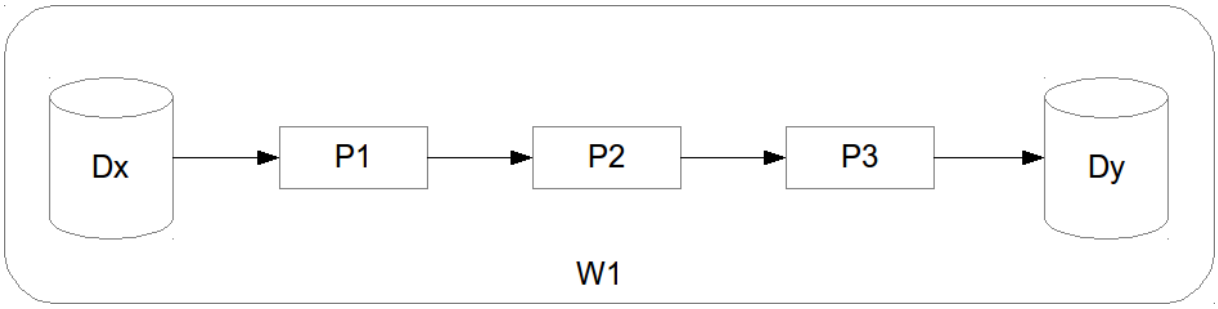


Figure 3.2: Example to illustrate provenance information, where rectangles are processes, and the arrows indicate data flow.

To illustrate these concepts, assume we have a workflow W1 composed of three processes – P1, P2 and P3 - as shown in Figure 3.2. That workflow has dataset Dx as input and Dy as output. The provenance information associated with Dy - retrieved from its TraceLog - is the sequence of processes executed when running W1 with Dx as input.

Assume the user wants to run W1, whose specification is already stored in the workflow repository. Assume also that a workflow designer already added quality dimensions to the workflow via the Quality Adapter. Then, when the workflow is executed, the following happens:

- Data is retrieved from Dataset Dx by process P1.
- P1 pre-processes the obtained data and delivers it to P2.

- P2 runs the experiment with the received data.
- P3 processes experimentation output for publication.
- Dy stores the results of the experiment.

In this case, QualityFlow stores not only that Dy was originated from Dx through execution of W1, but more detailed information, such as:

1. Dy was created from Dx using W1.
2. W1 is composed by P1, P2, P3.
3. W1 owner is U1.
4. Dx original metadata is Mdx, and has quality dimensions QA1 and QA2, added by U1.
5. W1 has quality dimensions QA3,QA4.

In this example, items 1. and 2. are directly computed by the WfMS's provenance mechanism and stored as TraceLog records by the Provenance Manager. Item 3. is obtained by the Quality Manager from our database middleware and items 4. and 5. are provided by the Quality Adapter and are available to the Quality Manager. Quality entries QA1, QA2, QA3 and QA4 are stored in QualityAnnotation table according to the schema shown in section 3.5

3.4 Data Model of the quality repositories

The data model for QualityFlow comprises the history of workflows executions and the related input and output data. This model has to consider also custom – per user - quality metrics.

Figure 3.3 shows the schema of the data model. This schema was conceived to allow the addition of quality information to the different elements – workflow, process, data source and data result – which later can be used by the Quality Manager.

The role of the data entities are the following:

- User - user credentials and type.
- Quality Metrics - quality functions and/or rules defined by users.
- Workflow - workflow description.

- TraceLog - provenance information.
- DataResult - output of workflow. In the database, the output can be stored as a path to a local file, a short text output or an external URI (that represents a complex result, like a video, an image, a time series).
- Process - processes that compose workflows.
- DataSource - data providers used in workflows.
- Quality Dimension - quality dimensions defined by user.
- Quality Annotation - quality values provided by workflow users.

Notice there is a mandatory one-to-one relationship between DataResult and TraceLog, which is the basis for obtaining quality information from provenance. This allows tracking result provenance and therefore deriving additional quality information.

This schema models a database - currently deployed in PostgreSQL.

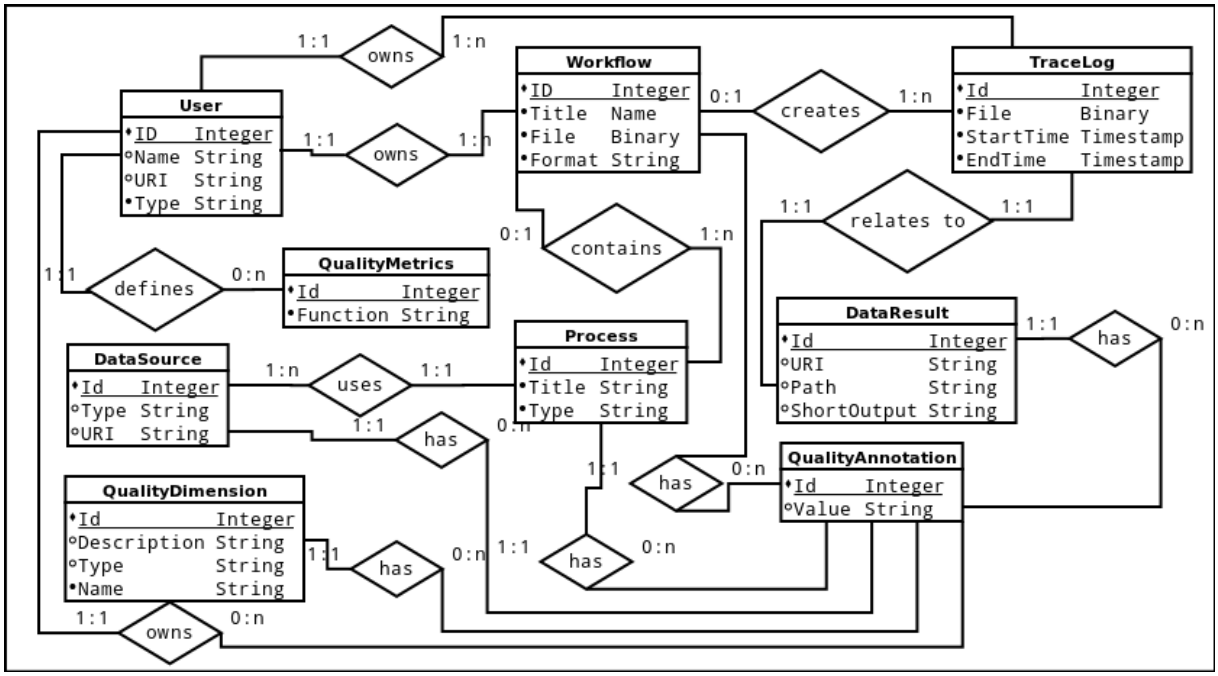


Figure 3.3: ER representation of the database schema.

3.5 Accumulating quality information

The model of section 3.4 allows to derive quality attributes from resources indicated by the user. For example, a quality dimension for a workflow can be derived given the quality annotations of its child processes. In the basic scenario, the user is able to view a workflow's result together with the quality information provided and the associated trace. This can become progressively more sophisticated. For instance, in a slightly more complicated case, the user can request the computation of specific quality dimensions at the end of a workflow's execution. Also, a user can request to view previous quality evaluations of the same execution, to compare assessments made by distinct experts.

With time, after several executions, as it accumulates a reasonable amount of quality data, the quality report of a given workflow will get richer. The system will be able to provide not only the immediate workflow quality annotations, but also each piece of quality information related to its processes, quality information provided by distinct users and also provenance of the quality information itself.

3.6 Example with a sample workflow

This section presents the use of a hypothetical workflow in QualityFlow to show the role of each component in more detail. Sections 4.3 and 4.4 of Chapter 4 presents more comprehensive experiments built upon real data.

