Allan Koch Veiga

A conceptual framework on biodiversity data quality

São Paulo
2017

Allan Koch Veiga

A conceptual framework on biodiversity data quality

Thesis presented to Escola Politécnica da Universidade de São Paulo to obtain the title of Doctor of Sciences

Concentration area:
Computer Engineering

Supervisor: Antonio Mauro Saraiva

São Paulo
2017

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Veiga, Allan Koch
A conceptual framework on biodiversity data quality / A. K. Veiga -- versão corr. -- São Paulo, 2017.
156 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Frameworks 2.Informática-Biodiversidade 3.Ciência da informação 4.Administração da qualidade 5.Qualidade de dados I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Dedico este trabalho aos meus pais Leila Adriana Koch e Jonas Marques Veiga,
à minha madrasta Miriam Mardegan, a meu irmão Felipe Koch Veiga
e à minha querida noiva Heloisa Rosa.

Acknowledgements

Agradeço a Deus, primeiramente, pois "O coração do homem planeja o seu caminho, mas o Senhor lhe dirige os passos"(Provérbios 16:9).

Agradeço especialmente, também, a todos que contribuíram diretamente e indiretamente para o conclusão deste trabalho, incluindo:

Minha família, a base fundamental da minha formação. Agradeço meu pai, Jonas Marques Veiga e minha mãe, Leila Adriana Koch, pela educação construída com muita disciplina, bons exemplos e umas boas chineladas; minha madrasta, Miriam Mardegan, por todo o incentivo e boa educação que me ofereceu durante todos esses anos; meu irmão, Felipe Koch Veiga, por ser um exemplo para mim e parceiro em todos os momentos.

Minha namorada, Heloisa Rosa, por me suportar e me ajudar a passar pelos momentos mais estressantes, tensos e difíceis com muita risada, companheirismo e carinho.

Meu orientador, prof. Antonio Mauro Saraiva, que deu todo o apoio e suporte necessários para o bom andamento desta pesquisa, além de ser um amigo e mentor, que contribuiu efetivamente no meu crescimento acadêmico, profissional e pessoal.

Meus tutores na Universidade de Harvard, doutores Paul Morris, Bob Morris e James Hanken, por terem sido excelentes anfitriões e ótimos ouvintes e críticos do meu trabalho. Contribuíram muito no amadurecimento desta tese e no enriquecimento de ideias relacionadas a ela.

Parceiros, apoiadores e críticos deste trabalho, Arthur Chapman, Dmitry Schigel, Tim Robertson, Christian Gendreau, que contribuíram imensamente para o desenvolvimento desta pesquisa.

Meus amigos do Laboratório de Automação Agrícola. O Etienne Cartolano, por ser uma referência para mim de profissional e de pessoa e por sempre ter me ajudado em questões de qualquer natureza; Wilian França, Michel Bieleveld, Raul Teruel, Daniel Lins, Edson de Souza, Suzano Bitencourt, Ivairton Santos e João Ferreira, por inúmeras conversas e discussões sobre a pesquisa (e outros assuntos de bar) - parceiros nessa vida bandida. Lourdes Keico por todo apoio e suporte na secretaria do laboratório e pelos lanchinhos da tarde. O convívio e a amizade com vocês foi muito bom e importante.

O BioComp, por todo o suporte e financiamento de equipamentos e viagens, que foram essenciais para o desenvolvimento deste trabalho; a CAPES, pelo financiamento da pesquisa com uma bolsa de doutorado; o programa Ciência sem Fronteiras, por financiar o programa de doutorado sanduíche nos Estados Unidos; e ao Programa de Pós-Graduação em Engenharia Elétrica da Escola Politécnica da USP, por todo o apoio administrativo e auxílios para participação de eventos.

"If I have seen further, it is by standing on the shoulders of giants."
(Isaac Newton, 1676)

Abstract

The increasing availability of digitized biodiversity data worldwide, provided by an increasing number of sources, and the growing use of those data for a variety of purposes have raised concerns related to the "fitness for use" of such data and the impact of data quality (DQ) on outcomes of analyses, reports and decisions making. A consistent approach to assess and manage DQ is currently critical for biodiversity data users. However, achieving this goal has been particularly challenging because of the idiosyncrasies inherent to the concept of quality. DQ assessment and management cannot be suitably carried out if we have not clearly established the meaning of quality according to the data user's standpoint. This thesis presents a formal conceptual framework to support the Biodiversity Informatics (BI) community to consistently describe the meaning of data "fitness for use". Principles behind data fitness for use are used to establish a formal and common ground for the collaborative definition of DQ needs, solutions and reports useful for DQ assessment and management. Based on the study of the DQ domain and its contextualization in the BI domain, which involved discussions with experts in DQ and BI in an iterative process, a comprehensive framework was designed and formalized. The framework defines eight fundamental concepts and 21 derived concepts, organized into three classes: DQ Needs, DQ Solutions and DQ Report. The concepts of each class describe, respectively, the meaning of DQ in a given context, the methods and tools that can serve as solutions for meeting DQ needs, and reports that present the current status of quality of a data resource. The formalization of the framework was presented using conceptual maps notation and sets theory notation. In order to validate the framework, we present a proof of concept based on a case study conducted at the Museum of Comparative Zoology of Harvard University. The tools FP-Akka Kurator and the BDQ Toolkit were used in the case study to perform DQ measures, validations and improvements in a dataset of the Arizona State University Hasbrouck Insect Collection. The results illustrate how the framework enables data users to assess and manage DQ of datasets and single records using quality control and quality assurance approaches. The proof of concept has also shown that the framework is adequately formalized and flexible, and sufficiently complete for defining DQ needs, solutions and reports in the BI domain. The framework is able of formalizing human thinking into well-defined components to make it possible sharing and reusing definitions of DQ in different scenarios, describing and finding DQ tools and services, and communicating the current status of quality of data in a standardized format among the stakeholders. In addition, the framework supports the players of that community to join efforts on the collaborative gathering and developing of the necessary components for the DQ assessment and management in different contexts. The framework is also the foundation of a Task Group on Data Quality, under the auspices of the Biodiversity Information Standards (TDWG) and the Global Biodiversity Information Facility (GBIF) and is being used to help collect user's needs on data quality on agrobiodiversity and on species distributed modeling, initially. In future work, we plan to use the framework to engage the BI community to formalize and share DQ profiles related to a number of other data usages, to recommend methods, guidelines, protocols,

metadata schemas and controlled vocabulary for supporting data fitness for use assessment and management in distributed system and data environments. In addition, we plan to build a platform based on the framework to serve as a common backbone for registering and retrieving DQ concepts, such as DQ profiles, methods, tools and reports.

Keywords: data quality. information quality. fitness for use. biodiversity informatics. conceptual framework. quality assurance. quality control. quality assessment.

Resumo

A crescente disponibilização de dados digitalizados sobre a biodiversidade em todo o mundo, fornecidos por um crescente número de fontes, e o aumento da utilização desses dados para uma variedade de propósitos, tem gerado preocupações relacionadas a "adequação ao uso" desses dados e ao impacto da qualidade de dados (QD) sobre resultados de análises, relatórios e tomada de decisões. Uma abordagem consistente para avaliar e gerenciar a QD é atualmente crítica para usuários de dados sobre a biodiversidade. No entanto, atingir esse objetivo tem sido particularmente desafiador devido à idiossincrasia inerente ao conceito de qualidade. A avaliação e a gestão da QD não podem ser adequadamente realizadas sem definir claramente o significado de qualidade de acordo com o ponto de vista do usuário dos dados. Esta tese apresenta um arcabouço conceitual formal para apoiar a comunidade de Informática para Biodiversidade (IB) a descrever consistentemente o significado de "adequação ao uso" de dados. Princípios relacionados à adequação ao uso são usados para estabelecer uma base formal e comum para a definição colaborativa de necessidades, soluções e relatórios de QD úteis para a avaliação e gestão de QD. Baseado no estudo do domínio de QD e sua contextualização no domínio de IB, que envolveu discussões com especialistas em QD e IB em um processo iterativo, foi projetado e formalizado um arcabouço conceitual abrangente. Ele define oito conceitos fundamentais e vinte e um conceitos derivados organizados em três classes: Necessidades de QD, Soluções de QD e Relatório de QD. Os conceitos de cada classe descrevem, respectivamente, o significado de QD em um dado contexto, métodos e ferramentas que podem servir como soluções para atender necessidades de QD, e relatórios que apresentam o estado atual da qualidade de um recurso de dado. A formalização do arcabouço foi apresentada usando notação de mapas conceituais e notação de teoria dos conjuntos. Para a validação do arcabouço, nós apresentamos uma prova de conceito baseada em um estudo de caso conduzido no Museu de Zoologia Comparativa da Universidade de Harvard. As ferramentas FP-Akka Kurator e BDQ Toolkit foram usadas no estudo de caso para realizar medidas, validações e melhorias da QD em um conjunto de dados da Coleção de Insetos Hasbrouck da Universidade do Estado do Arizona. Os resultados ilustram como o arcabouço permite a usuários de dados avaliarem e gerenciarem a QD de conjunto de dados e registros isolados usando as abordagens de controle de qualidade a garantia de qualidade. A prova de conceito demonstrou que o arcabouço é adequadamente formalizado e flexível, e suficientemente completo para definir necessidades, soluções e relatórios de QD no domínio da IB. O arcabouço é capaz de formalizar o pensamento humano em componentes bem definidos para fazer possível compartilhar e reutilizar definições de QD em diferentes cenários, descrever e encontrar ferramentas de QD e comunicar o estado atual da qualidade dos dados em um formato padronizado entre as partes interessadas da comunidade de IB. Além disso, o arcabouço apoia atores da comunidade de IB a unirem esforços na identificação e desenvolvimento colaborativo de componentes necessários para a avaliação e gestão da QD. O arcabouço é também o fundamento de um Grupos de Trabalho em Qualidade de Dados, sob os auspícios do Biodiversity Information Standard (TDWG) e do Biodiversity Information Facility (GBIF) e está sendo utilizado para

coletar as necessidades de qualidade de dados de usuários de dados de agrobiodiversidade e de modelagem de distribuição de espécies, inicialmente. Em trabalhos futuros, planejamos usar o arcabouço apresentado para engajar a comunidade de IB para formalizar e compartilhar perfis de QD relacionados a inúmeros outros usos de dados, recomendar métodos, diretrizes, protocolos, esquemas de metadados e vocabulários controlados para apoiar a avaliação e gestão da adequação ao uso de dados em ambiente de sistemas e dados distribuídos. Além disso, nós planejamos construir uma plataforma baseada no arcabouço para servir como uma central integrada comum para o registro e recuperação de conceitos de QD, tais como perfis, métodos, ferramentas e relatórios de QD.

Palavras-Chave: qualidade de dados. qualidade de informação. adequação ao uso. informática para biodiversidade. arcabouço conceitual. garantia de qualidade. controlde de qualidade. avaliação da qualidade.

List of Figures

List of abbreviations and acronyms

ALA	Atlas of Living Australia
API	Application Programming Interface
BDD	Biodiversity Data Digitizer
BDQ	Biodiversity Data Quality
BI	Biodiversity Informatics
BioComp	Research Center on Biodiversity and Computing
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
DOI	Digital Object Identifier
DQ	Data Quality
EPUSP	Escola Politécnica da Universidade de São Paulo
FP	FilteredPush
GBIF	Global Biodiversity Information Facility
GBIO	Global Biodiversity Informatics Outlook
GUID	Global Unique Identifier
IE	Information Element
IG	Interest Group
LSID	Life Science Identifiers
MCZ	Museum of Comparative Zoology
MCZbase	The Database of the Zoological Collections of Harvard Museum of Comparative Zoology
PCS	Departamento de Engenharia de Computação e Sistemas
PPGEE	Programa de Pós Graduação em Engenharia Elétrica
SMC	Simple Matching Coefficient
TDQM	Total Data Quality Management
TDWG	Biodiversity Information Standards

TG	Task Group
TQM	Total Quality Management
URL	Uniform Resource Locator

Contents

1 Introduction

This chapter presents the research justification, problem statement, the main and specifics objectives, the method used to achieve the objectives of this research and the structure of this document.

1.1 Justification

The Biodiversity Informatics (BI) community has successfully supported initiatives to capture and digitize standardized biodiversity-related data and to deliver platforms for free access to biodiversity data integrated from many data providers distributed around the world.

The research field called BI, which aims at applying informatics concepts, techniques and tools to research and development on biodiversity, has concentrated much effort into the digitization of standardized biodiversity data, the integration of such data and on making those data available by means of digital platforms on the Internet for being used in a myriad of usages. Very few examples of biodiversity data usages are distribution modeling, agrobiodiversity, impact assessment of environment changes and biodiversity conservation.

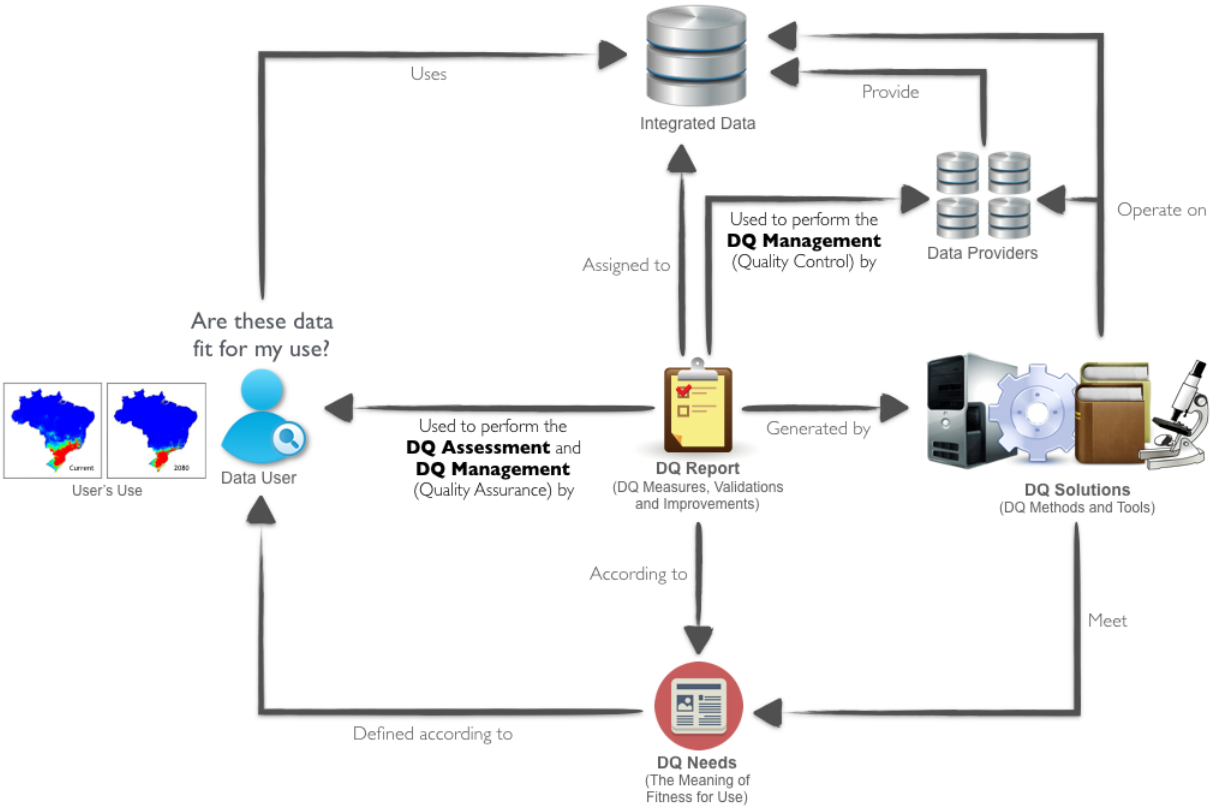
However, the increasing amount of freely available data from an also increasing amount of sources, which may have a different and unclear level of concern with the quality of their data, has risen interest related to the "fitness for use" of such data. Before using any data, the data users have to ask if the quality of data is fit for their particular uses, that is, to perform the **Data Quality (DQ) assessment**.

Performing DQ assessment became a critical issue in the BI context, specially because not enough information about the quality of data are provided. It makes difficult to select subsets of data that are fit for use for different specific purposes. Furthermore, the action of determining if data are fit for use is highly dependent on the "data use" and considering all the potential biodiversity data usages for dealing properly with DQ is impractical for the most the BI initiatives.

In this context, it is evident that any effort aiming at allowing DQ assessment necessarily requires determining what DQ needs means according to the data user's perspectives, as illustrated in the Figure 1. Due to the idiosyncratic nature of the concept of "quality", it is essential to understand what means "data fitness for use" according to the data user's perspective in order to enable DQ assessment.

Based on a well-defined user's DQ needs, DQ solutions must be delivered in order to meet the DQ needs. DQ solutions are everything able to perform DQ measures, validations or improvements in datasets or single records. The results obtained by DQ solutions must be reported to data users and their respective providers for assisting them to perform the DQ

Figure 1: Scenario and main components regarding DQ assessment and management.



Source: Author

assessment and management.

DQ reports, illustrated in Figure 1, are sets of DQ assertions assigned to a dataset or a single record, generated by DQ solutions according to users' DQ needs. The DQ report describes the current status of quality of a dataset or a single record according to the perspectives of data users. DQ reports contain DQ measures, validations and improvements (recommended or performed) that enable data users to perform an appropriate DQ assessment and, perhaps, the selection of a subset of the data from the original dataset that is fit for use, i.e. to perform the **DQ management by the DQ assurance approach**. A DQ report can also be used by data providers to improve their own data based on improvement recommendations or by just highlighting the current level of quality of their data, that is, to perform **DQ management by the DQ control approach**.

Scenario described above, illustrated in the Figure 1, demonstrates that three main components are necessary to enable the DQ assessment and management in a given context, for example in BI: user's DQ needs, DQ solutions that meet user's DQ needs and DQ reports that communicate the current status of DQ according to user's DQ needs.

Besides DQ assessment and management being a critical issue in the BI context, no formal model has been proposed yet to define and organize the required concepts for capturing and formalizing the user's DQ needs, describing shareable DQ solutions and for generating standardized DQ reports, focused on the biodiversity DQ assessment and management.

1.2 Problem Statement

It has become essential to tackle DQ in the current scenario of BI, a platform that biodiversity data users can collaboratively describe what means data with suitable quality for being used in specific tasks, where such requirements are linked to available solutions that generate standardized DQ measures, validations and improvements, or that highlights the lack of solutions for a specific requirement.

The core problem that this research tackle is the lack of conditions for data users to assess the quality of biodiversity data and determine if the data are fit for their uses. Determining how much quality a dataset that is available on the Internet has, and judging if such quality is suitable for the data to be used for a specific purpose, has become one of the most important and difficult challenges in the current scenario of BI.

Another central problem tackled in this research is DQ management guided by the user's DQ needs. Typically, in the BI community context, DQ management (quality improvement) is mostly driven by "the most common data errors" detection and correction. Highlighting common errors (e.g. invalid coordinates, misspelling errors, unknown scientific name) in data and trying to correct them are often the main methods used to deal with biodiversity DQ. That approach may improve the average quality, however, with no specific target to drive the efforts to make data fitter for a specific use, the applied efforts may generate scattered results that are not sufficient to tackle real DQ problems according to data user's perspective, i.e. no effort might be applied to real needs and much effort might be applied to solve nonexistent needs, in a given context.

In order to solve these two central problems, it is necessary to define a common ground to frame and guide the discussions toward the definition of consistent methods, protocols and standards which will support building up sharable and efficient tools for the DQ assessment and management of biodiversity data. In this context, the proposed conceptual framework aims to define a foundation and a common ground to serve as a guide to a proper DQ assessment and management in BI context.

1.3 Objective

To propose a comprehensive and formal conceptual framework that defines and organizes the necessary components to enable DQ assessment and management in the BI context and evaluate the conceptual framework with a study case in the Museum of Comparative Zoology

(MCZ) of Harvard University.

1.4 Materials and Method

The presented research problem statement does not allow us to clearly visualize a set of procedures to propose and test the hypothesis for achieving the research objective. Therefore it was used an exploratory approach, where a process of probing the problem's domain was recurrently applied in order to propose, test and improve potential valid hypothesis and ideas, always supported by bibliographic references of DQ and BI research areas and in close contact with domain experts.

The following four main steps define the method used to achieve the objectives of this research:

1. **Problem domain study**: Review of the literature regarding DQ, BI and other related areas, such as Computer Security, Reliability Engineering, Data Science, Statistical Methods, Statistical Ecological Modeling, Geospatial Modeling, Provenance, among others. Participation and organization of a number of conferences, symposiums, workshops, meetings, discussions, training, working groups, interchange and research projects on DQ and BI around the world for strengthening a more comprehensive, precise, and up-to-date understanding of the real international scenario of DQ in the BI domain. See Appendix C for more details. Furthermore, as an exploratory research, some activities for a deeper immersion in the problem domain were performed, such as countless meetings and discussions with experts of the international community, specially in the TDWG and GBIF scope, and leading one of the Task Groups of the TDWG/GBIF Biodiversity Data Quality. In addition, an interchange at the Harvard University was of a great value to discuss, test and improve the framework and related ideas, working together with experts in BI in the context of the MCZ and the Kurator project.

2. **Conceptual framework design**: To identify, organize and define the fundamental and derived concepts for enabling DQ assessment and management based on the problem domain study. The conceptual framework definition was guided by the objective of the research and supported by the problem domain study outcomes. During this phase, an iterative and incremental approach was used to design, test and improve the conceptual framework based on DQ principles and different scenarios found during the problem domain study. This process was performed until achieving a stable, robust and mature conceptual framework that reaches the objective of this research.

3. **Conceptual framework formalization**: To represent the conceptual framework using formal notations. Two formal notations were used to represents the conceptual framework into a mathematical and a visual representation. To represent mathematically the concepts and their relationships it was used a sets theory notation (DEVLIN, 1993), presented in

details in Appendix A. In addition, it was also used a conceptual maps notation to visually formalize human thinking into a machine-friendly representation (NOVAK; CAÑAS, 2008).

4. **Conceptual framework evaluation**: To evaluate the conceptual framework by means of a study case in the MCZ of Harvard University with the participation of experts on BI and DQ. The study case was conducted in order to evaluate if the components of the conceptual framework were complete and consistent to enable the DQ assessment and management in the MCZ of Harvard University, which is the institution responsible for managing a database about important zoological collections, the MCZbase. In addition, indirect evaluations were performed by international players of the BI community by means of several meetings, workshops and discussions during the period of this research.

1.5 Thesis Structure

This document is organized into six chapters:

- **Chapter 2** presents an introduction to DQ and BI researches based on the review of the literature, in order to highlight some relevant facts regarding the problem domain of this research;
- **Chapter 3** describes the fundamental and derived concepts of the conceptual framework and represents them using conceptual maps;
- **Chapter 4** presents and discusses the study case used to evaluate the conceptual framework;
- **Chapter 5** discusses the results of this research;
- **Chapter 6** concludes with final remarks, contributions and next steps.

In addition, three appendixes complement this thesis:

- **Appendix A - Mathematical Formalization**: Mathematical formalization of the conceptual framework using a sets theory notation;
- **Appendix B - Study Case (Conceptual Framework Evaluation)**: Detailed results of the case study performed in the MCZ of Harvard University;
- **Appendix C - List of Activities Related to the Problem Domain Study**: A short list of activities that helped to better understand the problem domain.

2 Topics on Data Quality and Biodiversity Informatics

This Chapter aims to present relevant facts regarding researches on DQ and BI in order to contextualize the conceptual framework in those research areas.

2.1 DQ Contextualization

DQ is a subject that permeates most research fields where data are an important asset for answering questions. Therefore, related researches on DQ, information quality or data fitness for use have been conducted and applied in a number of domains and under several aspects and approaches.

In this section, some comprehensive concepts, approaches and methods regarding DQ are presented to contextualize this research in the DQ research area.

2.1.1 DQ definitions

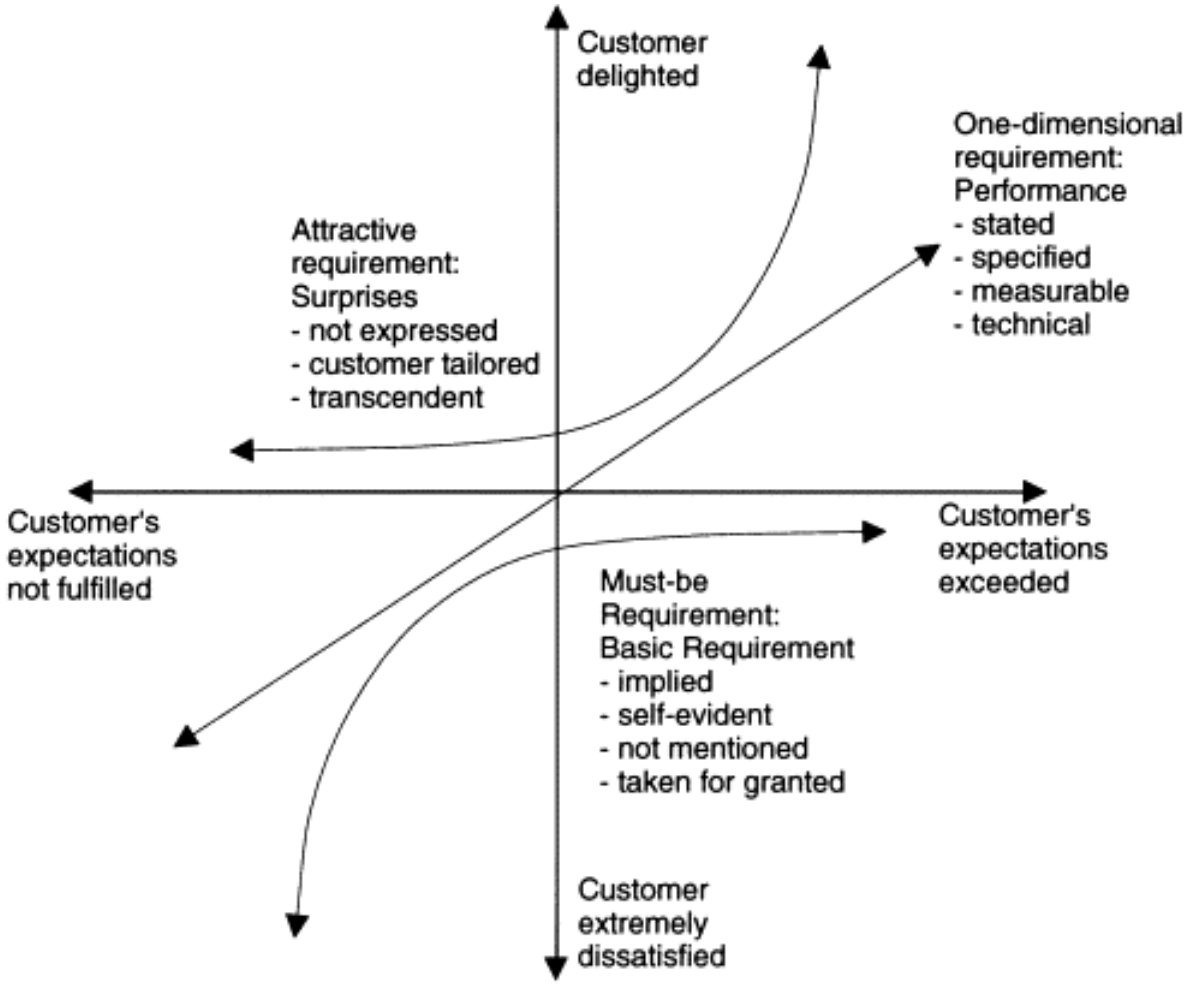
Quality is widely defined in literature as *requirement compliance* (CROSBY, 1995 apud WANG; KON; MADNICK, 1993). To define the meaning of quality, it is essential to deeply understand the true consumer's needs and, in addition, find out extra requirements that are perceived as necessary or desired by consumers but were not previously mentioned by them (ROSE, 1994).

Another related definition of quality has to do with *consumers satisfaction*. Whether a consumer is satisfied with a service or a product, this service or product has the quality to this consumer. Defining the meaning of quality according to these definitions means gathering and deeply understanding a set of requirements that affect consumers satisfaction. As illustrated in Figure 2, the Kano Requirements Model can be used to capture consumers requirements in order to maximize their satisfaction (MAZUR, 1993; BOLT; MAZUR, 1999)

Kano Requirements Model defines three types of requirements, as described next:

- **Normal Requirements:** These are stated requirements. The consumers' satisfaction increases when this type of requirements is met and it decreases when this type of requirements is not met.
- **Expected Requirements:** These are basic requirements which are normally not stated by consumers because of their triviality. The consumers' satisfaction remains unchanged when this type of requirements is met, however when it is not met, consumers satisfaction is strongly decreased.

Figure 2: Kano's model of customer satisfaction.



Source: (SAUERWEIN et al.,)

- **Exciting Requirements:** These are unstated and unexpected requirements which delight consumers. It is more difficult to find out this type of requirements, but when they are met, the satisfaction of consumers is strongly increased, but when they are not met, the satisfaction of consumers remains unchanged.

In order to increase consumers satisfaction, Kano Requirements Model can be used for gathering the important requirements that will define a product or a service with quality according to consumers expectations. Accordingly, that definition of quality, usually applied to product or services, can also be applied to data or information (WANG; KON; MADNICK, 1993).

The framework *Total Data Quality Management* (TDMQ) supports the thesis that information must be managed as a product (WANG; STRONG, 1996; WANG et al., 1998; WANG, 1998; WANG et al., 2003). Table 1 compares the manufacturing process of *product* and *information*.

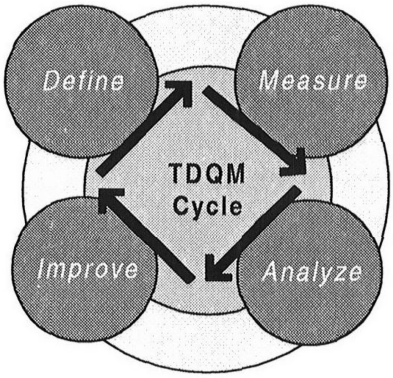
Table 1: Product vs. Information manufacturing process

	Product	**Information**
Input:	Raw Material	Raw Data
Process:	Assembly Line	Information System
Output:	Physical Products	Information Products

Source: (WANG; ZIAD; LEE, 2001)

Considering the premise that exists an analogy between quality issues in product manufacturing and those in information manufacturing, TDQM was designed based on the principles, guidelines and techniques for product quality of the Total Quality Management (TQM) (WANG; ZIAD; LEE, 2001). Based on the widely practiced Deming cycle for quality enhancement (Plan, Do, Check and Act), TDQM defines an analog cyclical process to *define* important DQ requirements, *measure* the quality according to the defined requirements, *analyse* the impact of poor quality and its root causes and *improve* quality of data in order to meet DQ requirements, resulting in better quality measures. The TDQM Cycle is illustrated with the Figure 3.

Figure 3: TDQM Cycle.



Source: (ALABRI, 2010)

Another approach for defining DQ is founded on the capacity of the information to adequately represent conditions of the real world, that is, *compliance with the real world* (WANG; KON; MADNICK, 1993). The quality of data created to represent a *thing* (the most general class in an ontology; including physical objects, virtual objects, events, collections or any other abstract or concrete elements)(NOY; MCGUINNESS, 2016), is proportional to how good (complete, precise, accurate, current, among other DQ dimensions relevant in the context) a data user is able to represent back the original *thing*.

Independently of the approach used to define the quality or DQ, most definitions in the literature consider the idiosyncratic nature inherent to the concept of quality. DQ only can be determined based on the point of view of data user (STRONG; LEE; WANG, 1997), different users can have a different perspective of quality about the same data. Accordingly, the most

accepted and used definition of DQ in the literature is related to *fitness for use* (STRONG; LEE; WANG, 1997; WANG; KON; MADNICK, 1993). Data have quality if they are fit for use.