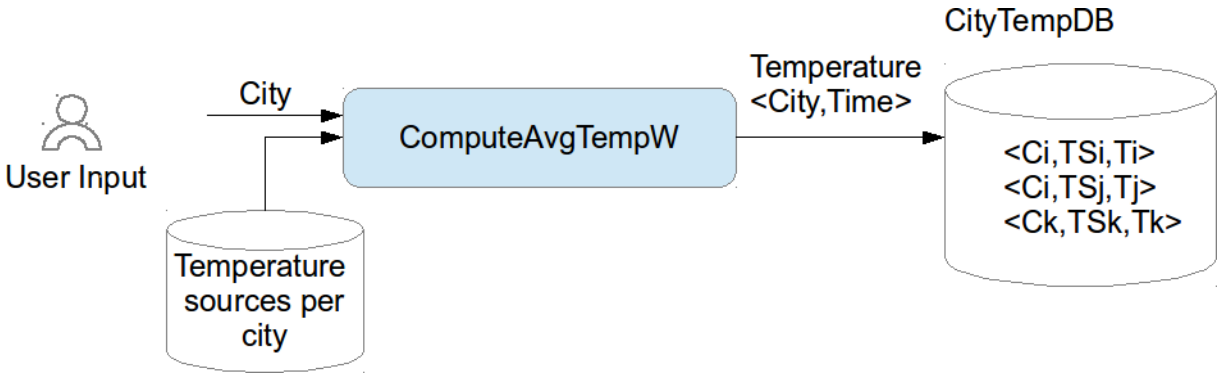


Figure 3.4: Workflow example

Figure 3.4 shows a sample workflow, called ComputeAvgTempW (CTW). It receives as input the name of a city and computes the average temperature at that city at the time (TS) of the request, storing that average in database CityTemperatureDB. It uses

as input a table that, for each city, indicates the temperature sources to check. Thus, one city may have one single source (e.g., weather station) while another may have many sources.

Assume that ComputeAvgTempW is stored in the Workflow Repository and that, using the Quality Adapter, expert user ExpertA adds to its data sources, for a given city C, the dimensions of Table 3.1. ExpertA also defines a custom quality metric QM1 as shown in equation 3.1.

Dimension / Source	Reputation (of owner) (R)	Coverage (C)	Precision (P)
University	0.8	0.20	0.99
Airport	0.6	0.40	0.90
ResearchInstitute	[UserInput]	0.80	0.99

Table 3.1: Quality dimensions

$$QM1 = 0.5 * P + 0.3 * R + 0.2 * C \tag{3.1}$$

In this scenario, regular user UserX performs the execution of this workflow in the WfMS. The execution trace can be directly obtained from the WfMS after the execution and stored in the TraceLog, being processed by the Provenance Manager. Since the trace was generated by ComputeAvgTempW, the quality dimensions are automatically 'connected' to the trace as well, through the database relationships.

Another regular user UserY can later use the results stored in CityTempDB. As explained in section 3.3, the provenance records provide information to connect the temperature data result to its originating workflow. This user can use the Quality Manager to visualize the quality metadata associated to each entry - due to the relationship between TraceLog and DataResult - and request the available quality metrics. In this case, quality metrics QM1 can be calculated, if the user provides the missing reputation quality attribute.

Furthermore, another expert user ExpertB may define additional quality dimensions and metrics (or new values for the existing dimensions). Again, those can be associated to the same workflow, or even, to a given workflow instance (DataResult and TraceLog). Thereupon, the history of quality assessment of ComputeAvgTempW will be constructed by uniting the analyses of each user.

Figure 3.5 illustrates the use of QualityFlow's components for this example. For example, in step (1.1) ExpertA's request to add quality dimensions is processed by the Quality Adapter, which accesses the Workflow repository to reference the annotated workflow.

The tables from 3.2 to 3.13 show what happens with QualityFlow tables in each step. Tables 3.2, 3.3, 3.4 and 3.5 show the initial state, with the registered users, the workflow, its composing process and the available data sources.

ID	Name	URI	Type
1	ExpertA	-	expert
2	UserX	-	regular
3	UserY	-	regular
4	ExpertB	-	expert

Table 3.2: Users in QualityFlow

ID	Title	File	Format	User_FK
1	ComputeAvgTempW	W.xml	t2flow	1

Table 3.3: Workflow added by ExpertA

ID	Title	Type	Workflow_FK
1	ComputeAvgTempP	-	1

Table 3.4: Process within ComputeAvgTempW

ID	Title	URI	Process_FK
1	UniversityTemp	-	1
2	AiportTemp	-	1
3	ResearchInstTemp	-	1

Table 3.5: Data source used by ComputeAvgTempW

Tables 3.6, 3.7, 3.8 and 3.9 show the quality dimensions, annotations and metrics added by ExpertA. Quality dimensions are presented directly, instead of the foreign keys, to improve readability.

ID	Name	Description	Type	User_FK
1	Reputation	-	decimal	1
2	Coverage	-	decimal	1
3	Precision	-	decimal	1

Table 3.6: Quality dimensions added by ExpertA

Tables 3.10 and 3.11 show the TraceLog and DataResults added by UserX.

ID	Dimension*	Value	User_FK
1	Reputation	0.8	1
2	Coverage	0.2	1
3	Precision	0.99	1
4	Reputation	0.6	1
5	Coverage	0.4	1
6	Precision	0.9	1
7	Coverage	0.8	1
8	Precision	0.99	1

Table 3.7: Quality annotations added by ExpertA

QA_ID	DataSource_ID	User_ID
1	1	1
2	1	1
3	1	1
4	2	1
5	2	1
6	2	1
7	3	1
8	3	1

Table 3.8: QualityAnnotation for DataSource by ExpertA

ID	Function	User_FK
1	CalcAverage	1

Table 3.9: Quality Metrics

ID	File	StartTime	EndTime	Workflow_FK	User_FK
1	TL1	t1	t2	1	2

Table 3.10: TraceLog added by UserX

ID	URI	Path	ShortOutput	Tracelog_FK
1	-	-	R1	1

Table 3.11: DataResult added by UserX

Finally, tables 3.12 and 3.13 show the annotations added by UserY, including the one generated from a custom quality metric.