Regarding all the mentioned DQ definitions in the literature, in this thesis, DQ is defined as *fitness for use according to data user's needs*.

2.1.2 Approaches on DQ researches

Most researches on DQ falls into answering these following questions: how to assess DQ? how to manage DQ? and how does DQ impact organizations? Accordingly, most of DQ investigations can be classified into three branches: DQ assessment, DQ management and contextual DQ (GE; HELFERT, 2007).

Research on DQ assessment aims at the measurement, validation and classification of the quality of data (REDMAN, 2001; HUANG; LEE; WANG, 1999; LEE et al., 2002; PIPINO; LEE; WANG, 2002; STVILIA et al., 2007). DQ management research aims at the improvement of the usefulness of data by means of error prevention and correction using quality control or quality assurance approach (HUANG; LEE; WANG, 1999; WANG, 1998; EPPLER, 2006). Contextual DQ research aims at the evaluation of the impact of DQ in organizations (CHENGALUR-SMITH; BALLOU; PAZER, 1999; BERNDT et al., 2001; KAPLAN et al., 1998).

According to the scope of this research, a brief overview of DQ assessment and DQ management is presented as follows.

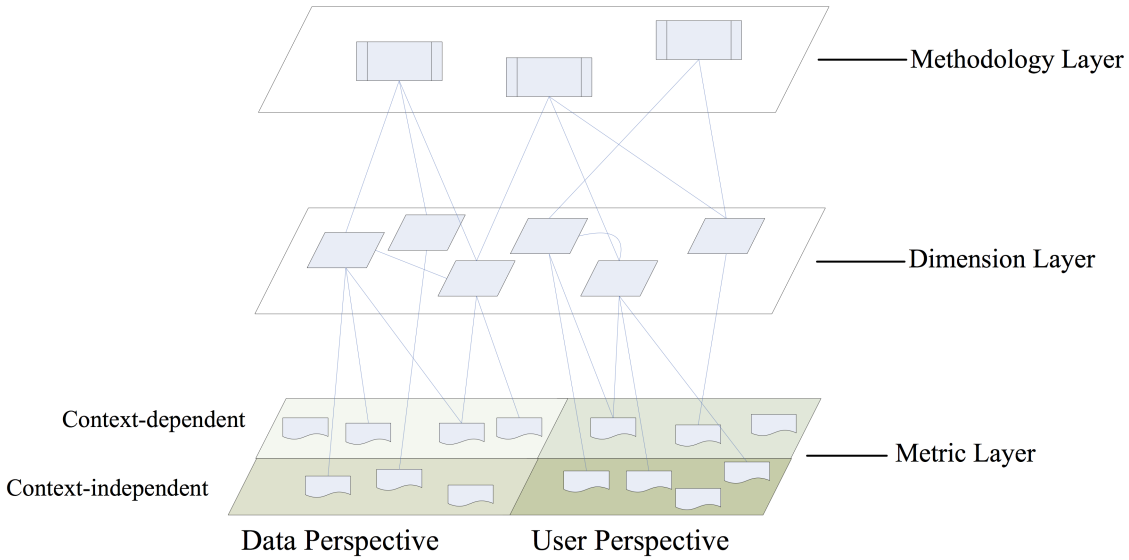
2.1.2.1 DQ Assessment

According to Ge and Helfert (GE; HELFERT, 2007), there are three key components in DQ assessment: DQ problems, DQ dimensions and DQ assessment methodology. In this context, DQ assessment can be organized into three layers: metric layer, dimension layer and methodology layer, as illustrated with Figure 4 (GE; HELFERT, 2007).

The metric layer represents DQ problems that degrade the quality of data. The dimension layer represents the quality aspects of data. The metric layer and the dimension layer are connected by links between quality dimensions and the problems that degrade the quality of these dimensions. For example, the *problem of incorrect value* (in the metric layer) can be linked to the *dimension of accuracy* (in dimension layer) because the quality of this dimension may be degraded by the mentioned problem. The methodology layer contains the DQ assessment models, frameworks and methodologies to perform the assessment. The methodology layer elements are linked to a set of dimensions (in dimension layer) (GE; HELFERT, 2007).

A brief description of DQ problem, DQ Dimension and DQ assessment methodology is presented next.

Figure 4: DQ assessment layers.



Source: (GE; HELFERT, 2007)

2.1.2.1.1 DQ problem

DQ problems represent the measurable degradation of the quality of data. DQ problems can be defined as classes of instances of errors or the generalization of similar errors occurrences that degrades DQ.

DQ problems can be classified according to a two-by-two model, as showed with Table 2. The columns capture problems from "data perspective" and "user perspective", and the rows capture problems as "context-independent" and "context-dependent". According to this model, DQ problems can be classified into four classes, represented by the quadrants of Table 2 (CHEN et al., 2009).

- **Data Perspective/Context-independent quadrant**: represent DQ problems in the database;
- **Data Perspective/Context-dependent quadrant**: represent DQ problems that violate the business rules;
- **User Perspective/Context-independent quadrant**: represent DQ problems that may happen in processing data;
- **User Perspective/Context-dependent quadrant**: represent DQ problems that may make data unsuitable to use.

Classifying DQ problems can help choose better methods to resolve each class of problems. Data perspective problems can be resolved through, for example, data cleansing algorithms,

Table 2: Classification of DQ problems

	Data Perspective	**User Perspective**
Context-independent	Spelling error Missing data Duplicate data Incorrect value Inconsistent value Inconsistent data format Outdated data Incomplete data format Syntax violation Unique value violation Violation of integrity constraints Text formatting	The information is inaccessible The information is insecure The information is hardly retrievable The information is difficult to aggregate
Context-dependent	Violation of domain constraint Violation of organization's business rules Violation of company and government regulations Violation of constraints provided by the database administrator	The information is no based on fact The information is of doubtful credibility The information presents an impartial view The information is irrelevant to the work The information consists of inconsistent meanings The information is incomplete The information is compactly represented The information is hard to manipulate The information is hard to understand

Source: (GE; HELFERT, 2007)

data mining rules, statistical process control or dictionary matching routines (GE; HELFERT, 2007). User perspective problems often are harder to be resolved by automated processes and require optimization of resource allocation, analysis of business issues, re-engineering process, or aligning data flow with manufacturing systems.

2.1.2.1.2 DQ dimension

DQ dimension is a key concept in most DQ researches. Various studies show that DQ itself is a multi-dimensional concept (DALCIN, 2005; BALLOU; PAZER, 1985; HUANG; LEE; WANG, 1999; REDMAN, 1996; WAND; WANG, 1996; WANG; STRONG, 1996; GE; HELFERT, 2007). DQ dimensions are measurable attributes of quality. They are used to measure specific aspects of quality, such as completeness, accuracy, precision, credibility and consistence, for instance.

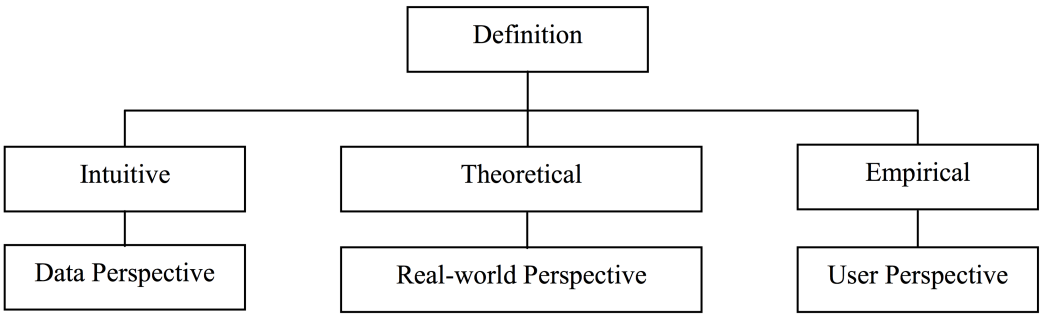
In a practical way, DQ dimensions can be considered the opposite of DQ problems, based

on the premise that data free of defects are considered data with quality (REDMAN, 2001). In this perspective, a set of selected DQ dimensions should describe what means DQ in a given context, and a set of selected DQ problems should describe what means lack of DQ in a given context. In other words, identify and properly define a set of relevant DQ dimensions should define the meaning of DQ itself in a given context, while identify and define a set of relevant DQ problems define the opposite of DQ.

Although researches on identifying different sets of dimensions of DQ are being developed over more than three decades, there is no consensus on an ideal composition set of dimensions or even definitions for these dimensions. The relevance and the meaning of each dimension can be perceived differently according to the data use context.

To identify a suitable set of DQ dimension and give a definition for each dimension in a given context is essential to understand the meaning of DQ. Identification and definition of DQ dimensions can be conducted using three approaches according to Wang and Strong (WANG; STRONG, 1996): intuitive, theoretical and empirical, as illustrated in Figure 5.

Figure 5: Definitions of DQ dimensions from different approaches.



Source: (GE; HELFERT, 2007)

Identification of a suitable set of dimensions in a given context using intuitive approach is centered on the experience of researchers from the application context. Using the theoretical approach, the DQ dimensions are identified based on the data deficiencies, for example, observing the inconsistencies between the real-world system and information system. The empirical approach focuses into identify relevant DQ dimensions based on the data user's perspective, selecting a set DQ dimensions that allow evaluating whether data are fit for use (GE; HELFERT, 2007; WANG; STRONG, 1996).

Definition of DQ dimensions using the intuitive approach is performed based on the data perspective. For instance, Ballou and Pazer (BALLOU; PAZER, 1985) define completeness as "all values for a certain variable are recorded" (GE; HELFERT, 2007). Using the theoretical approach, DQ dimensions are defined based on the real-world perspective. For instance, Wand and Wang (WAND; WANG, 1996) define completeness as "the ability of an information system to represent every meaningful state of the represented real world-system" (GE; HELFERT, 2007). Empirical approach defines DQ dimensions according to data user's perspective. For

instance, Wang and Strong (WANG; STRONG, 1996) define completeness as "the extent to which data are of sufficient breath, depth, and scope of the task at hand" (GE; HELFERT, 2007).

Select and define relevant DQ dimensions for a given context has been the base in most research on DQ for allowing data users to assess DQ according to their perspective of quality.

2.1.2.2 DQ assessment methodology

DQ assessment is the deed of judging the status of quality of data in a given context. For instance, DQ assessment may happen when data users or automated systems run rules to select a subset of data that have enough quality for a specific use, or when they judge that a specific record has high quality and another record has low quality, or even when they assign a numeric value that indicates the level of the overall quality of a record or dataset in a specific context.

At this point, it is important to highlight that DQ assessment is not the same as DQ measurement. DQ measurement indicates a measure of quality according to a specific DQ dimension, *e.g.* coordinates "completeness" of a dataset is equal to 98%. DQ assessment, often uses DQ measures in the process of judgment of quality, but others aspects beyond DQ measures can be taken into account for the assessment. In fact, DQ assessment can be performed in two ways: objectively or subjectively (PIPINO; LEE; WANG, 2002).

Objective assessment reveals clearly the extent to which information conforms to quality specification and references (*e.i.* comparing the current data value with the optimal data value). This approach can use software and strict rules (defined by the data users) to perform automated or semi-automated DQ assessment.

Subjective assessment happens when data users believe that some data have enough quality to be used, mostly based on their needs and experience. This approach may use measuring instruments, data provenance information or other complementary metadata (LEE; STRONG, 2003), but it may also be entirely based on the data user intuition and knowledge regarding the data and the context of use (GE; HELFERT, 2007). This approach can not be automated, which can lead to an unpractical use of time and specialized human resources; in addition, it may generate inconsistent results, where the assessment of a data user is different from the assessment of another data user.

Several DQ assessment methodologies have been proposed using objective, subjective or even a combination of both DQ assessment approaches (REDMAN, 1996; HUANG; LEE; WANG, 1999; LEE et al., 2002; PIPINO; LEE; WANG, 2002; STVILIA et al., 2007; GE; HELFERT, 2007). Usually, these methodologies define processes to identify and define DQ dimensions and propose or use tools and models for specific or general uses and for practical or theoretical applications.

In this context, this research does not intend to propose a complete methodology for tackling DQ. It proposes a conceptual framework that considers and uses relevant concepts used

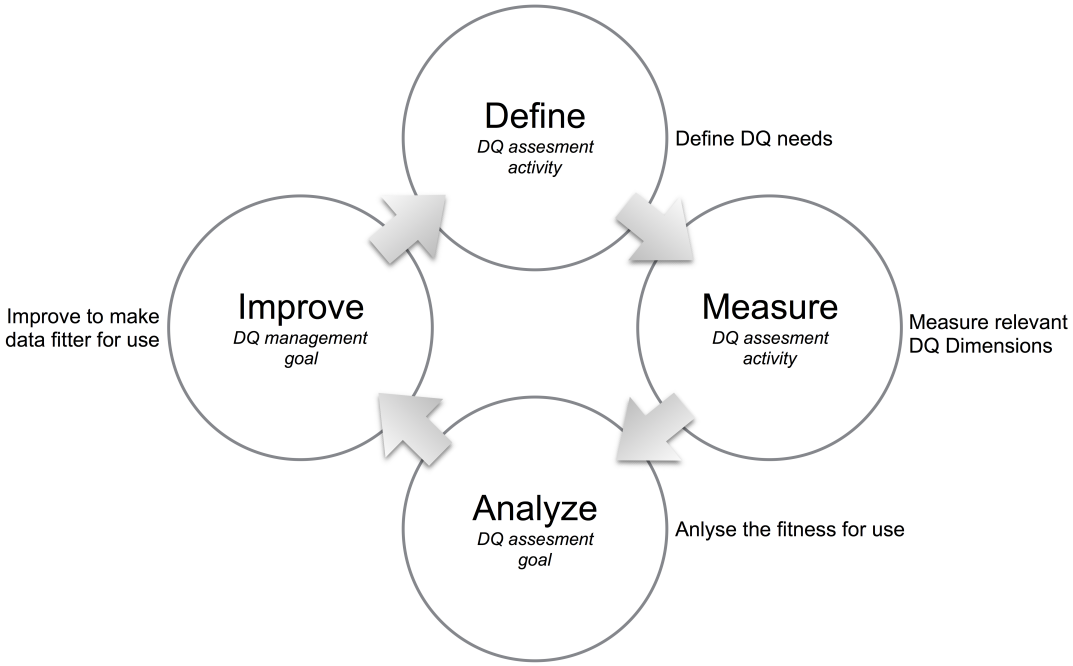
in DQ researches to guide thinking into a consistent way of gathering, defining and organizing the necessary components for enabling data users to assess DQ in a globally distributed data scope.

2.1.2.3 DQ Management

The overall aim of DQ management is to improve DQ, that is, make data fitter for use. Considering that the meaning of DQ and the current status of quality of particular data come from the DQ assessment outcomes, it is critical the DQ management be founded on DQ assessment (EPPLER, 2006). Management uses assessment outcomes as indicators for acting toward DQ improvement according to data user point of view, in order to obtain further better assessment outcomes in a virtuous circle.

Total Data Quality Management (TDQM), a comprehensive and structured approach to organizational management to improve the quality, is strongly based on DQ assessment outcomes (SearchCIO, 2005; WANG, 1998; LEE, 2006). In this method, *analyses*, which can be interpreted as DQ assessment outcomes, are used to *improve* DQ. *Analyze* are defined according to the relevant *measures*, which are based on dimensions *defined* according to DQ needs. As illustrated in Figure 6, this DQ management method encompasses not only the management itself, but DQ assessment related activities are embedded in the greater part of the cycle (define, measure and analyze).

Figure 6: DQ assessment in TDQM.



Source: Author

Another approach uses data associated with the knowledge of the process (know-what, know-how, know-why) to improve DQ (KNORR-CETINA, 1981 apud LEE; STRONG, 2003).

This approach is supported by the confirmed hypothesis that declares that knowledge generally reflects in data working performance (data production, data maintenance and data use), *i.e.* better knowledge will produce better data (LEE; STRONG, 2003).

In the literature we can find a wide range of approaches, principles, methods and guidelines used for DQ management. In several of these researches two basic fundamental concepts are implicitly or explicitly considered for DQ improvement: DQ control and DQ assurance (DALCIN, 2005; MCGILVRAY, 2008; VEIGA; SARAIVA; CARTOLANO, 2012; VEIGA; SARAIVA; CARTOLANO, 2014).

DQ control focuses effort on reducing the problems that cause quality degradation in the relevant DQ dimensions by preventing error, correcting error or recommending correction (MCGILVRAY, 2008). With DQ control, no data are discarded but it can not be assured that data will have suitable quality for a specific use, but it will try to improve the quality as better as it is feasible and, as result, data might achieve suitable quality for other different specific uses.

DQ assurance intends to assure that the available data have enough quality to be used according to data user's criteria. This approach accepts only data that are considered fit for use according to the DQ assessment and disregards unsuitable data for use. Figure 7 illustrates DQ control and DQ assurance.

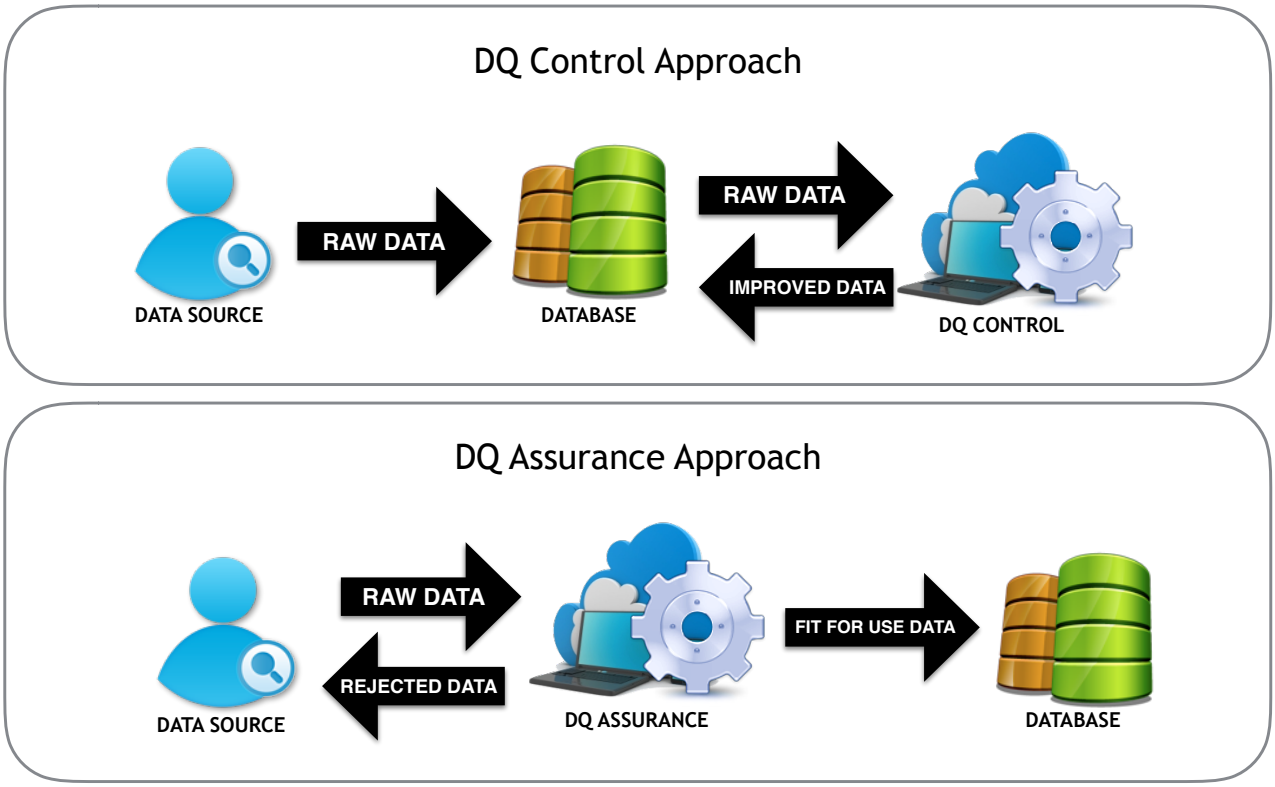
Another principle used in products and services quality management which is applicable to DQ management is that DQ improvement can be performed by preventing problems or correcting problems (DALCIN, 2005). As in products and services quality management, it is highly recommended to try to prevent problems whenever it is feasible and correct only when it was not feasible. Indeed, prevention is considered superior to correction, since correction involves errors detection, which is often costly and can not guarantee success on the detection stage or even on the correction stage itself (DALCIN, 2005).

It is highly important to improve DQ, however, seek correcting common errors without a solid foundation on the DQ assessment, can lead to impaired efficiency of DQ management, since the improvement efforts may not be aligned with the real DQ needs of data users.

2.2 BI Contextualization

It is notable that the concept of diversity is important in many areas of biology (MACLAURIN; STERELNY, 2008). But the term "biodiversity" comes particularly from conservation biology area. "Biodiversity" was coined in 1985 by Walter G. Rosen for the conference "The National Forum on BioDiversity" (HARPER; HAWKSWORTH, 1994 apud MACLAURIN; STERELNY, 2008). Its proceedings were edited by Edward O. Wilson, the same author of the seminal book *The Diversity of Life* written in 1992 (WILSON; REEDER, 1992). The aim of this book was to highlight the serious consequences of species loss, in particular to the loss of species caused by human activity.

Figure 7: DQ Control vs DQ Assurance.



Source: Author

An important contribution of Wilson's work is the concern with the number of different species. However, the biodiversity is not only comprised of the species number. Biodiversity can be defined as the variability among living organisms from all sources, such as, terrestrial, marine and other aquatics ecosystems, and the ecological complex which they depend on, including diversity among species, inter-species and ecosystems (RASHID et al., 2003; STEINHAGE, 2003; CDB,).

The use of biodiversity information is critical for decision-making in a wide range of domains (CANHOS et al., 2004). There is an increasing demand for understanding and solving complex environmental problems (SARAIVA; CANHOS, 2012). Similarly, as the development of capacity to forecast climatic events increase, it is also necessary to develop the capacity to forecast ecological impact based on changes, such as climatic changes and population growth.

To achieve such goal, it is essential, in order to perform comprehensive and useful analysis, to access and use integrated biodiversity information in a broad geographical and temporal scope, coming from different data sources and associated to additional information such as data about ecosystem and climate global changes, carbon life cycle and abiotic data, such as precipitation, humidity, temperature, among others (CANHOS et al., 2004).

In this context, the fledgling field of BI has joined biologists and computer scientists to

pull together, in a rising tide of coherence and organization, the major advances in globalization and interoperability of biodiversity information resources (BISBY, 2000).

BI is the application of computing concepts, techniques and tools to biodiversity information for improved capture, cleaning, management, improvement, analysis and interpretation in a global extent. Biodiversity information encompasses a wide extent of types of data, such as morphological data, genomic data, geospatial data, taxonomic data, nomenclatural data, environmental data and so long (PETERSON, 2014).

In this range of types of biodiversity information, the called "primary data" have been one of the most worked data in the BI community, since it supposes to offer information without subjectivity, assumptions, interpretations, or information loss. These characteristics allow data to be used for a wide range of purposes and, as result, "secondary data" may be generated, which involves analysis, interpretations and assumptions. Primary biodiversity data describe a specific taxon at a specific place and at a specific time with appropriate documentation (PETERSON, 2014; CHAPMAN, 2005b).

In order to enable interoperability and sharing of primary biodiversity data derived from a myriad sources stored in various format on many distinct platform around the world, Darwin Core was proposed in 1999 as a loosely defined set of terms and progressed through several iterations until it reached sufficient maturity and convergence in the BI community to be ratified as a standard by Biodiversity Information Standard (TDWG) in October 2009.

The Darwin Core standard, derived from previous standards work (e.g., Dublin Core), describes core sets of characteristics of biodiversity, such as location and taxonomic core sets, which are applicable in many biological domains, such as Paleontology and Botany, as illustrated with Figure 8 (WIECZOREK et al., 2012). In addition, this standard can be extended to cover details of specific sub-disciplines, such as Genetic Resources, Herbaria, Taxonomic Checklists.

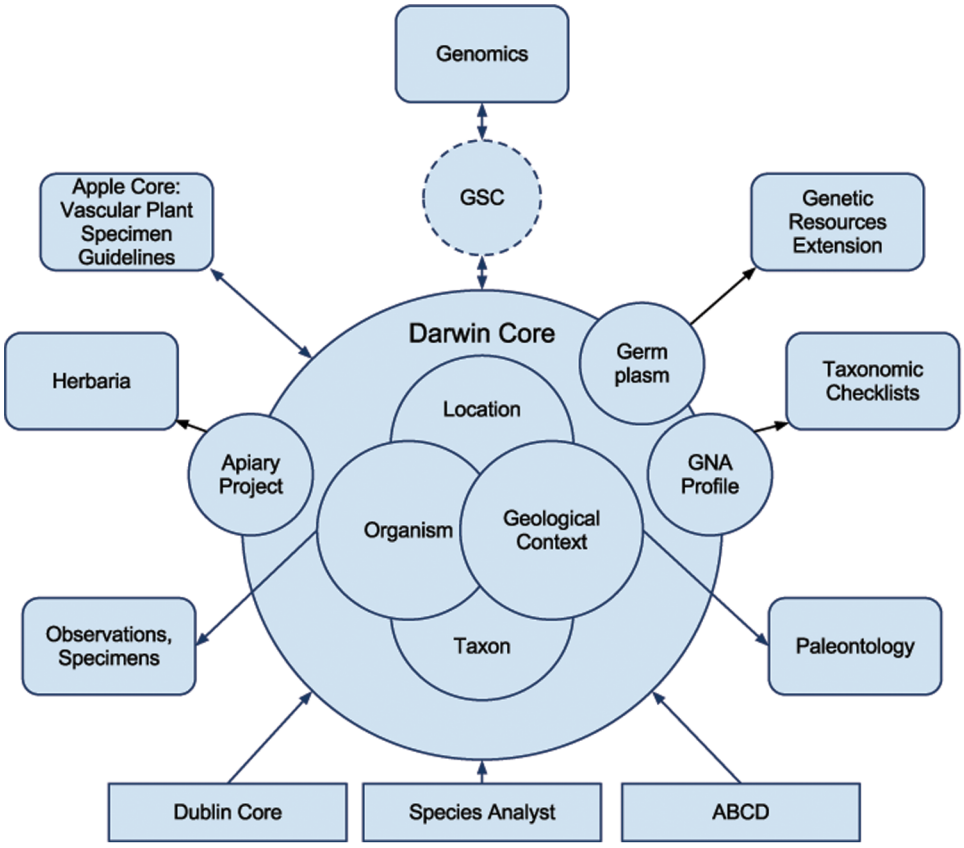
With this standard and many other efforts of the BI community, specially those ones involved with TDWG, several organizations were able to consistently achieve success on capturing, digitizing, managing, integrating, sharing and analyzing standardized biodiversity data.

One of the outstanding organizations in the context of BI is the Global Biodiversity Information Facility (GBIF), which currently indexes more than 600 million Darwin Core-formatted records published by more than 800 organizations from 54 countries.

The integration of data from a number of institutions and the subsequent publishing of those data brought to light many gaps in quality that previously were not evident. Currently, this is a critical issue and a great challenge to be tackled (HOBERN et al., 2013).

Despite the success of making primary biodiversity data digitally available, the quality of part of such data is so poor that makes unpractical the use of them. And the part of data that have a rich quality is not easily distinguished from data that have a poor quality rendering those rich quality data unreachable, lost and mixed in poor quality data. Researches show that quality of many of the data being available is currently not satisfactory for many of the desired

Figure 8: Scope of Darwin Core standard.



Source: (WIECZOREK et al., 2012)

purposes such as conservation and sustainable a distribution modeling, crop wild relatives and agrobiodiversity (BECK et al., 2014; MALDONADO et al., 2015).

For over many years several ideas for determining DQ and consequent fitness for use have been proposed and put into practice. For example, in 2005, GBIF commissioned the publication of manuals on DQ and data cleaning (CHAPMAN, 2005b; CHAPMAN, 2005a) in an attempt to inform and educate institutions on the importance of DQ and its relationship to fitness for use and to provide a range of methodologies that could be used to improve the quality of data in their institutions. Also in 2005, another seminal research on DQ applied to taxonomic database was published (DALCIN, 2005)

The evident concern with DQ has triggered effort also in the development of many tools to address DQ in different organizations; for very few examples see (VEIGA; SARAIVA; CARTOLANO, 2014; VEIGA; SARAIVA; CARTOLANO, 2012; DOU et al., 2012; ALA, 2012; CHAPMAN, 2004; OTEGUI et al., 2013).

However, to date, no consistent framework has been developed for assessing biodiversity DQ or documenting that quality; thus, it is difficult for users of biodiversity data to adequately compare the quality of one dataset to another and determine its fitness for use.

In this context, with the multiplicity of efforts and approaches raised to tackle DQ issues in BI community, it became evident the need for a common ground when speaking about DQ. What does quality mean? How is quality addressed by each institution or researcher? Which are processing techniques and tools used to accomplish certain quality characteristics? For which purposes? What are the results?

It is important that these and many other questions be asked and answered according to a common understanding of the underlying concepts they embed. It is necessary to engage a truly global collaborative effort that is founded on a common, comprehensive and consistent understanding of DQ and related principles in order to make the available biodiversity data properly assessable and manageable and to communicate stakeholders the status of quality of data with clarity and objectiveness.

In order to fill this gap, GBIF joined effort with TDWG to create the Biodiversity Data Quality Interest Group (BDQ-IG). In the BDQ-IG scope, three task groups were created. Task Group (TG) 1 was created based on the partial results of this research. TG1 aims to develop a conceptual framework that serves as a common ground for a collaborative development of solutions (encompassing tools, policies and concepts) for DQ assessment and management (BDQIG, 2015). TG 2 addresses tools as DQ solutions and TG 3 addresses use cases as DQ needs, both based on the conceptual framework proposed in this research.

2.3 Chapter Final Remarks

DQ is a subject that permeates most research fields where data are an important asset for answering questions. BI is one of these intensive research fields. Biodiversity quality data are defined in this thesis as fitness for use according to data user's point of view.

To determine data fitness for use in a given context, DQ assessment must be performed. Founded mostly on DQ dimensions, which are used to define and measure DQ, (or their opposite, e.i. DQ problems) data fitness for use can be assessed objectively and subjectively according to data user's perspective.

Based on the DQ assessment specifications and outcomes, the quality of data can be improved using DQ control and DQ assurance approaches of DQ management, by preventing error, correcting error or recommending corrections, in order to reduce the problems that degrade quality of the relevant DQ dimensions.

In BI research field, one of the current most challenging issues is the DQ assessment, and consequently, an appropriate and efficient DQ management. Due to the idiosyncrasy of the concept of DQ and the variety of types of biodiversity data and potential uses for them, it is difficult to define the meaning of DQ and how to allow the appropriate assessment and management in a global and distributed environment.

Next chapter presents a conceptual framework proposed to support the BI community

to define and organize the meaning of DQ and use these components to assess and manage DQ.

This conceptual framework formalizes human thinking into machine-friendly mind maps and makes it possible to communicate and reuse formalized needs and solutions among data user communities, data curators, and other stakeholders.

The framework developed in this document provides a context for describing biodiversity data quality, allowing users to make an informed assessment of DQ and its subsequent fitness for their use as well as allowing institutional assessment of DQ for management and improvement. This framework will provide a common ground for the collaborative development of solutions for DQ assessment and management based on data fitness for use principles.

The framework is the foundation for a joint DQ interest group (BDQIG, 2015) that TDWG and GBIF have established to serve as a common ground for a collaborative development of solutions for the BI stakeholders.