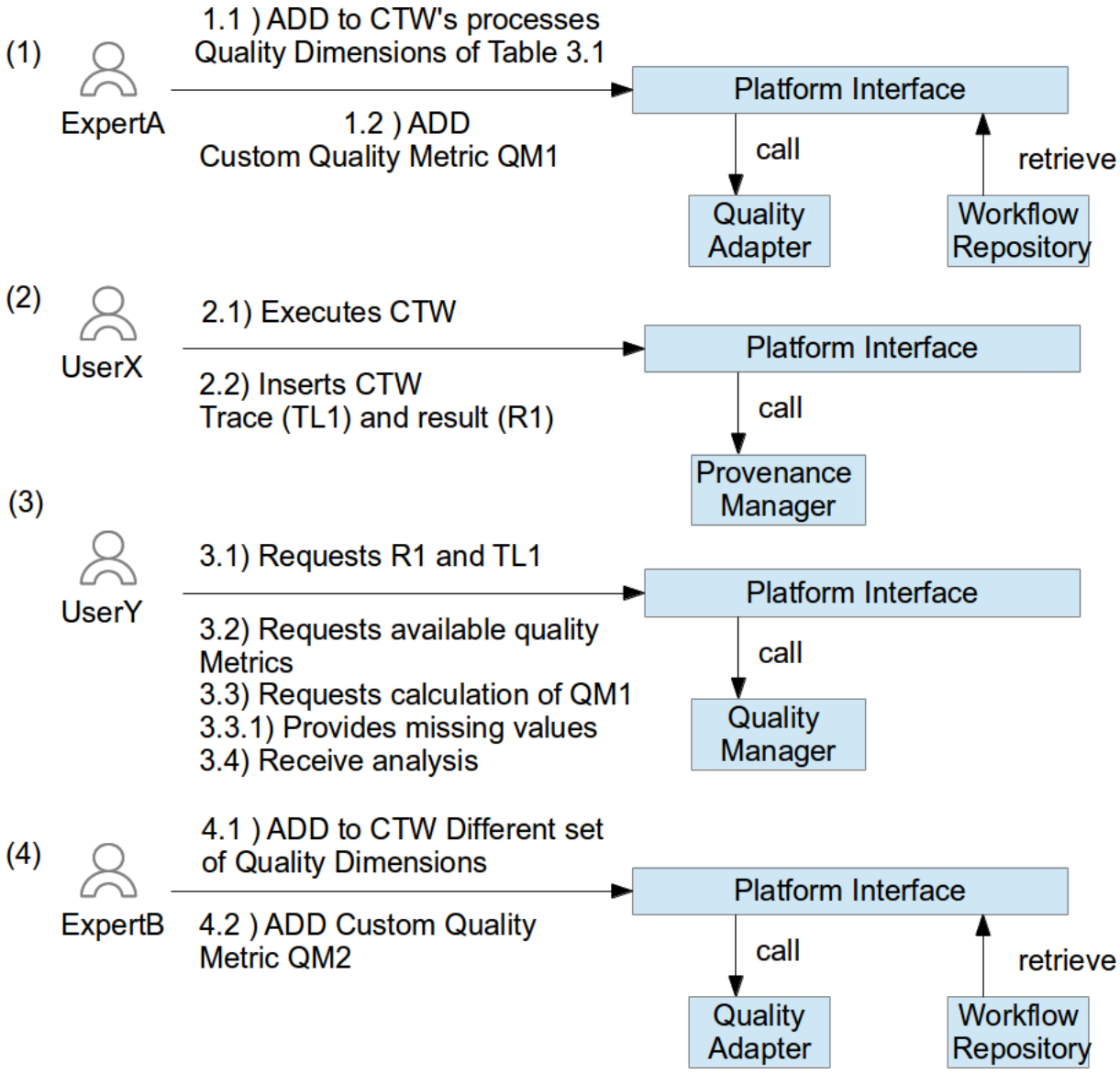


Figure 3.5: Illustration of platform use

ID	Dimension*	Value	User_FK
9	Reputation	0.7	3
10	CalcAverage	0.658	3
11	CalcAverage	0.6	3
12	CalcAverage	0.788	3

Table 3.12: Quality annotations added by UserY

QA_ID	DataSource_ID	User_ID
9	3	3
10	1	3
11	2	3
12	3	3

Table 3.13: QualityAnnotation for DataSource by UserB

3.7 Summing up

In this chapter, we have proposed QualityFlow - a collaborative platform to manage the problem of data quality in scientific experiments. It takes advantage of the WfMS environment to coordinate experiments and collaboration. It uses the provenance information inherent to the workflow environment. It also provides features to improve the quality information available to scientists - the addition of quality metadata and definition of custom quality metrics.

The underlying data model of QualityFlow supports custom quality metrics. Each user can define a set of quality properties that are relevant for their goals. Moreover, a set of quality metrics may be shared by other users.

A novel feature of this new platform is that the system itself generates new provenance information, which is stored for each quality annotation created/shared/modified by users. Moreover, the platform keeps track of users' individual assessments, thereby helping to compare distinct assessments (and even different quality criteria).

In short, this proposal groups three different, but related, aspects of data quality assessment: the use of quality dimensions, the concept of fitness for use (in custom quality metrics) and provenance information.

Chapter 4

Implementation aspects

This chapter provides an overview of the implementation of QualityFlow, and describes two experiments.

4.1 Implementation overview

The original idea was to materialize QualityFlow via implementation of a plug-in for the Taverna WFMS. In the beginning of this project, we noticed that the technical challenges to achieve the research goals would surpass the scientific challenges, therefore we decided to validate our work in a decoupled web platform. We were able to develop a simple interface for interchanging information between the workflow environment and our platform by combining workflow description files, workflow annotations and provenance log files. The platform was implemented using Django [2], a web development framework in Python [4]. Figure 4.1 shows where each technology was used.

4.2 Object-relational mapping

A core feature of the Django framework is its object-relational mapping, and we have taken advantage of it to simplify our implementation. Listing 4.1 shows the declaration of QualityDimension and QualityAnnotation Entities. Notice the class declarations correspond to the schema entities in Figure 3.3. Django automatically assigns an integer primary key to each entity by default, therefore it is omitted in the code. For instance, lines 2,3 and 4 show the attribute declaration of the QualityDimension entity. Lines 5 and 11 declare helper functions for visualization.

Listing 4.1: QualityFlow's object-relational mapping code

```
class QualityDimension(models.Model):
    name = models.CharField(max_length=100)
    description = models.TextField(blank=True)
    value_type = models.CharField(max_length=100)
    def __unicode__(self):
        return unicode("{} ({}): {}".format(self.name,self.
            value_type, self.description))

class QualityAnnotation(models.Model):
    dimension = models.ForeignKey(QualityDimension)
    value = models.CharField(max_length=100)
    def __unicode__(self):
        return unicode("{} = {}".format(self.dimension.name,self.
            value))
```