3 Conceptual Framework

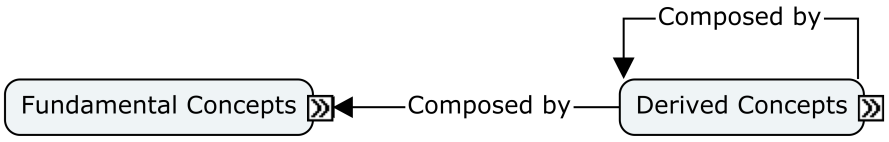
The proposed conceptual framework is comprehensive and involves a number of inter-dependent concepts and relationships that may be seen complex at the first read. For that reason, we present the framework using conceptual maps to illustrate the concepts and their relationships, in an attempt to mitigate the inherent complexity of the framework (NOVAK; CAÑAS, 2008). In Appendix A we use a mathematical notation to formalize the framework concepts. In Appendix B we provide some practical examples for a better understanding of how the framework can be used in a more practical way.

A conceptual framework can be defined as the way ideas are organized to achieve a research project's purpose (SHIELDS; RANGARJAN, 2013). This framework was designed to support the BI community to collaboratively reuse and design solutions for allowing data users to assess and manage the fitness for use of biodiversity data. Not included in the framework, although supported by it, is the design, development, and application of DQ software and processes.

In this sense, our understanding of "assessment" is the deed performed by data users or curators to judge the extent of the fitness for use of data (single record or dataset) for some specific purpose (PIPINO; LEE; WANG, 2002), and "management" is the deed performed by any actor (software, people, institution) to improve DQ to make data fitter for use for a wider range of uses (LEE, 2006). Loosely, DQ assessment involves, for instance, a data consumer selects a subset of records that conforms to that consumer's data fitness needs; while data quality management involves, for instance, making quality improvements to increase the fitness of the data for some specific usage.

The conceptual framework is composed of eight fundamental concepts and twenty-one derived concepts, which are composed of the relationships among the fundamental and other derived concepts, as illustrated in Figure 9. For example, the fundamental concepts Criteria (e.g. data must be in an acceptable format), Information Element (e.g. event date), and Resource Type (e.g. single record) are combined into the derived concept Contextualized Criteria (e.g. event date values of single records must follow the ISO 8601 standard format).

Figure 9: Relationships between the fundamental and derived concepts.



Source: Author

There are three classes of concepts: concepts concerning DQ Needs (colored in light yellow in the following figures), concepts concerning DQ Solutions (in blue) and concepts concerning DQ Report (in light green), as illustrated in Figure 10. The small squares with two arrows inside, attached to the concepts, represents, in this conceptual map notation, that the concept may encompass other inner concepts (NOVAK; CAÑAS, 2008).

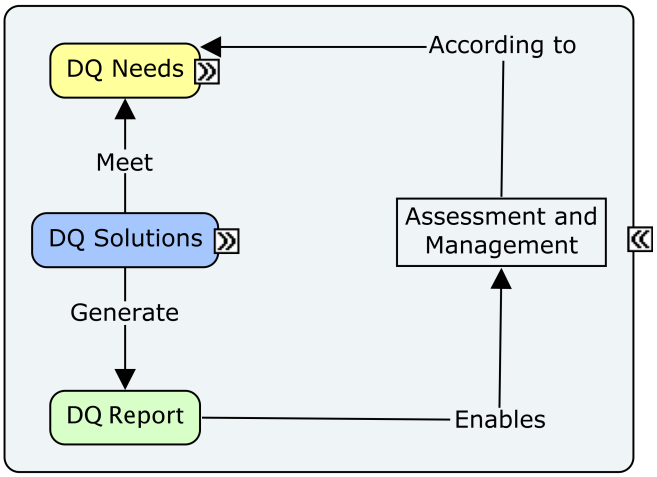
The DQ Needs class covers fundamental and derived concepts related to DQ requirements in a context delimited by the concept Use Case (see Section 3.1.1.1). The DQ Needs concepts describe how data must be addressed to enable the assessment and management of fitness for use of data according to a particular context.

The DQ Solutions class covers fundamental and derived concepts related to the methods and artifacts to enforce the compliance of DQ Needs. Therefore, DQ Solutions concepts define methods, tools and any other resource used to meet DQ Needs.

The DQ Report class covers fundamental and derived concepts related to outcomes of applying DQ Solutions to a specific data resource (a single record or a dataset) (see Section 3.1.3.1). Therefore, this class encompasses concepts related to assertions generated by the processing of a data resource, *i.e.*, a DQ validation assertion, measurement assertion or improvement assertion. A selected set of assertions should enable data users to assess and manage the fitness of the data resource for a particular use.

Splitting the concepts into these three classes is convenient. It allows users to gather their DQ requirements within the DQ Needs concepts, whereas developers can catalog available DQ tools with the DQ Solutions concepts. These two efforts can be conducted incrementally and separately to avoid dependency between them. This process promotes a synergy between them because both concept classes are linked and compatible. The DQ Report concepts map the assertions generated when DQ Solution are applied to data to comply with DQ Needs, which will enable fitness for use assessment and management.

Figure 10: Classes of concepts.



Source: Author

Those concepts will be described below in two subsections: fundamental concepts and derived concepts.

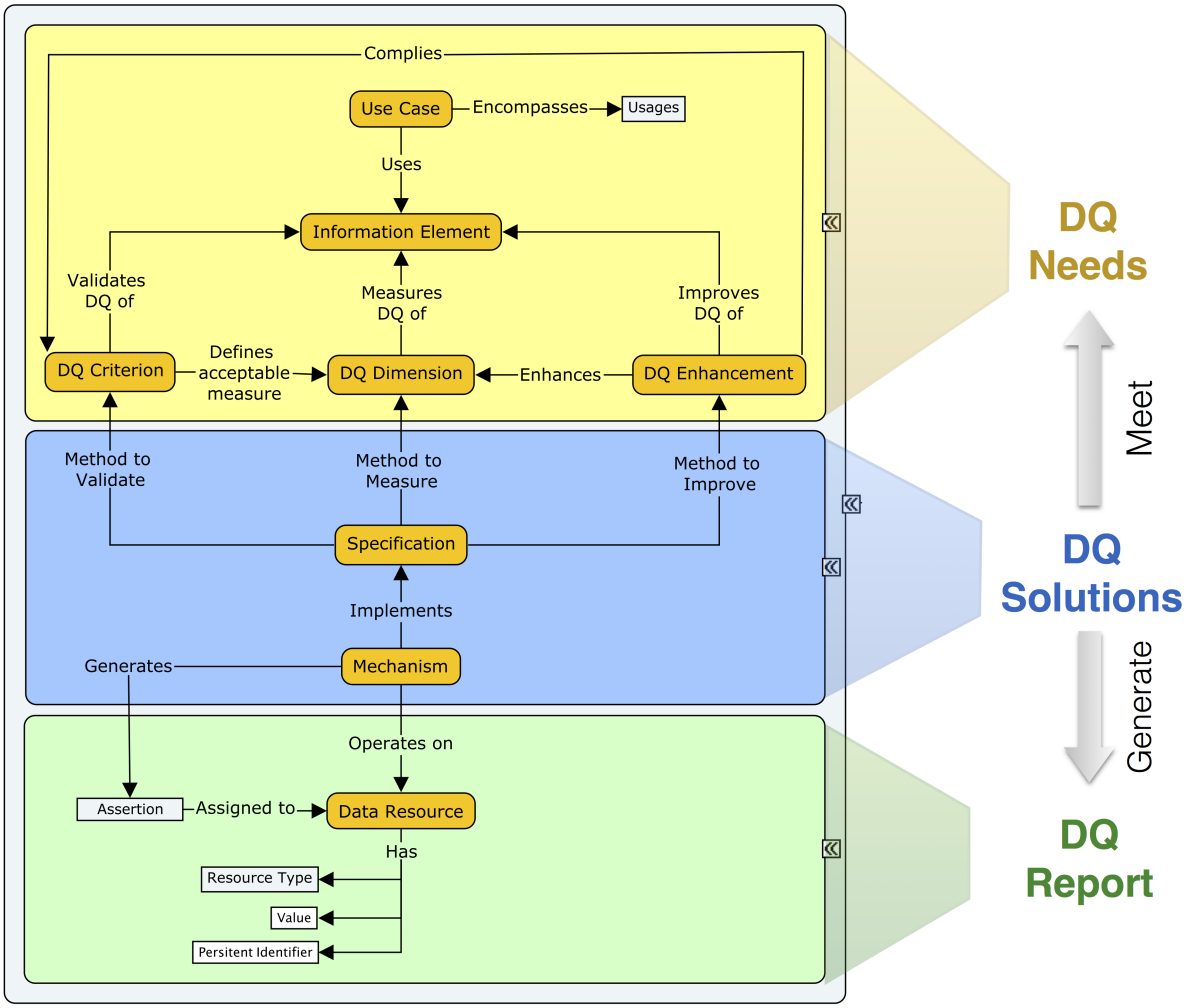
3.1 Fundamental Concepts

Fundamental concepts define elements with an independent meaning, that is, the definition of fundamental concepts instances are independent of relationships. This enables to foster the reuse of these concepts for different purposes and contexts.

The fundamental concepts and their relationships are represented by the conceptual map illustrated in Figure 11. The fundamental concepts are represented in dark-yellow boxes with rounded edges. Some fundamental concepts have certain properties (represented by light gray boxes with straight edges), e.g., Usages and Resource Type.

Each concept is in a higher-level box: light-yellow, blue or light-green, which represents the DQ Needs, DQ Solutions and DQ Report classes, respectively.

Figure 11: Conceptual map of fundamental concepts.



Source: Author

We now briefly describe the meaning of these fundamental concepts and present some examples in the BI context.

3.1.1 DQ Needs

The following concepts belong to the class DQ Needs, which means they are used to describe DQ requirements according to data consumers.

3.1.1.1 Use Case

By definition, DQ is related to **data fitness for use** (CHRISMAN, 1983). Therefore, it is necessary to clearly define what the **uses** are and what is meant by **fitness** in those **data**. A Use Case defines a scope delimitation concerning DQ Needs. This scope defines which DQ features data must be fit for use in a specific Use Case context. Each Use Case is composed of one or more specific Usages. These Usages represent specific tasks performed with data by data users, e.g., to generate a distribution model for the wild bee *Tetragonisca angustula s.l.* in Brazil.

Every attempt to assess and manage fitness for use should be guided by the data Usages described in a Use Case. The Use Case must be used as a reference to define any effort regarding DQ, including aspects such as DQ measurement, validation and improvement.

To define a Use Case in an institutional context, for instance, it is necessary to map stakeholders who handle data (collect, manipulate, use) and to list the Usages performed by them for each identified user. Based on the usages identified, a Use Case can be created by grouping usages based on the similarity of the input and output of the usages in the list.

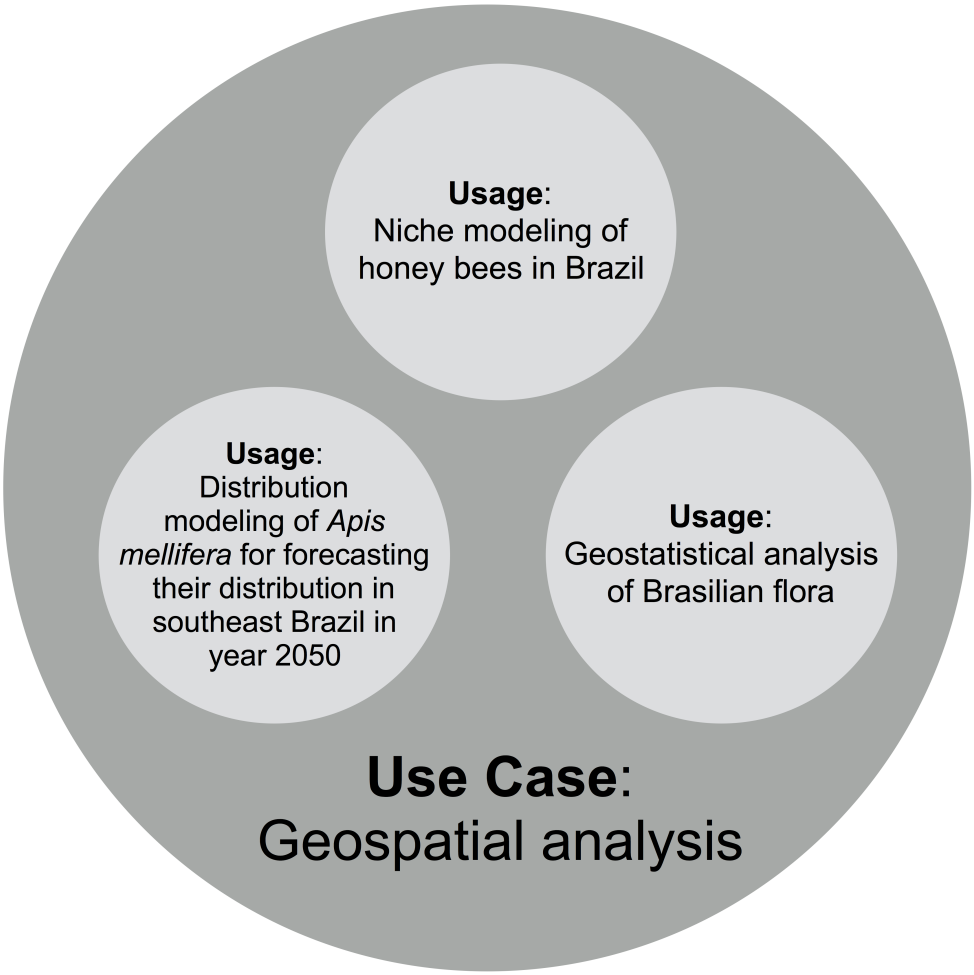
For example, in a given institution context, a Use Case for "geospatial analysis" purposes could be defined based on the Usages: "niche modeling of honey bees in Brazil", "distribution modeling of *Apis mellifera* for forecasting their distribution in southeast Brazil in year 2050", and "geostatistical analysis of Brazilian flora", as illustrated with Figure 12. In this context, DQ assessment and management could be performed to meet DQ requirements of these Usages, represented by the Use Case "geospatial analysis". See the Mathematical Formalization Introduction and Mathematical Formalization A in Appendix A and some related examples in Tables A, B and H in Appendix B for details.

3.1.1.2 Information Element

An Information Element (IE) is an abstraction that represents a relevant content in the Use Case context. An IE can be a single element or a set of elements that may represent an event, an object, an abstract data concept such as a GUID (Global Unique Identifier), or an entity of the real world, and has some importance in a data use context.

It can be classified as a single IE or a composed IE. For example, "decimal latitude" could be a single IE that represents, in decimal degrees, the position from the Equator to the

Figure 12: Example of a Use Case based on three related Usages.



Source: Author

north (positive values) or to the south (negative values) with valid values between -90 and 90, inclusive. "decimal coordinates" could be a composed IE that comprises decimal latitude, decimal longitude, Datum and uncertainty in meters, which represent the a specific position on the surface of the Earth using decimal degrees (CHAPMAN; WIECZOREK, 2006).

There is a subset of IE that is required for each usage; therefore, this subset should be the target of DQ efforts, either for quality measurement, validation or improvement. See the Mathematical Formalization Introduction in Appendix A and examples in Table C in Appendix B for details.

3.1.1.3 DQ Dimension

DQ is a multidimensional concept, that is, the DQ concept is defined by a set of Dimensions that describes important quality aspects in some context (DALCIN, 2005; MCGILVRAY, 2008; WANG; REDDY; KON, 1995; STRONG; LEE; WANG, 1997). Dimensions are measurable quality aspects of data (WANG; STRONG, 1996). When the quality of some data is measured,

a set of Dimensions is used to obtain this quality measurement. For example, in a given context, data with high quality could mean data that are complete, precise, credible and accurate, so in this context, the quality of data will be proportional to the measure of those DQ dimensions.

The relevance of a dimension for a specific purpose is relative (DALCIN, 2005; MCGILVRAY, 2008). In the mentioned example, DQ could be considered poor if the most important dimension was the timeliness and the measure for timeliness was considered low.

There are a number of classical DQ dimensions cited in the literature that can be used as reference (ASKHAM et al., 2013; WAND; WANG, 1996; FOX; LEVITIN; REDMAN, 1994; CAI; ZHU, 2015), but any measurable attribute useful for measuring the quality of data in the context of a Use Case can be adopted as a DQ Dimension. For reference, we can present some of the commonly accepted and widely used DQ Dimensions, such as: timeliness, credibility, accuracy, consistency, integrity, completeness, readability, fitness, accessibility, precision and believability (CAI; ZHU, 2015). See the Mathematical Formalization Introduction in Appendix A and examples in Table D in Appendix B for details.

3.1.1.4 DQ Criterion

A Criterion is a statement that describes principles or standards under which data are judged regarding quality. This concept is used to perform three types of validation:

- **Error detection**: Validates whether data instances are correct;
  - Ex.: "Decimal latitude must be between -90 and 90 degrees."
  - Ex.: "Coordinates of a record must be consistent with the country."
- **Acceptable measure requirement**: Validates whether the measurement of a Dimension is acceptable;
  - Ex.: "Measure of the completeness of point coordinates in a dataset (dimension) must be greater than 90%."
- **Compliance with business requirement**: Validates whether data meets some business requirement.
  - Ex.: "Scientific name must be in the national species checklist."

A set of Criteria must describe how data should be presented to meet the DQ needs of a particular Use Case. See the Mathematical Formalization Introduction in Appendix A and examples in Table E in Appendix B for details.

3.1.1.5 DQ Enhancement

Enhancements are statements that describe activities required to improve DQ. An Enhancement can be a description of a procedure, tool, software, equipment, protocol, a best practice or anything that can be used to improve DQ. There are four types of Enhancements:

- **Prevention**: for preventing incidents (errors);
  - Ex.: "Suggest similar and valid scientific names while typing."
- **Correction**: for correcting errors;
  - Ex.: "Correct taxon name based on the most similar name according to a nomenclatural authority."
- **Recommendation**: for recommending corrections.
  - Ex.: "Recommending coordinates based on the locality description."
- **Enrichment**: for enriching the data.
  - Ex.: "Associate known distribution maps and pictures of species to correspondent species occurrences."

An Enhancement can be classified into multiple types; for example, an Enhancement could be designed to prevent errors by recommending correct values; this features a prevention and a recommendation Enhancement simultaneously. See the Mathematical Formalization Introduction in Appendix A and examples in Table F in Appendix B for details.

3.1.2 DQ Solutions

The fundamental concepts below belong to the class DQ Solutions. These concepts are used to describe how to proceed to comply with DQ Needs.

3.1.2.1 Specification

Specifications are statements used to describe, in a formal or informal way, how to perform DQ measurement, validation and improvement. Specifications can be highly formal, using any of several formal languages defined for that purpose, or can also be informal, using, for example, natural languages to describe how to proceed (BISHOP, 2003). Dimensions, Criteria and Enhancements can be addressed in several ways. For example, the same "precision" Dimension could be measured using a numerical or a statistical approach; the Criterion "Scientific name must be able to be related to a nomenclatural act" could be validated using several different string similarity algorithms and parameters; and the Enhancement "Fill taxonomic hierarchy based on the most specific name" could be performed using a number of different taxonomic authority sources.

Thus, the concept Specification describes the method, technique or algorithm used to obtain the value of the DQ measurement, validation or improvement. See the Mathematical Formalization Introduction in Appendix A and examples in Table S in Appendix B for details.

3.1.2.2 Mechanism

A Mechanism can be software, hardware, a technique, a method, a procedure, a tool, a best practice or any other artifact that implements one or more Specifications to perform a DQ measure, validation or improvement. A Mechanism generates the values for the measurements, validations and improvements, called DQ Assertions.

Critical for understanding the conceptual framework is the distinction among Criterion, Dimension and Enhancement, and Mechanisms. Dimension, Criterion and Enhancement are statements that describe DQ Needs for a given context, concerning measurement, validation and improvement, respectively. Mechanisms are technical or nontechnical artifacts used to enforce compliance with the DQ Needs, which will indeed execute the measurement, validation and improvement (BISHOP, 2003).

There are two types of Mechanism based on its coverage:

- **Broad**: implements multiple Specifications.
  - Ex.: DwC-A Validator 2.0 is software that implements several DQ Specifications of validation and improvement (GBIF, 2015a). DwC-A Validator 2.0 could be considered a Broad Mechanism.
- **Precise**: implements a single Specification.
  - Ex.: A specific Web Service from Google Maps Application Programming Interface (API) could be a precise Mechanism designed to implement a specific improvement Specification that recommends the country name according to coordinates.

In some cases, we use more than one Mechanism to implement a Specification. To take advantage of reuse, we can define Mechanisms that depend on (or use) other existing Mechanisms.

For instance, a Specification for "validating coordinates" defines that latitude and longitude must be numbers, not empty, must not have invalid characters and must be in the correct range. In this case, we could implement a Mechanism that uses five other Mechanisms: (1) Check whether it is a number; (2) Check whether it is not empty; (3) Check whether it is free from invalid characters; (4) Check whether Latitude is in the range; and (5) Check whether Longitude is in the range.

Reusing existing mechanisms wherever possible is strongly encouraged to avoid duplication of effort. See the Mathematical Formalization Introduction in Appendix A and examples in Table T in Appendix B for details.

3.1.3 DQ Report

The following concept belongs to the class DQ Report. From the perspective of fundamental concepts, there is only one fundamental concept: Data Resource. This concept is the target of the DQ Solution concept. Any assertion about DQ that helps users assess and manage the fitness for use is assigned to a Data Resource.

3.1.3.1 Data Resource

Data Resource is an instance of data. This concept represents the data that have their quality measured, validated and improved. In the context of this research, Data Resources have three properties: Persistent Identifier, Resource Type and Value.

The Persistent Identifier is a global, unique and persistent identifier of the Data Resource, such as DOI, LSID, or URL (RICHARDS et al., 2011).

Resources Type, in the context of this research, can be "Single Record" or "Dataset". This property is important because it affects the method for measuring, validating and improving a Data Resource. For example, Coordinates Completeness of a "Single Record" could be measured qualitatively, checking whether the Latitude and Longitude of the record are filled or not, whereas the Coordinates Completeness of a "Dataset" could be measured quantitatively, measuring the proportion of Latitude and Longitude of all records of a dataset that are filled. Both measurements are for Coordinates Completeness, but they are measured in different ways due to the different Resource Type. In the context of this research, a single record may refer to a database tuple in the strict sense, that is, a row in a concrete table, or to a row in a flattened view or query across several related tables. Likewise, structured (e.g. Star Schema) DarwinCore representations of an Occurrence (e.g. an Occurrence with a determination history comprising a current identification and 5 previous identifications), are referred to here as a “Single Record”, even though it is composed of several records of structured data.

Value is the property that describes the content of the Data Resource. For example, it could be the content of a DarwinCore record that describes a single species occurrence or an entire dataset with records of a checklist of plants of a specific country. See the Mathematical Formalization Introduction and the Mathematical Formalization B in Appendix A and examples in Tables G and Z in Appendix B for details.

3.2 Derived Concepts

Derived concepts are defined by relationships between fundamental concepts or other derived concepts. These relationships define new concepts, e.g., "Valuable IE" is a derived concept that is composed of a subset of "IE" that is relevant (valuable) for a specific "Use Case".

To represent each derived concept, we use a conceptual map representation, where dark-green boxes represent Derived Concepts and dark-yellow boxes represent Fundamental Concepts.

The higher-level box represents the class to which the derived concept belongs (light-yellow for DQ Needs, blue for DQ Solutions and light-green for DQ Report).

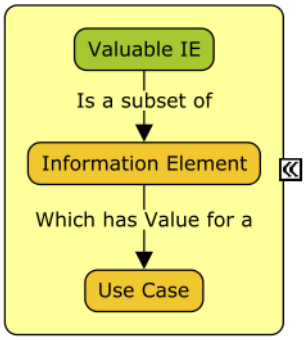
3.2.1 DQ Needs

The following derived concepts belong to the class DQ Needs. These derived concepts are used to describe DQ Needs according to data users.

3.2.1.1 Valuable IE

Valuable IE is a derived concept defined by a subset of IE that is valuable for the purposes of a particular Use Case, as illustrated in Figure 13. Each Use Case should have a definition of Valuable IE that represents the information that must have an appropriate level of quality to be fit for use in the context of the referred Use Case. Valuable IE represents the IE that are the targets of the fitness for use assessment and management in the context of a given Use Case.

Figure 13: Conceptual map of the derived concept Valuable IE.



Source: Author

For instance, a Use Case "Niche Modeling" could define as Valuable IE two IEs: "Coordinates" and "Taxon Name", which indicates that for Niche Modeling purposes, Coordinates and Taxon Name must be fit for use.

See the Mathematical Formalization C in Appendix A and examples in Table I in Appendix B for details.

3.2.1.2 Contextualized Dimension

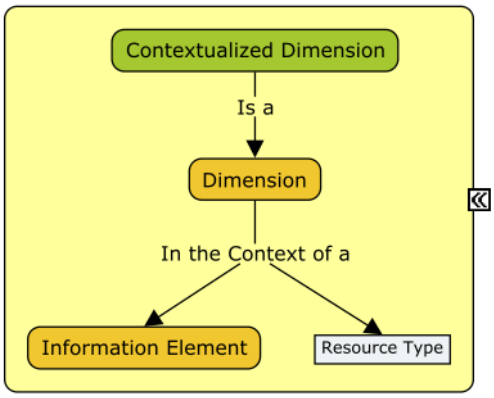
The same fundamental Dimension can have different meanings when it is related to different IEs. For example, "coordinate precision" and "taxon precision" have the same fundamental Dimension (precision), however, when they are associated to different IEs (coordinates and taxon) the meaning of "precision" changes. For instance, the fundamental Dimension "precision" could be defined as "the granularity or resolution of data", but when it is associated with the IE "coordinates", the meaning of "precision" could be "the number of significant digits of coordinates"

and when it is associated to the IE “taxon”, the meaning of "precision" could be "the taxon rank level".

The Resource Type can also modify the meaning and the way of measuring a fundamental Dimension. For example, "coordinate completeness of single records" could mean that all of the latitude, longitude, and datum fields have their values filled (qualitative measurement) in the record and "coordinate completeness of datasets" could mean the proportion of records with all of latitude, longitude, and datum that have their values filled (quantitative measurement).

Contextualized Dimension is a composition of a fundamental Dimension with an IE that this Dimension measures and the type of resource that the Dimension measures (dataset or single record), as illustrated in Figure 14.

Figure 14: Conceptual map related to the Contextualized Dimension.



Source: Author

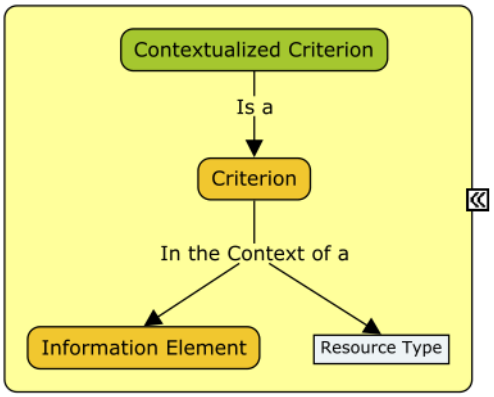
See the Mathematical Formalization D in Appendix A and examples in Table J in Appendix B for details.

3.2.1.3 Contextualized Criterion

Similarly to Dimensions, Criteria can also have different meanings and be validated differently according to the IE and Resource Type. Therefore, the Contextualized Criterion is a composition of a fundamental Criterion with an IE and a type of resource that it validates, as illustrated in Figure 15. For example, the fundamental Criterion "Human uncertainty must be assigned to data" could be used to define a Contextualized Criterion based on the Taxon Verification Status defined on pages 21-24 of (CHAPMAN, 2005b) to assign an uncertainty measure to the IE "Taxon Information" of Single Records; or with the same fundamental Criterion, another Contextualized Criterion could be defined based on a point-radius method to estimate the uncertainty in meters of the IE "Coordinates" of Single Records, according to (ALA, 2012).

s

Figure 15: Conceptual map related to Contextualized Criterion.



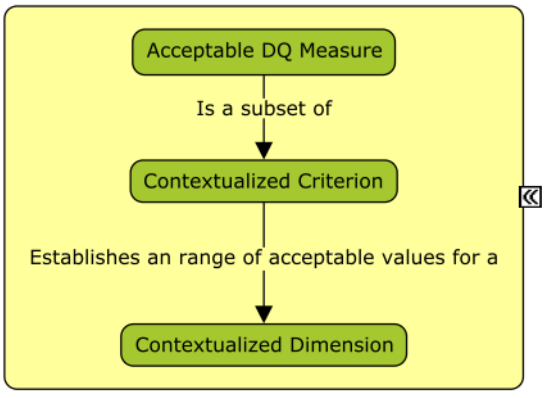
Source: Author

See the Mathematical Formalization E in Appendix A and examples in Table K in Appendix B for details.

3.2.1.4 Acceptable DQ Measure

A Contextualized Criterion may define an acceptable measure for a Contextualized Dimension. For example, the Contextualized Criterion "Datasets must have at least 90% of Completed Coordinates" could use the Contextualized Dimension "Coordinates Completeness of Dataset" to validate whether the dataset has an Acceptable DQ Measure. Acceptable DQ Measure is defined according to the relationship illustrated in Figure 16.

Figure 16: Conceptual map related to the Acceptable DQ Measure derived concept.



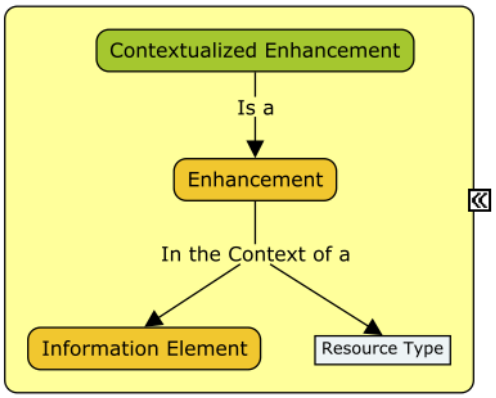
Source: Author

See the Mathematical Formalization F in Appendix A and examples in Table L in Appendix B for details.

3.2.1.5 Contextualized Enhancement

Similar to Dimensions and Criteria, the Enhancement can have different meanings and the procedure to improve DQ can also be different according to the IE and the Resource Type. Therefore, Contextualized Enhancement is a composition of a fundamental Enhancement with an IE and a type of resource that it improves, as illustrated in Figure 17. For example, a fundamental Enhancement "Recommend valid values" could be used to define a Contextualized Enhancement that uses the Catalog of Life (CoL, 2015) database to recommend similar names for the IE "Scientific Name" of Single Records; or could use the controlled vocabulary recommended by DarwinCore <http://tdwg.github.io/dwc/terms/index.htm#basisOfRecord> to recommend a value for the IE "Basis of Record" of Single Records.

Figure 17: Conceptual map related to the Contextualized Enhancement derived concept.



Source: Author

See the Mathematical Formalization G in Appendix A and examples in Table M in Appendix B for details.

3.2.1.6 Improvement Target

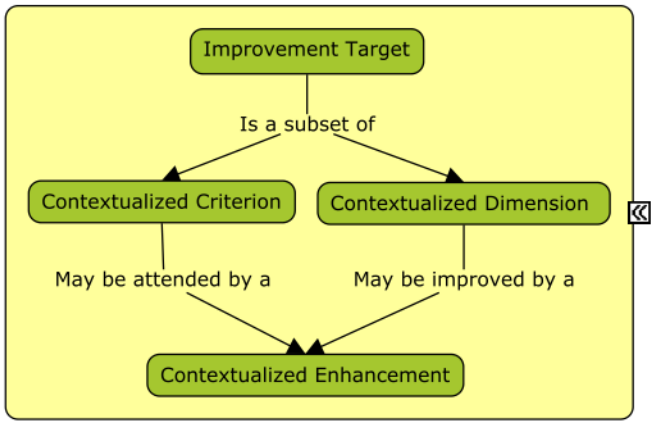
A Contextualized Enhancement may improve the measure of a subset of Contextualized Dimension and comply with a subset of Contextualized Criterion, as illustrated in Figure 18. For example, the Contextualized Enhancement "Recommending coordinates based on the locality description" can improve the Contextualized Dimension "Completeness of Coordinates of Single Records" and comply with the Contextualized Criterion "Species Occurrences records must be georeferenced".