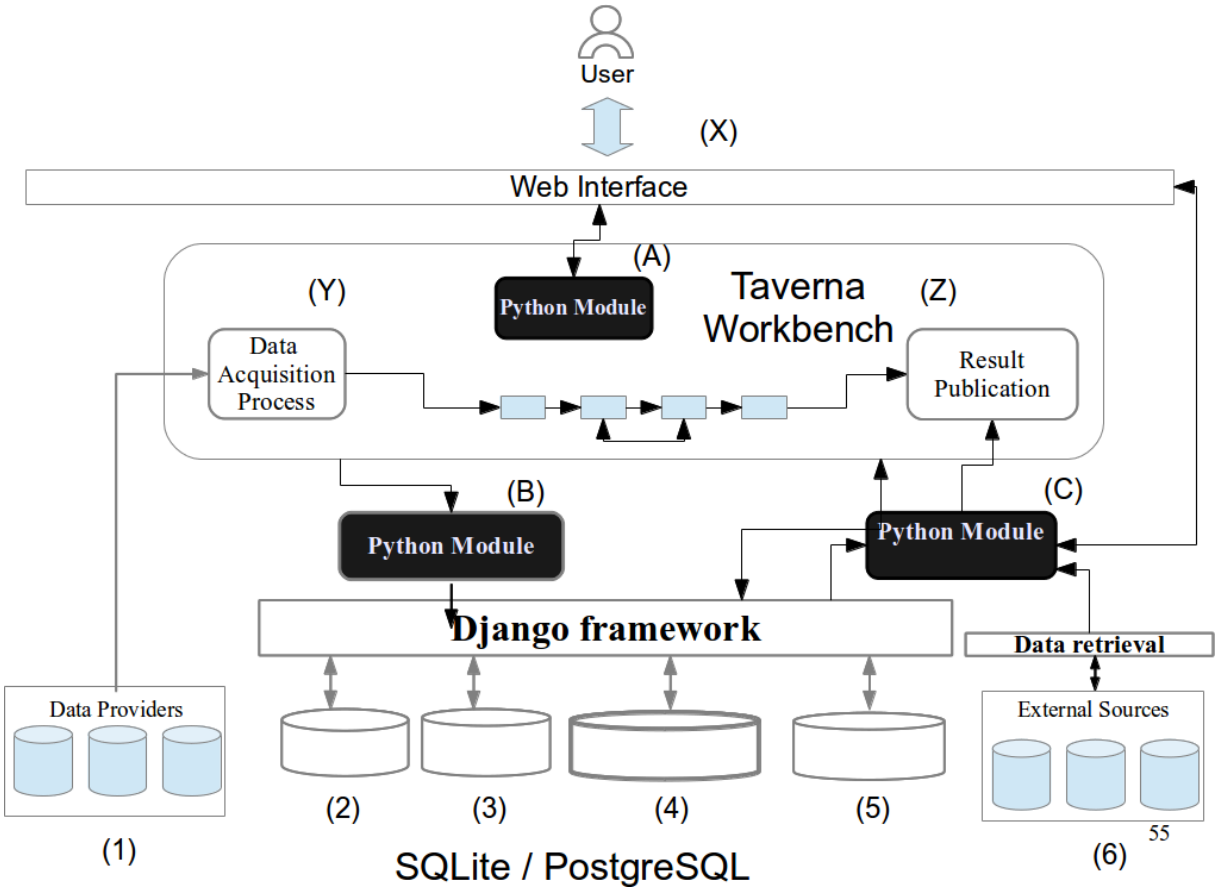


Figure 4.1: Implementation aspects of QualityFlow architecture

We have validated our work with two experiments in different knowledge domains. The first is geared towards supporting metadata-based retrieval in biodiversity observation databases; it shows usage of QualityFlow to augment workflows with quality metadata and custom metrics. The second explores a small workflow that gets weather information for cities, to show in detail every aspect of QualityFlow.

4.3 Experiment 1 - Preservation of Animal Sounds

This experiment was geared toward supporting metadata-based retrieval, where long term accessibility is associated with long term metadata curation; it was implemented considering FNJV quality issues. This experiment was published in the Workshop on Long Term Preservation for Big Scientific Data at ICDE [36] and it was based on Cugler et al. [12]. This study was conducted in a preliminary version of QualityFlow. The results were used to create the present version. Two directions were explored:

i. Improving quality by deriving and/or checking the contents of metadata fields using external authoritative sources;

ii. Enhancing data preservation by extending the set of metadata attributes, not originally contemplated by the scientists, thereby augmenting the scope of queries that can be supported, and increasing the chances of reuse of the associated data sets.

4.3.1 Overview of the study

The goal of the specified workflow was to detect outdated species names by contrasting such names with authoritative organizations which publish and maintain official species names lists, in our case the Catalogue of Life [1]. For this kind of activity of metadata curation, experts are interested in finding out the accuracy of the original metadata. Figure 4.2 shows a partial screenshot of the results. This prototype was implemented as part of Cugler's thesis [11].

Given a species name, if it is no longer valid, the Catalogue of Life web service informs what is the current up to date species name used. For each record in which the species name was detected as outdated in the FNJV database, the prototype persists the updated species name in a separate table and creates a reference between the original metadata record and the species name. This strategy is important in order to maintain the original collection unchanged – given, for instance, that several papers concerning that recording and the outdated name have been written. It also provides a historical log of metadata modifications. Before such names are persisted in the database, they are flagged to be checked by biologists. Cugler's prototype shows the progress of name checking, publishing

the number of distinct species names in the database, the number of records processed, the number of species names which were detected as outdated and the respective updated names.

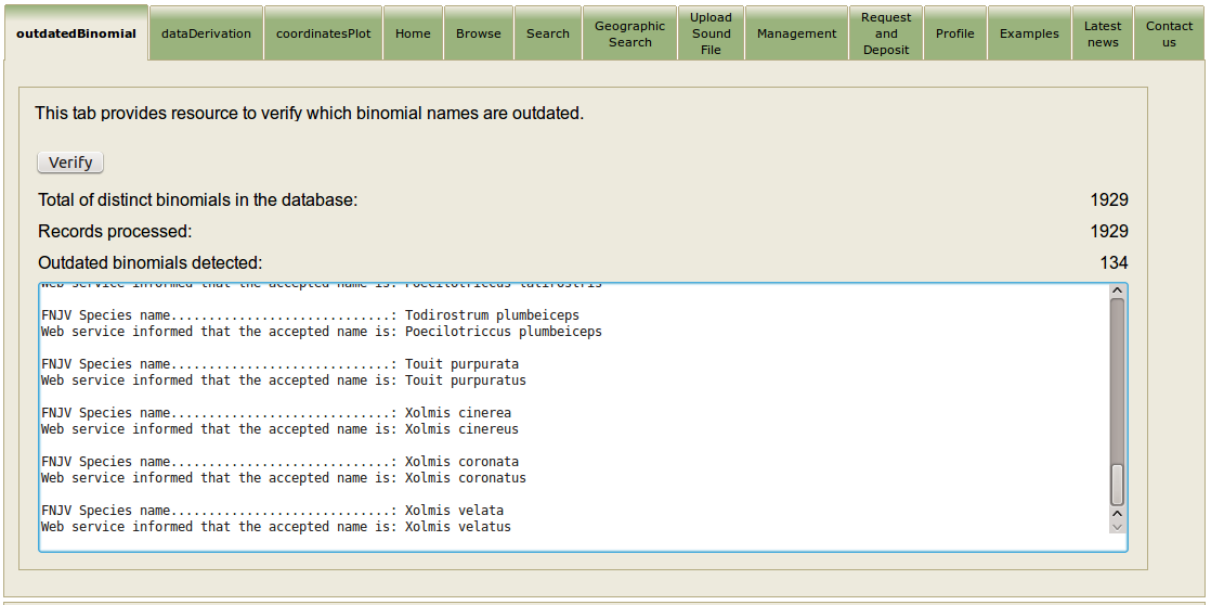


Figure 4.2: Prototype for detection of outdated species names.

Cugler's experiments were executed over a total of 11898 records, with 1929 distinct species names analyzed. As shown in Figure 4.2, 134 distinct species in the collection (7% of the species analyzed) had their scientific names changed along time.

Listing 4.2: Excerpt from Taverna's workflow description file

```
1
2  <processor>
3  <name>Catalog_of_life</name>
4  <annotations>
5   ...
6   <text>
7     Q(reputation): 1;
8     Q(availability): 0.9;
9   </text>
10        </annotationBean>
11        <date>2013-11-12 19:58:09.767 UTC</date>
```

```
      <creators />
      <curationEventList />
    </net.sf.taverna.t2.annotation.AnnotationAssertionImpl>
  </annotationAssertions>
</annotations>
```