See the Mathematical Formalization H in Appendix A and examples in Table N in Appendix B for details.

3.2.1.7 DQ Measurement Policy

DQ Measurement Policy consists of a subset of Contextualized Dimensions that are used to measure DQ of a specific Use Case, as illustrated by Figure 19. This subset of Contextualized

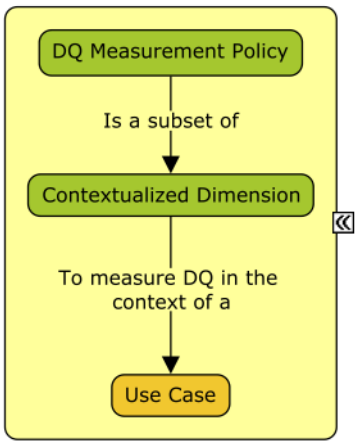
Figure 18: Conceptual map related to the Improvement Target derived concept.



Source: Author

Dimensions must be formed to support data users in assessing the data fitness for use in the context of a Use Case. This policy describes all pertinent DQ measurements that should be performed in a Use Case context and represents which quality aspects (Dimensions) are desirable for data users in the Use Case scope.

Figure 19: Conceptual map related to the DQ Measurement Policy derived concept.



Source: Author

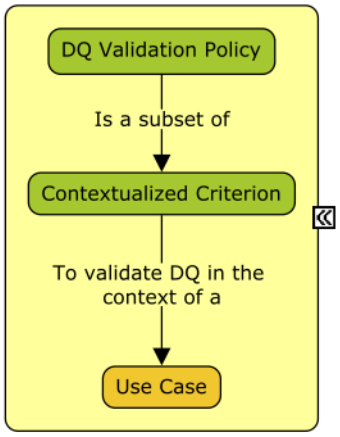
See the Mathematical Formalization I in Appendix A and examples in Table O in Appendix B for details.

3.2.1.8 DQ Validation Policy

DQ Validation Policy is a subset of Contextualized Criteria that are used to validate DQ in the context of a specific Use Case, as illustrated in Figure 20. DQ Validation Policy is a set of rules that shapes how data should be presented to be fit for use in the Usages defined in

a specific Use Case. This subset of Contextualized Criteria must be formed to support data users to assess the fitness for use of data and check for potential issues to be corrected in the context of a Use Case. This policy describes all pertinent DQ validation efforts that should be applied in the context of a Use Case to split data that have desirable quality from data that have undesirable quality.

Figure 20: Conceptual map related to the DQ Validation Policy.



Source: Author

See the Mathematical Formalization J in Appendix A and examples in Table P in Appendix B for details.

3.2.1.9 DQ Improvement Policy

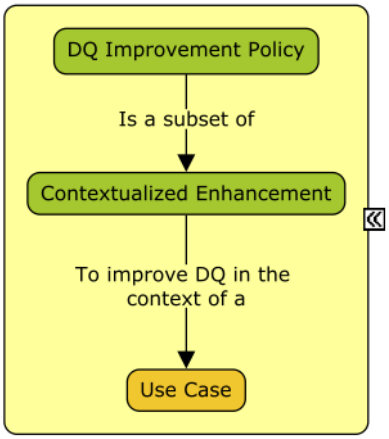
DQ Improvement Policy consists of a subset of Contextualized Enhancements that are used to improve the DQ in the context of a specific Use Case, as illustrated in Figure 21. This subset of Contextualized Enhancements must be formed to support data users to manage the data fitness for use by DQ control or assurance in the context of a Use Case. This policy describes all pertinent DQ improvement efforts that should be applied in the context of a Use Case.

See the Mathematical Formalization K in Appendix A and examples in Table Q in Appendix B for details.

3.2.1.10 DQ Profile

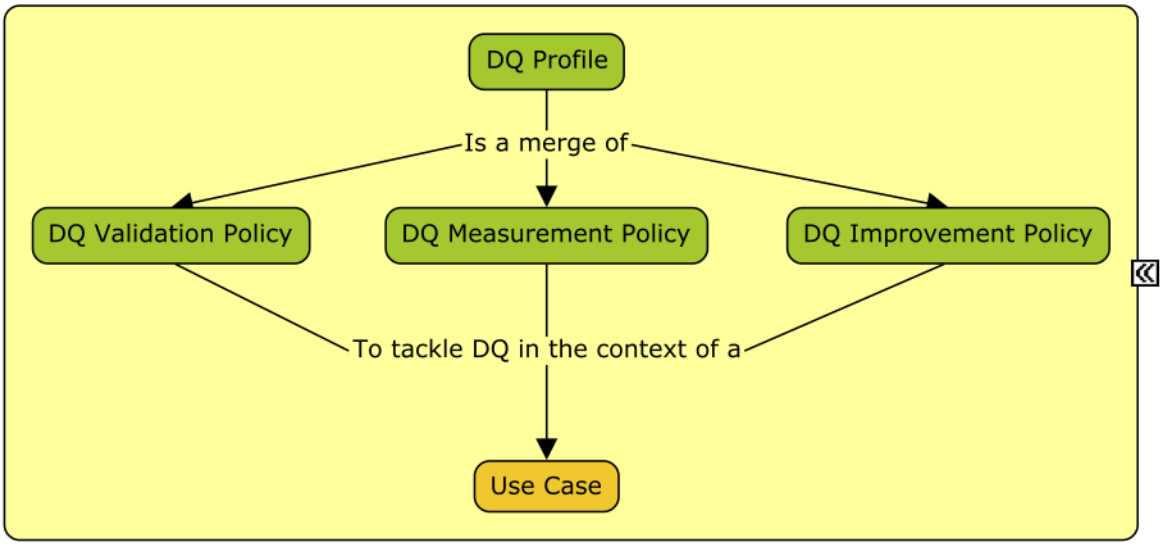
DQ Profile consists of the union of the DQ Measurement, Validation and Improvement Policies defined for a specific Use Case context, as illustrated in Figure 22. All of the efforts concerning DQ, in terms of measurement, validation and improvement, for enabling the data fitness for use in the assessment and management in the context of a Use Case should be described with a DQ Profile. Briefly, the DQ Profile describes the meaning of DQ needs in the context of a Use Case.

Figure 21: Conceptual map related to the DQ Improvement Policy.



Source: Author

Figure 22: Conceptual map related to the DQ Profile.



Source: Author

Ultimately, the DQ Profile encompasses all of the concepts regarding DQ Needs in the context of a Use Case. To define DQ Needs for a particular Use Case, it is necessary to define a DQ Profile and its dependent concepts (e.g., the "DQ Improvement Policy", their respective "Contextualized Enhancements", the related "Enhancement", "IE", "Resource Type" and "Improvement Target" for each "Contextualized Enhancement").

See the Mathematical Formalization L in Appendix A and examples in Table R in Appendix B for details.

3.2.2 DQ Solutions

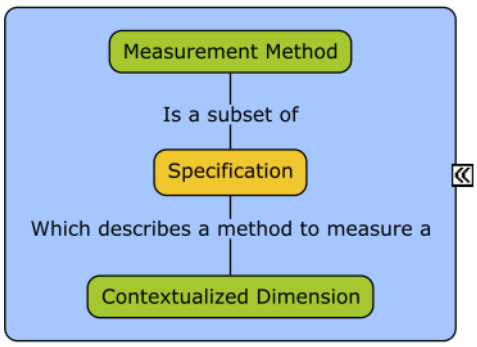
The following derived concepts belong to the class DQ Solutions. These derived concepts are used to describe how to proceed to comply with the DQ needs defined by a DQ Profile concept.

3.2.2.1 Measurement Methods

This is a subset of Specifications that describes how to perform a measurement of a specific Contextualized Dimension, as illustrated in Figure 23. A specific Contextualized Dimension could be measured using different algorithms, methods, languages or approaches, described by several Specifications.

For example, the Contextualized Dimension "Numerical Precision of Coordinates" could be measured by "Counting the number of character after the first '.' or ','" or by "Counting the number of numeric characters after the last '.'". Although the Contextualized Dimension is the same, the measured results may be different because the Specifications use different methods to measure. If data users know which Specification was used, they will be able to assess the fitness for use of data in a more conscious way.

Figure 23: Conceptual map related to the Measurement Methods.



Source: Author

See the Mathematical Formalization M in Appendix A and examples in Table U in Appendix B for details.

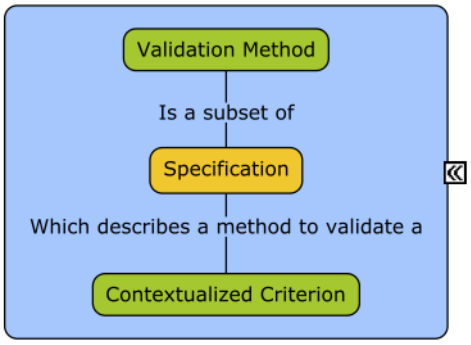
3.2.2.2 Validation Methods

Similarly to the Measurement Specification, there is a subset of Specifications that describes how to validate DQ according to a specific Contextualized Criterion, as illustrated in Figure 24. A specific Contextualized Criterion can be validated using different algorithms, methods, languages or approaches and can be described by several Specifications.

For example, the Contextualized Criterion "Basis of Record must be in the controlled vocabulary" could be validated by "Checking whether the value is exactly equal to at least one

item of the controlled vocabulary list recommended by DarwinCore" or by "Checking whether the uppercase value is exactly equal to at least one item also in uppercase transformation of the controlled vocabulary recommended by DarwinCore", and another variation of this last Specification could consider also "trimmed values (value with no extra spaces characters before or after the string)". Although the Contextualized Criterion is the same, the validation results may be different because the Specifications describe different methods to validate. If data users know which Specification was used to validate, they will able to assess and manage the fitness for use of data in a more conscious way.

Figure 24: Conceptual map related to the Validation Methods.



Source: Author

See the Mathematical Formalization N in Appendix A and examples in Table V in Appendix B for details.

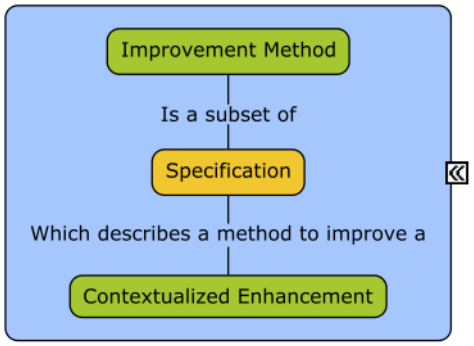
3.2.2.3 Improvement Methods

Similar to the measurement and validation Specifications, there is a subset of Specifications that describes how to improve DQ according to a specific Contextualized Enhancement, as illustrated in Figure 25. A specific Contextualized Enhancement could be performed using different algorithms, methods, languages or approaches and can be described by several Specifications.

For example, the Contextualized Enhancement "Recommend a similar and valid Scientific Name" could be performed by "Recommend the three most similar names in the Catalog of Life using the Levenshtein distance as a string-similarity metric" or by "Recommend the most similar name in the Catalog of Life using the Simple Matching Coefficient (SMC) as string-similarity metric". Although the Contextualized Enhancement is the same, the improvement results may be different because the Specifications describe different methods to improve. If data users know which Specification was used, they will able to assess and manage the fitness for use of data in a more conscious way.

See the Mathematical Formalization O in Appendix A and examples in Table W in Appendix B for details.

Figure 25: Conceptual map related to the Improvement Methods.



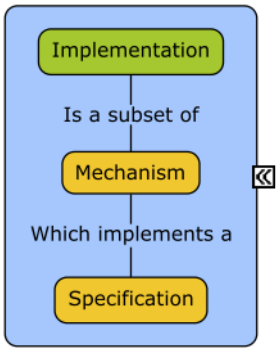
Source: Author

3.2.2.4 Implementation

A single Specification can be implemented in different ways by different institutions or developers and using different technologies and approaches, as illustrated in Figure 26. For example, the Specification "recommending the three most similar names in the Catalog of Life according to the Levenshtein distance" could be implemented by the Biodiversity Data Digitizer (BDD) tool (Cartolano, 2010) using a fuzzy matching library of PostgreSQL (PostgreSQL, 2015; LEVENSHTEIN, 1966). However, the same Specification could also be implemented by a specific tool of GBIF that could use a MapReduce approach running in a Hadoop ecosystem to do the same thing with more scalable technologies.

This derived concept defines a subset of Mechanisms that implements a specific Specification. A Specification can be implemented by several Mechanisms.

Figure 26: Conceptual map related to Implementation.



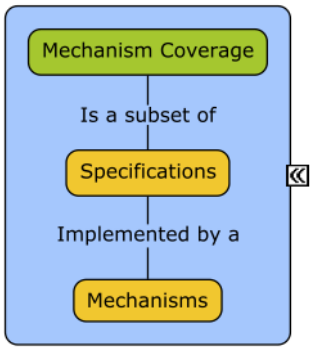
Source: Author

See the Mathematical Formalization P in Appendix A and examples in Table X in Appendix B for details.

3.2.2.5 Mechanism Coverage

Usually, tools are implemented to accomplish more than one task. For example, the ALA Sandbox (ALA, 2012) implements several Specifications of validation and improvement. These Specifications are the coverage of the Mechanism ALA Sandbox. This derived concept defines a subset of Specifications that a specific Mechanism implements, as illustrated in Figure 27.

Figure 27: Conceptual map related to Mechanism Coverage.



Source: Author

See the Mathematical Formalization Q in Appendix A and examples in Table Y in Appendix B for details.

3.2.3 DQ Report

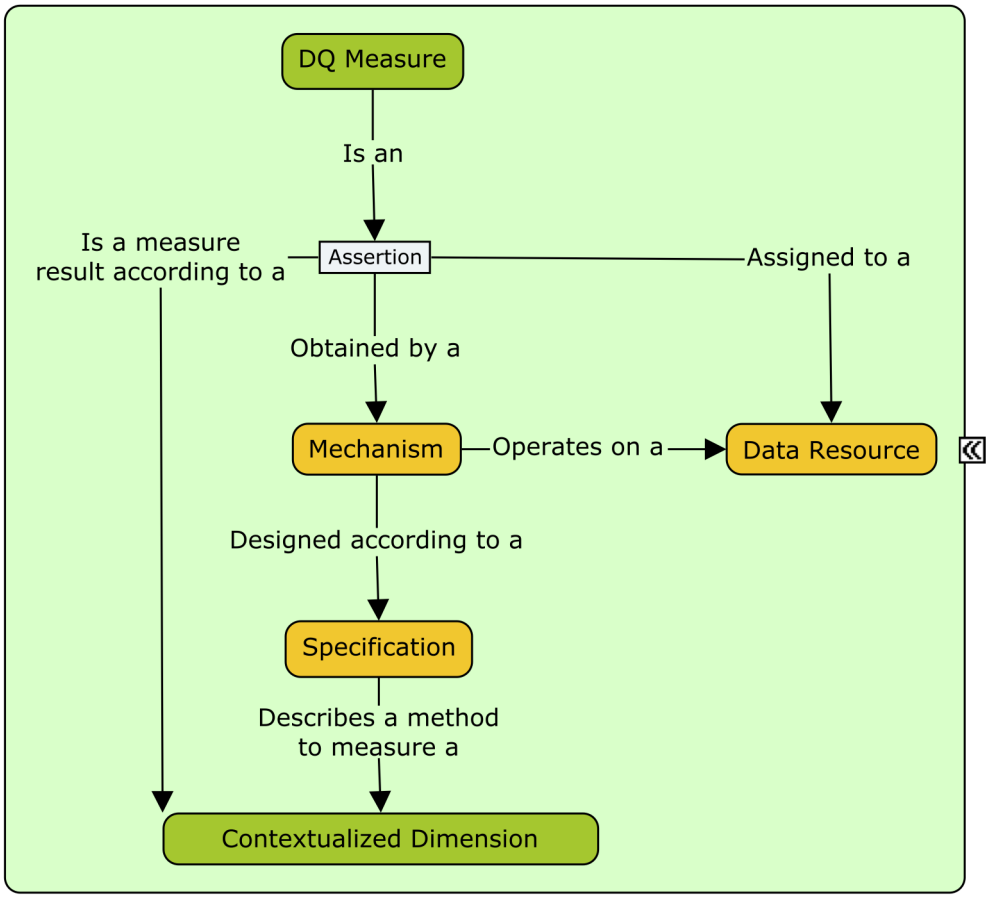
The derived concepts below belong to the class DQ Report. These derived concepts are generated by Mechanisms to enable data users to assess and manage the fitness for use.