4.3.2 Quality assessment using the preliminary QualityFlow architecture

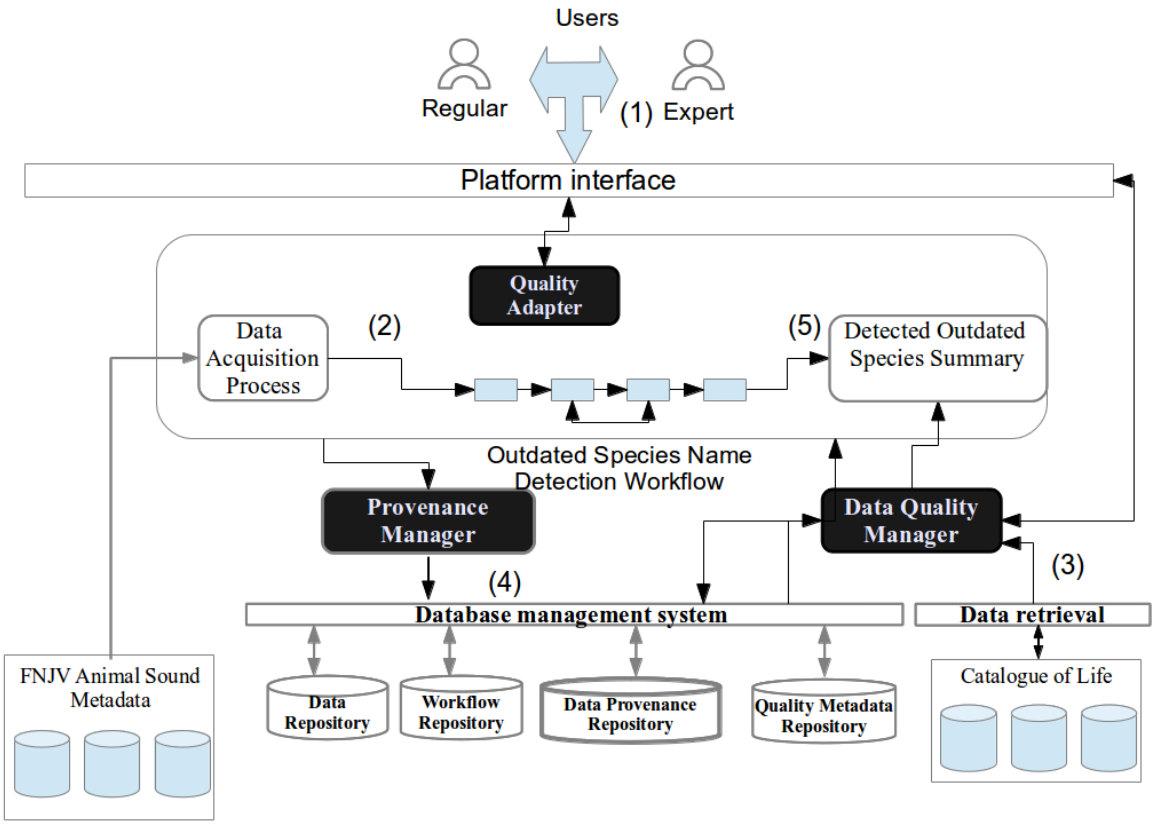


Figure 4.3: Architecture instance for the experiment

Figure 4.3 shows the initial version of QualityFlow used for this experiment.

The metadata curation process follows these steps:

1. Experts access QualityFlow to add quality metadata to the workflow, via the Quality Adapter module. They also define the Accuracy quality metrics via the Quality Manager.

2. The Outdated Species Name Detection Workflow receives FNJV sound metadata as input, and checks for outdated names, using the Catalogue of Life external data source;

3. After the workflow is executed, the Provenance Manager stores provenance information from the data source, workflow description and execution logs;

4. The workflow output is a summary of updated species names (see Figure 4.2).

The workflow (composed of a simple process implemented in Java) was run using the Taverna [21] workflow management system. QualityFlow is used to insert quality annotations for process and data sources before the workflow is executed - via Quality Adapter. Listing 4.2 shows an excerpt of the annotated workflow specification, where, for instance, the reputation of the Catalogue of Life is 1 (maximum) and its availability is 0.9 (since there are several connection problems) - lines 7 and 8 of the listing.

Taverna exports provenance information using the OPM (Open Provenance Model) model [30]. The Provenance Manager merges this information with Taverna's annotated workflow, and maps the result into the Provenance Repository.

The Data Quality Manager is used to define a quality metrics to compute the accuracy of species name metadata, defined as a percentage of correct names. Moreover, it outputs the availability and reputation of the Catalogue of Life. As a result, the end user can see that the original FNJV metadata, compared with an external authoritative source (reputation 1, availability 0.9) is 93% accurate. These results are shown to users, helping them to better understand their data. The Quality Manager accesses both the Provenance Repository and workflow output to provide such quality dimensions.

4.4 Experiment 2 - Get Weather Information Workflow

Get Weather Information (GWI) [17] is a freely available workflow in myExperiment website [3]. This workflow was used to validate the model and features of QualityFlow prototype.

4.4.1 Description of the experiment

GWI is composed by a few processes that are easy to understand and so facilitates to show QualityFlow usage in detail. Figure 4.4 shows the design of this workflow. It takes a country as input, gets a list of cities of the country, then queries and returns the weather for some of these cities. This section describes an experiment that uses QualityFlow to manage quality information in the usage of this workflow for a experiment.

Assume that Ana - an expert user in QualityFlow - is a researcher that designed and published GWI to allow any interested scientist to use it in order to have a collaborative quality analysis of the workflow. Bruno is a regular user that decided to use GWI for his own study. Carol and Diego are respectively expert and regular users that also use GWI.

4.4.2 Interaction with QualityFlow

Ana starts by registering GWI in QualityFlow. She uploads the workflow description file and fills the required fields. Figure 4.5 shows the workflow insertion screen. Besides publishing GWI, she also adds the following Quality Dimensions: Consistency, Freshness and Coverage, as shown in Figure 4.6. The user interface handles the insertion of the dimensions to the Quality Adapter. Notice she only defines the dimensions but not the values.

Bruno retrieves the GWI specification from QualityFlow and runs it a few times, registering in the system the associated DataResults and TraceLogs - e.g., Figure 4.7. The figure shows one of his runs concerns Ana's workflow, GWI. He associates some values for the quality dimensions defined by Ana (Figure 4.8). The figure shows, for instance, that he associated distinct coverages for each result.

Afterwards, Carol annotates the process GetCitiesByCountry with {Coverage = 1.0}. She also adds the quality metrics Weather Coverage - which defines that the coverage of GWI is equal to the coverage of GetCitiesByCountry.

Finally, when Diego wants to check out GWI quality information, he can observe the Quality Report generated by Quality Manager, as shown in Figure 4.9. In this report, there is quality information directly assigned to the workflow - e.g., the Freshness dimension and value - and also assigned to its results from distinct executions.

4.5 Limitations

The main limitation of this implementation is the possibility of automatic execution of workflows via QualityFlow. The platform should be able to execute the workflow specification directly, and automatically store results and provenance records. A naive implementation of this feature is straightforward. However, it would require significant development

to provide the user interaction with the WFMS - for instance, different kinds of errors may occur and for each kind of error different actions must be provided to the user: abort, ignore and continue, report and continue, and so on. Thus, in the present version, users have to execute the workflow locally and update results and provenance logs manually.

Another limitation is the capability of parsing both workflow specification files and provenance files. In workflow description files, it is not trivial to separate core processes from complementary processes like iterations or retries. The automatic parsing of these files results in many "artificial" processes that have no real application execution semantics. Since our main goal is to support quality assessment we have manually removed irrelevant processes and data sources. The alternative would be to design selective parsers to choose meaningful information to store. This is left for future work.

4.6 Summing up

QualityFlow provides an accessible platform for sharing eScience experiments with data quality and provenance. Although it has some limitations, the main concepts behind it are held in the implementation, which provides a first effort to face the problem of data quality in the WFMS environment. In this chapter, we have presented two experiments performed during our work. The first one use a real problem to illustrate the benefits of this proposal; it made a limited use of QualityFlow, as it was performed when the platform was still under development. The second one - using a hypothetical problem with a real workflow - was able to showcase the majority of the concepts and features that composes QualityFlow.

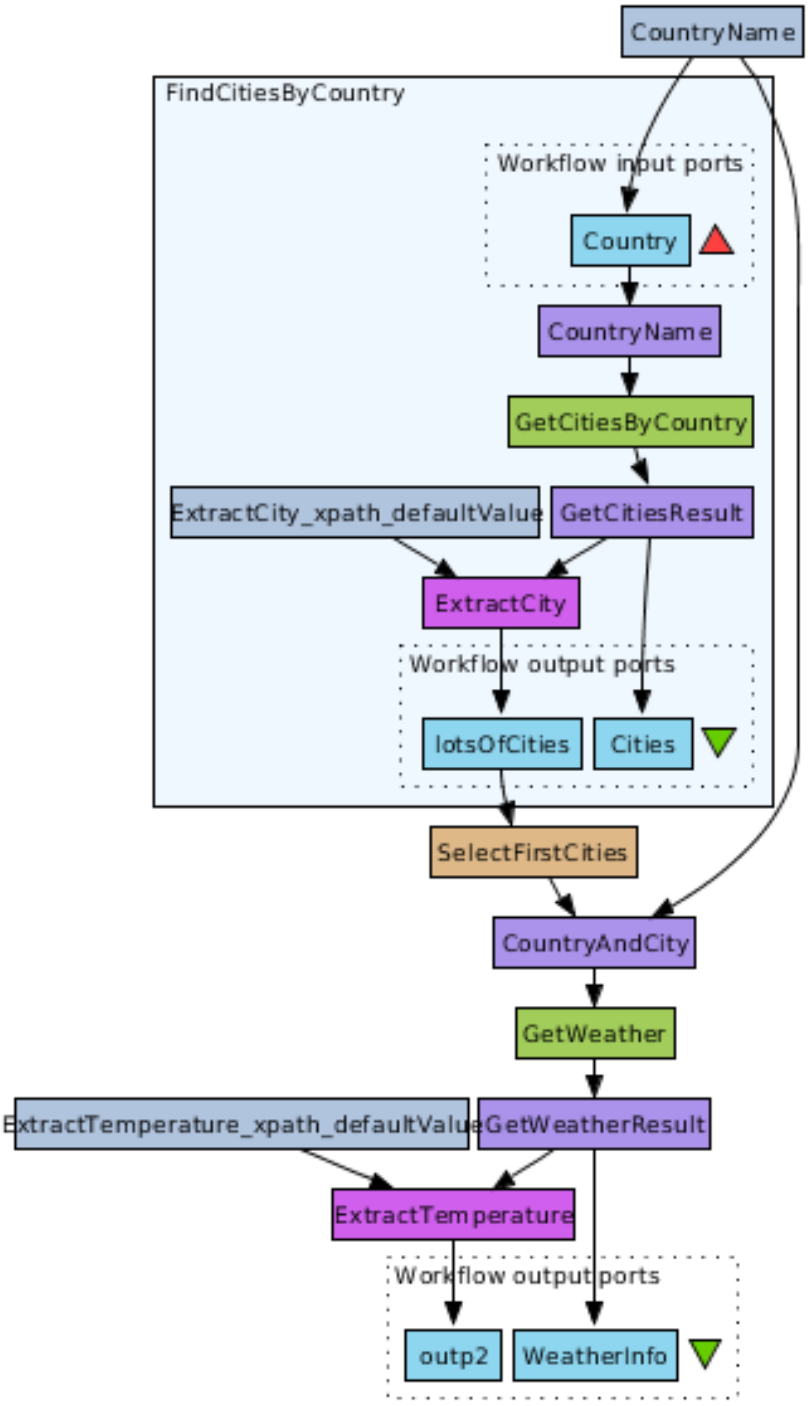


Figure 4.4: Get Weather Information Workflow

Figure 4.5: GWI added by Ana in QualityFlow insertion screen

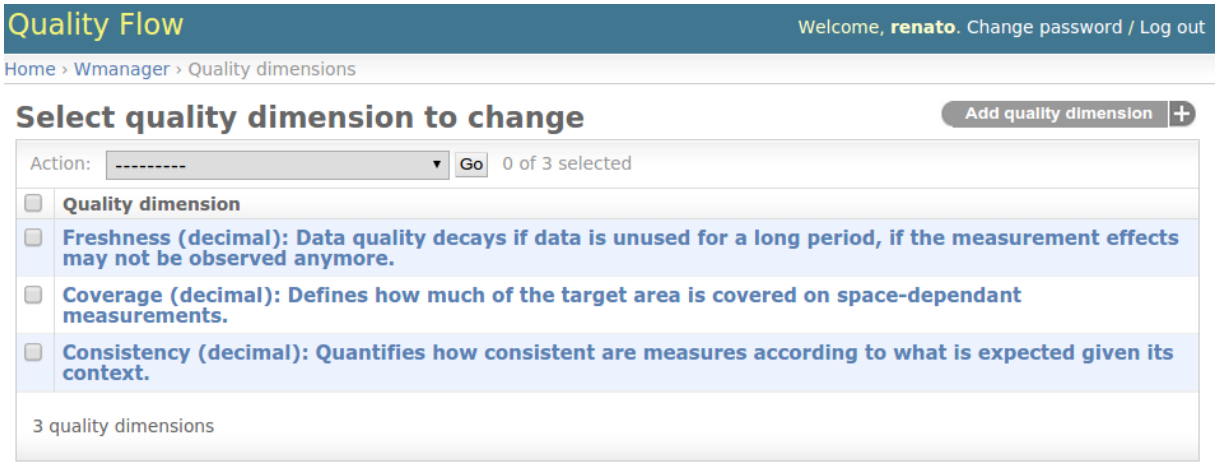


Figure 4.6: Quality Dimensions added by Ana via QualityFlow

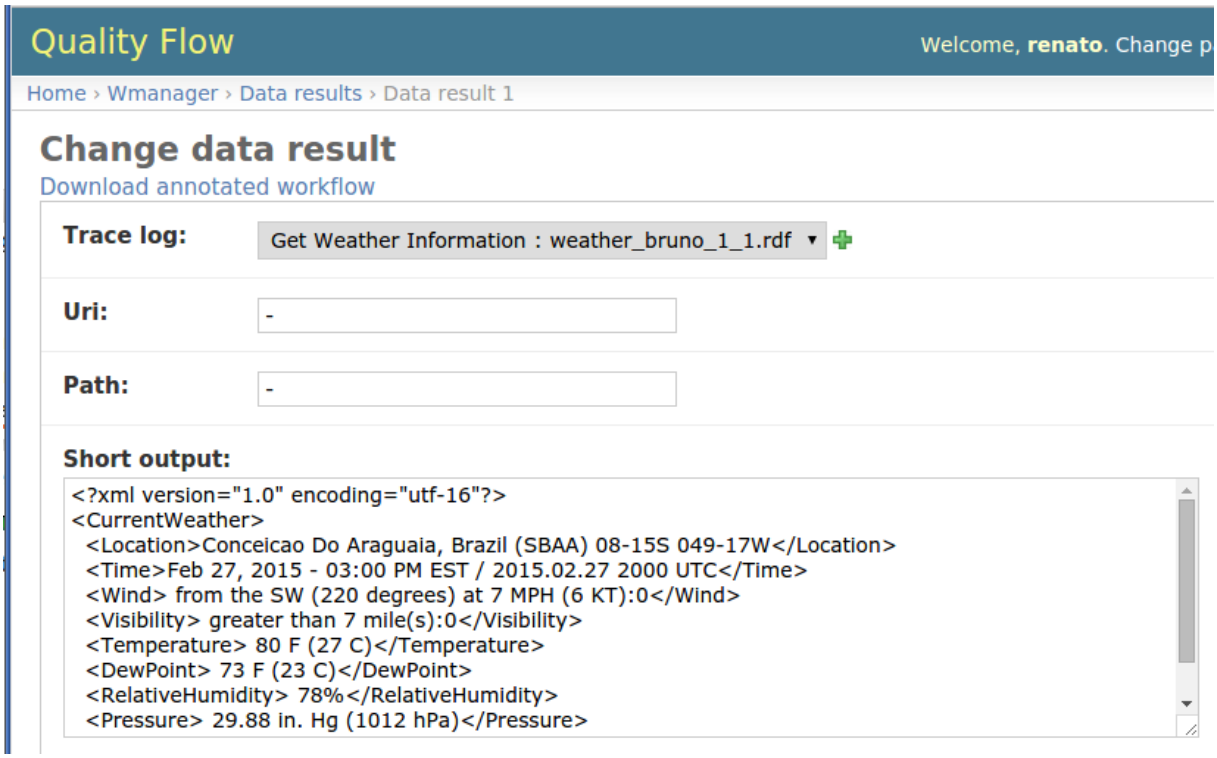


Figure 4.7: Data result added by Bruno in QualityFlow

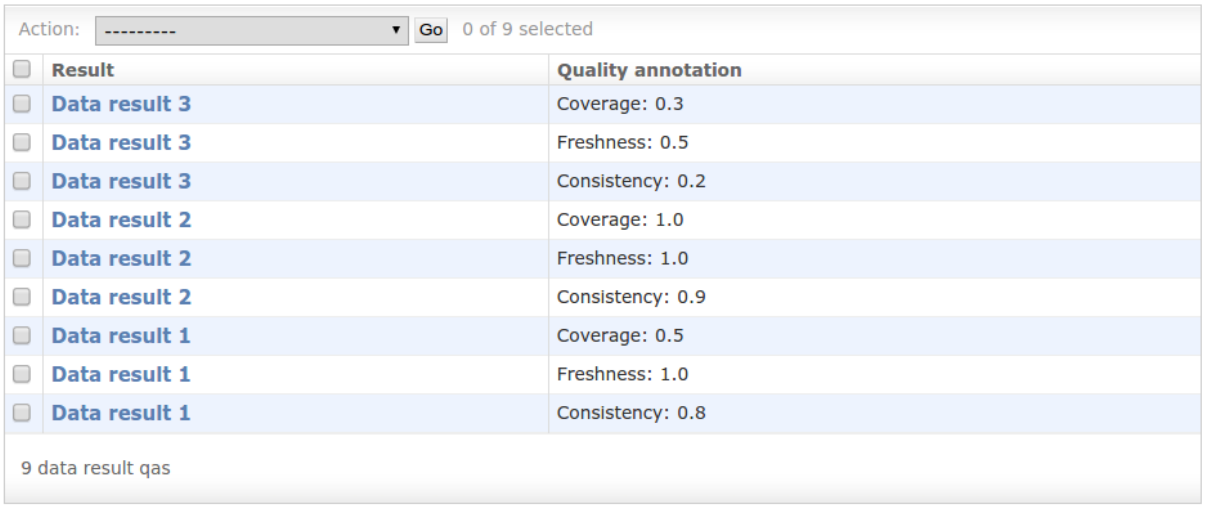


Figure 4.8: Quality Annotations added by Bruno in QualityFlow

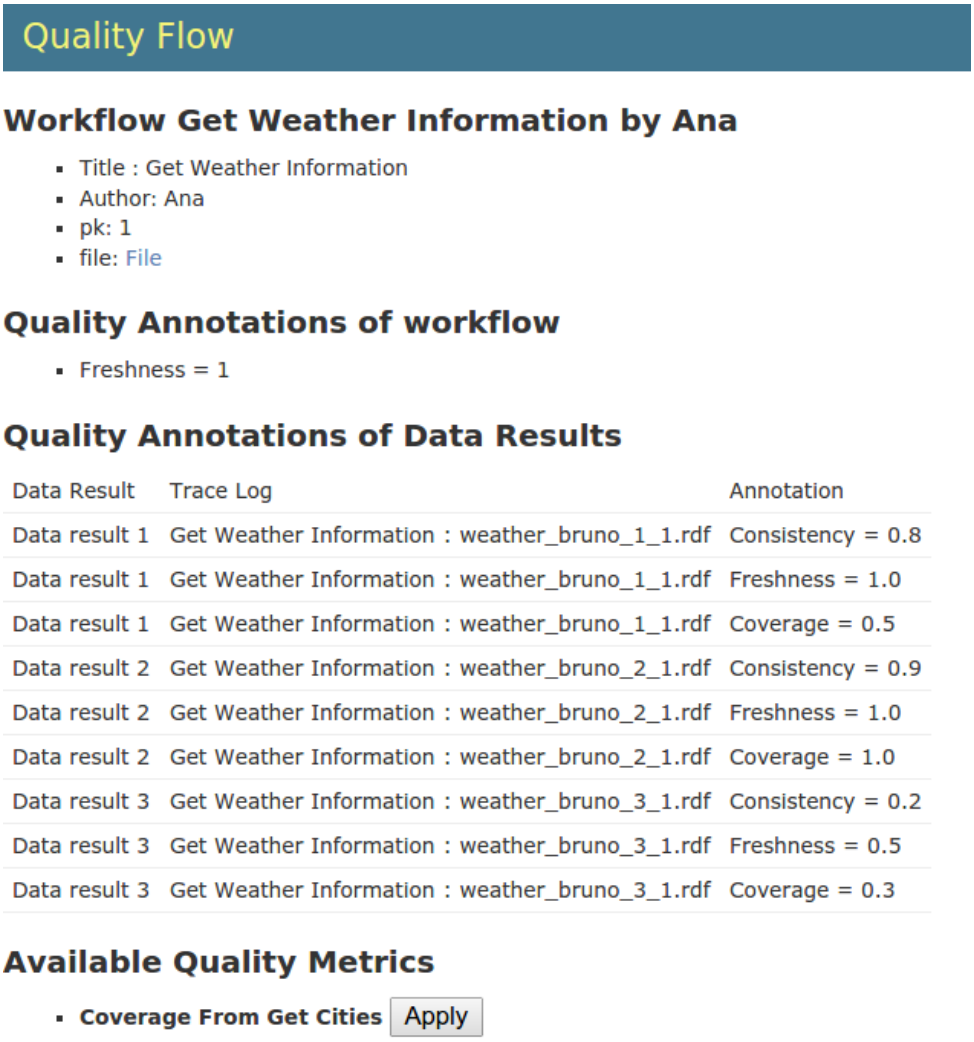


Figure 4.9: Quality Summary observed by Diego

Chapter 5

Conclusions and extensions

Our main problem is to provide means to assess data quality, given that current mechanisms are not able to materialize the concept of fitness for use. Considering this problem, our main goals are: to propose a data quality assessment mechanism, that supports flexible and multifaceted data quality analysis and that is able to generate quality information from data provenance.

5.1 Conclusions

The main contribution of this dissertation is the proposal of QualityFlow: a quality-aware collaborative platform to manage quality assessment for eScience applications. QualityFlow is based on Malaverri's work [28], a provenance-based approach for data-quality assessment. Although the Quality Manager is part of Malaverri's proposed framework, its architecture was not specified. Her thesis is centered on the Provenance Manager, and the Quality Manager was indicated as a component of the architecture, given its importance for better quality assessment. Our work specifies and provides a prototype of the Quality Manager, extends the Provenance Manager, and adds the Quality Adapter modules, thus providing useful functionaliy for filling such gaps.

In particular, more specific contributions of QualityFlow are:

- to support the creation of quality-aware scientific workflows, allowing users to add quality information to workflow specifications;
- to allow scientists to customize data quality dimensions and metrics collaboratively, so that the result of running a given workflow can have distinct assessments, depending on the user;

- to derive data quality information using a combination of provenance records and attributes defined by scientists;
- to support these contributions via the implementation of QualityFlow platform.

5.2 Extensions

There are many possible extensions to this dissertation. Among them, we point out the following:

- Add support to context-adaptative workflows to take advantage of quality information
- Develop an algorithm to suggest the most relevant processes of a workflow to receive quality annotations.
- Create or use an existing formal language for the definition of quality metrics
- Calculate a quality index of users (reputation) according to the quality values attributes to their artifacts by other users.
- Create or extend an existing ontology to maintain a canonical repository of quality dimensions
- Create plugins for the most used WFMSs to use the QualityFlow features directly
- Implement a seletive parser for TraceLog files to allow automatic storage of the relevant steps of workflow execution. The direct implementation generates irrelevant steps like iterations.

Bibliography

[1] Catalogue of life. http://www.catalogueoflife.org, accessed in October 2013.

[2] Django project. https://www.djangoproject.com/, accessed in March 2015.

[3] My experiment workflow repository. http://www.myexperiment.org/, accessed in October 2014.

[4] Python. https://www.python.org/, accessed in March 2015.

[5] Taiseera Hazeem Al Balushi, Pedro R Falcone Sampaio, and Pericles Loucopoulos. Eliciting and prioritizing quality requirements supported by ontologies: a case study using the elicito framework and tool. *Expert Systems*, 30(2):129–151, 2013.

[6] Derik Barseghian, Ilkay Altintas, Matthew B Jones, Daniel Crawl, Nathan Potter, James Gallagher, Peter Cornillon, Mark Schildhauer, Elizabeth T Borer, Eric W Seabloom, et al. Workflows and extensions to the kepler scientific workflow system to support environmental sensor data access and analysis. *Ecological Informatics*, 5(1):42–50, 2010.

[7] C. Batini, M. Lenzerini, and S.B. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys (CSUR)*, 18(4):323–364, 1986.

[8] Shawn Bowers. Scientific workflow, provenance, and data modeling challenges and approaches. *Journal on Data Semantics*, 1(1):19–30, 2012.

[9] E. Bertino C. Dai, D. Lin and M. Kantarcioglu. An approach to evaluate data trustworthiness based on data provenance. In *Proceedings of the 5th VLDB Workshop on Secure Data Management*, pages 82–98. Springer Berlin / Heidelberg, 2008.

[10] A.D. Chapman. Principles and methods of data cleaning–primary species and species-occurrence data. *Report for the Global Biodiversity Information Facility*, (version 1.0), 2005.

[11] Daniel Cugler. *Supporting Management of Biological Observation Databases*. PhD thesis, Instituto de Computação - Universidade Estadual de Campinas, 2014.

[12] Daniel Cintra Cugler, Claudia Bauzer Medeiros, and Luís Felipe Toledo. An architecture for retrieval of animal sound recordings based on context variables. *Concurrency and Computation: Practice and Experience*, 2012.

[13] Shashi Shekhar Daniel Cintra Cugler, Claudia Bauzer Medeiros and Luís Felipe Toledo. A geographical approach for metadata quality improvement in biological observation databases. In *9th IEEE International Conference on e-Science*. IEEE, 2013.

[14] Alexandre Carvalho de Alencar. Qualidade de dados em aplicações geográficas. Master's thesis, Instituto de Computação - Unicamp, March 2000.

[15] Lorena Etcheverry, Verónika Peralta, and Mokrane Bouzeghoub. Qbox-foundation: a metadata platform for quality measurement. In *proceeding of the 4th Workshop on Data and Knowledge Quality (QDC'2008)*, 2008.

[16] Tony Fountain, Sameer Tilak, Paul Hubbard, Peter Shin, and Lawrence Freudinger. The open source dataturbine initiative: Streaming data middleware for environmental observing systems. In *International Symposium on Remote Sensing of Environment*, 2009.

[17] Franch Tanoh. Get weather information. http://www.myexperiment.org/workflows/242.html, accessed in January 2015.

[18] Matthew Gamble and Carole Goble. Quality, trust, and utility of scientific data on the web: Towards a joint model. *Web Science Trust*, 2011.

[19] Tony Hey and Anne E Trefethen. The uk e-science core programme and the grid. *Future Generation Computer Systems*, 18(8):1017–1031, 2002.

[20] Sonja Holl, Olav Zimmermann, and Martin Hofmann-Apitius. A new optimization phase for scientific workflow management systems. In *E-Science (e-Science), 2012 IEEE 8th International Conference on*, pages 1–8. IEEE, 2012.

[21] Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R Pocock, Peter Li, and Tom Oinn. Taverna: A tool for building and running workflows of services. *Nucleic acids research*, 34(suppl 2):W729–W732, 2006.

[22] Rubens Camargo Lamparelli Joana E. G. Malaverri, Claudia Bauzer Medeiros. A provenance approach to assess the quality of geospatial data. In *Applied Computing, 27th Symposium on*, mar. 2012.

[23] Gilberto Zonta Pastorello Jr. *Managing the lifecycle of sensor data: from production to consumption.* PhD thesis, Universidade Estadual de Campinas, Campinas, São Paulo, 2008.

[24] Zoé Lacroix, CRL Legendre, and S Tuzmen. Reasoning on scientific workflows. In *Services-I, 2009 World Conference on*, pages 306–313. IEEE, 2009.

[25] Fernando Lemos. *Infrastructure and Algorithms for Information Quality Analysis and Process Discovery.* PhD thesis, Ingénierie des Systèmes d'Information, 2013.

[26] João Guilherme Souza Lima. Gerenciamento de dados climatológicos heterogêneos para aplicações em agricultura. Master's thesis, Universidade Estadual de Campinas, Campinas, São Paulo, Outubro 2003.

[27] Bertram Ludäscher, Ilkay Altintas, Chad Berkley, Dan Higgins, Efrat Jaeger, Matthew Jones, Edward A Lee, Jing Tao, and Yang Zhao. Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience*, 18(10):1039–1065, 2006.

[28] Joana Esther Gonzales Malaverri. *Supporting data quality assessment in eScience: a provenance based aproach.* PhD thesis, Universidade Estadual de Campinas, Campinas, São Paulo, 2013.

[29] T. Milo and S. Zohar. Using schema matching to simplify heterogeneous data translation. In *Proceedings of the International Conference on Very Large Data Bases*, pages 122–133. Citeseer, 1998.

[30] Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, et al. The open provenance model core specification (v1. 1). *Future Generation Computer Systems*, 27(6):743–756, 2011.

[31] Aisa Na'im, Daniel Crawl, Maria Indrawan, Ilkay Altintas, and Shulei Sun. Monitoring data quality in kepler. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pages 560–564. ACM, 2010.

[32] Tom Oinn, Mark Greenwood, Matthew Addis, M Nedim Alpdemir, Justin Ferris, Kevin Glover, Carole Goble, Antoon Goderis, Duncan Hull, Darren Marvin, et al.

Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, 2006.

[33] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[34] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. Germany, 2000.

[35] Michael Reiter, Uwe Breitenbucher, Oliver Kopp, and Dimka Karastoyanova. Quality of data driven simulation workflows. In *E-Science (e-Science), 2012 IEEE 8th International Conference on*, pages 1–8. IEEE, 2012.

[36] Renato Beserra Sousa, Daniel Cintra Cugler, Joana Esther Gonzales Malaverri, and Claudia Bauzer Medeiros. A provenance-based approach to manage long term preservation of scientific data. In *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*, pages 162–133. IEEE, 2014.

[37] W3C. The prov ontology, 2012.

[38] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, pages 5–33, 1996.

[39] Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher, et al. The taverna workflow suite: Designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic acids research*, 2013.

[40] Peng Yue and Lianlian He. Geospatial data provenance in cyberinfrastructure. In *Geoinformatics, 2009 17th International Conference on*, pages 1–4, aug. 2009.