3.2.3.1 DQ Measure

DQ Measure is an Assertion generated when a Mechanism processes a Data Resource for measuring a Contextualized Dimension according to a specific Specification, as illustrated in Figure 28. For example, a specific dataset (Data Resource) could have an Assertion equal to "0.78" for the Contextualized Dimension "Coordinates Completeness of Dataset", which was obtained by the "Biodiversity Data Quality Tool (BDQ-Tool)", which implements the Specification "Calculate the proportion of records with supplied Latitude and Longitude." DQ Measure is a derived concept that allows users to know the measure (Assertion) of a determined quality aspect (Contextualized Dimension) assigned to specific data (Data Resource) and the method (Specification) and tool (Mechanism) used to obtain it.

See the Mathematical Formalization R in Appendix A and examples in Table AA in Appendix B for details.

Figure 28: Conceptual map related to DQ Measure.



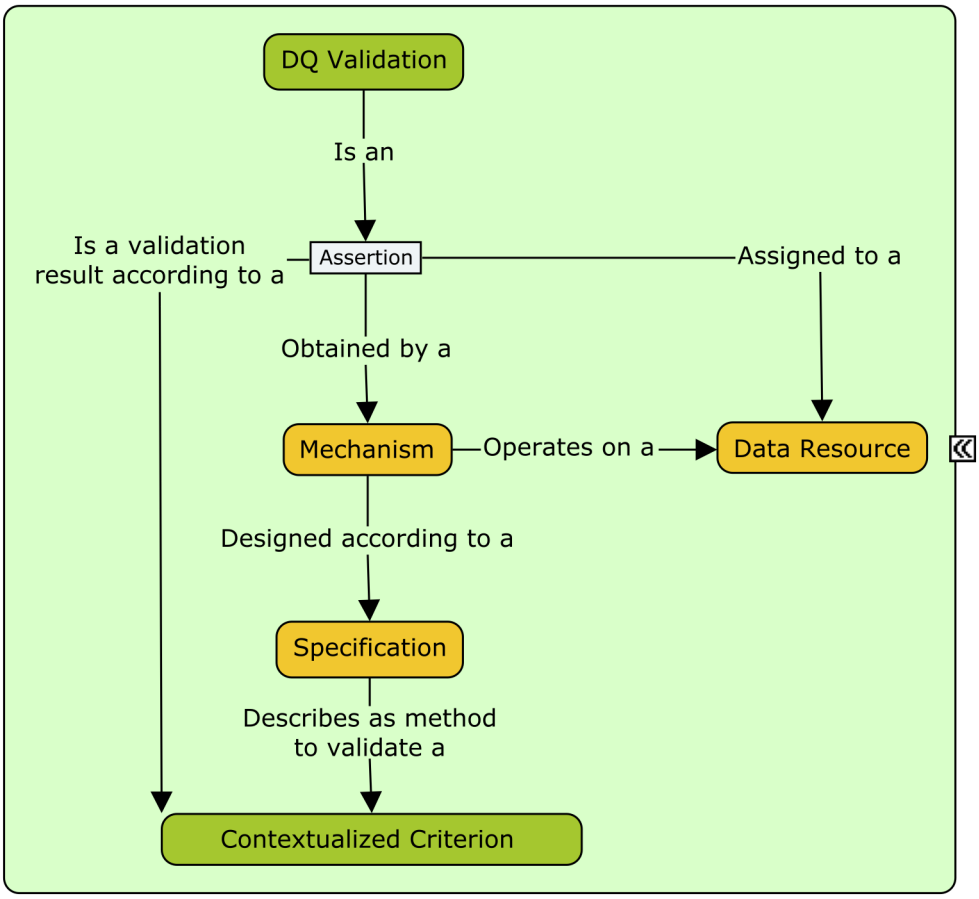
Source: Author

3.2.3.2 DQ Validation

DQ Validation is an Assertion generated by a Mechanism, according to a Specification, when it validates whether a Data Resource is in compliance with a Contextualized Criterion, as illustrated in Figure 29. For example, a specific record (Data Resource) could have an Assertion of "Not Valid" for the Contextualized Criterion "Basis of record must be well formed", which was obtained by the "Biodiversity Data Quality Tool (BDQ-Tool)", which implements the Specification "Check whether the supplied value of Basis of Record is in the DarwinCore controlled vocabulary." DQ Validation is a derived concept that allows users to know whether a determined quality criterion (Contextualized Criterion) was met (Assertion) for specific data (Data Resource) and which method (Specification) and tool (Mechanism) were used to validate it.

See the Mathematical Formalization S in Appendix A and examples in Table AB in Appendix B for details.

Figure 29: Conceptual map related to the Contextualized Dimension.



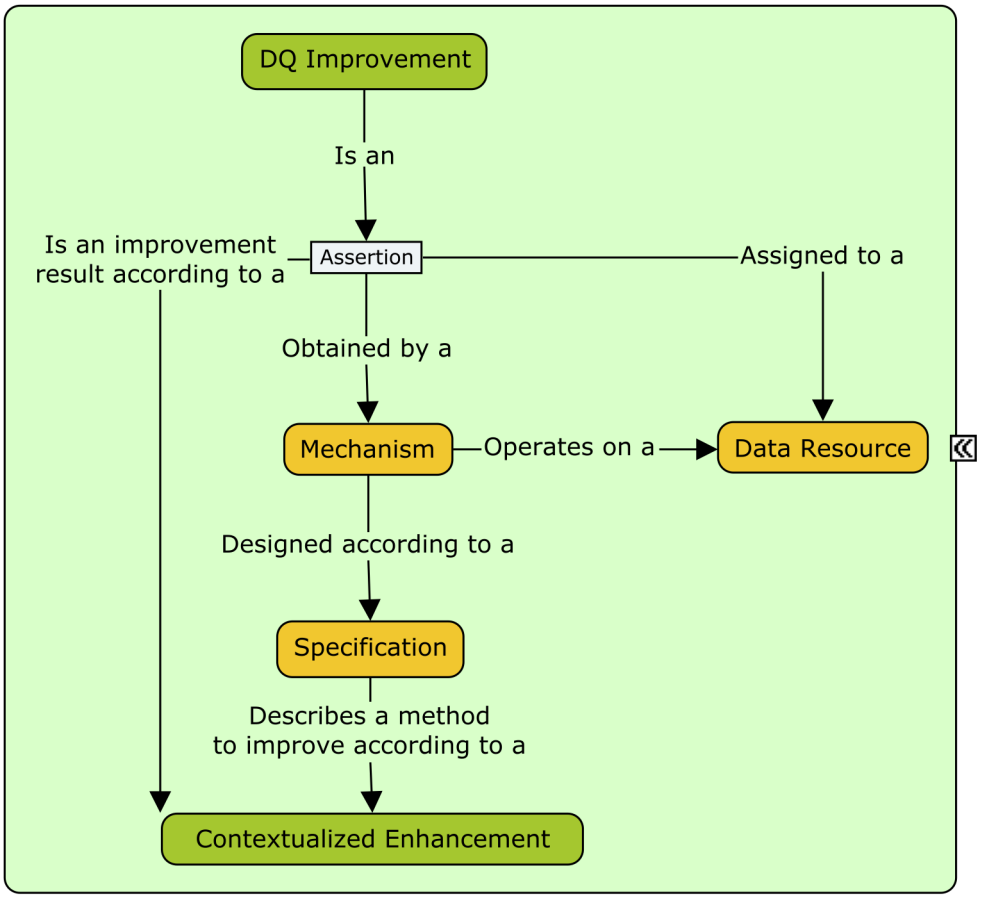
Source: Author

3.2.3.3 DQ Improvement

DQ Improvement is an Assertion generated by a Mechanism, according to a Specification, which improves a Data Resource based on a Contextualized Enhancement, as illustrated in Figure 30. For example, a specific record (Data Resource) with values "municipality=Rio de Janeiro; latitude=null; longitude=null; coordinatesUncertaintyInMeters=null" could have an Assertion of "municipality=Rio de Janeiro; latitude=-22.90278; longitude=-43.2075; coordinatesUncertaintyInMeters=80000" obtained according to the Contextualized Enhancement "Recommend coordinates based on the municipality centroid" and the Mechanism "Biodiversity Data Quality Tool (BDQ-Tool)", which implements the Specification "If coordinates are missing, recommend coordinates of the centroid of the supplied municipality and assign a consistent coordinatesUncertaintyInMeters based on the area of the municipality." DQ Improvement is a derived concept that allows users to improve DQ (Contextualized Enhancement) based on some correction, recommendation or enrichment (Assertion) of specific data (Data Resource) according to a determined method (Specification) and tool (Mechanism).

See the Mathematical Formalization T in Appendix A and examples in Table AC in Appendix B for details.

Figure 30: Conceptual map related to DQ Improvement.



Source: Author

3.2.3.4 DQ Assessment

DQ Assessment refers to the judgment of users about the fitness of data for use in a specific Use Case context. To support users to assess the data fitness for use in the context of a Use Case, the concept DQ Assessment defines a subset of DQ Measures, Validations and Improvements for a specific Data Resource, as illustrated by Figure 31.

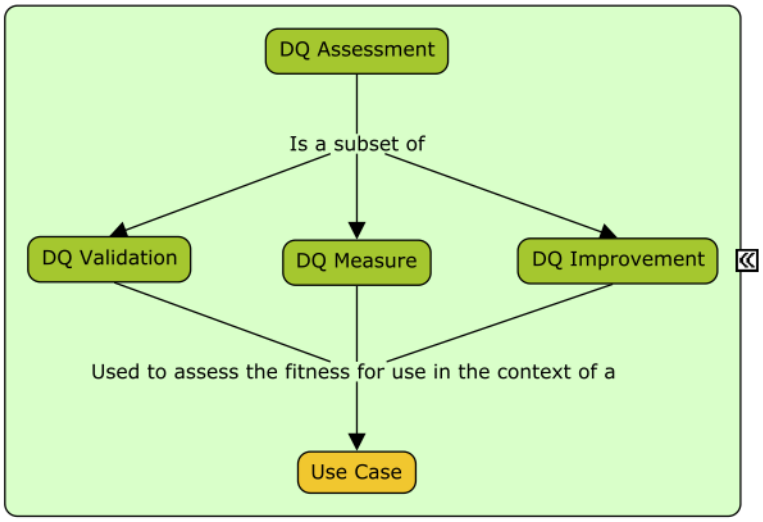
A data user is able to judge if a Data Resource is fit for use for a specific Use Case based on all of the DQ Measures, Validations and Improvements associated with the Use Case.

See the Mathematical Formalization U in Appendix A and examples in Tables AD, AE, AJ and AK in Appendix B for details.

3.2.3.5 DQ Management by Quality Control

The goal of DQ Management using a Quality Control approach is to reduce problems that can degrade DQ by preventing and correcting errors and recommending corrections, making data fitter for use for a wider range of usages. To support users to manage the data fitness for use, we provide a subset of DQ Validations and Improvements with respect to a specific Data

Figure 31: Conceptual map related to DQ Assessment.

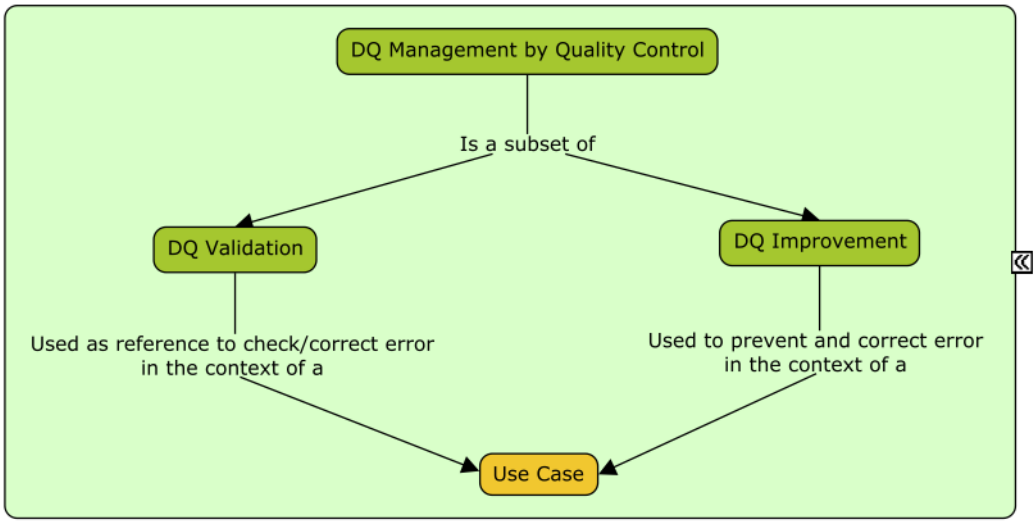


Source: Author

Resource, as illustrated by Figure 32.

Controlling the quality of a Data Resource can be performed based on the DQ Validations that have unsatisfactory Assertions and the DQ Improvements associated with a Use Case, to prevent and correct errors and accept recommendations.

Figure 32: Conceptual map related to DQ Management by Quality Control.



Source: Author

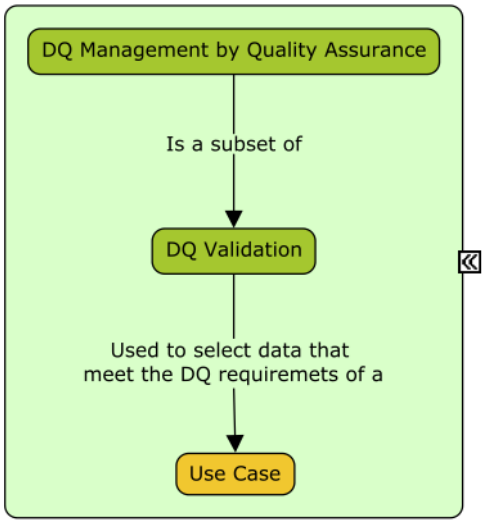
See the Mathematical Formalization V in Appendix A and examples in Tables AF, AG, AL and AM in Appendix B for details.

3.2.3.6 DQ Management by Quality Assurance

The goal of DQ Management using the Quality Assurance approach is to reduce problems that can degrade DQ by disregarding data that are not compliant with DQ needs, and consequently make data fitter for use for a wider range of uses. The goal of this approach is to accept only data that are in accordance with a subset of Contextualized Criteria, as illustrated in Figure 33. Only data that have satisfactory Assertion of DQ Validations are accepted.

Ensuring DQ of a Data Resource involves accepting or filtering only data that are compliant to the DQ Validations according to a Use Case context.

Figure 33: Conceptual map related to DQ Management by Quality Assurance.



Source: Author

See the Mathematical Formalization W in Appendix A and examples in Table AH and AI in Appendix B for details.

4 Framework Validation

For validating the conceptual framework, we present a proof of concept that addresses the instantiation of all the framework concepts for enabling the assessment and management of fitness for use of a real dataset. Details of the proof of concept are available in Appendix B.

4.1 Material and Methods

The proof of concept is based on a real application of the conceptual framework at the MCZ of Harvard University. The MCZ is a center for research and education focused on the comparative relationships of animal life. Its collections are comprised of approximately 21-million extant and fossil invertebrate and vertebrate specimens and almost 2-million digitized records (MCZ, 2016; MCZBASE, 2016). In this context we mapped and formalized the DQ Needs, DQ Solutions and DQ Reports required for the assessment and management of fitness for use of biological collections data at an institutional level. The scope of the proof of concept does not intend to cover all the aspects of DQ related to the curation of biological data, for the sake of conciseness, but it does cover all aspects necessary to stress all the framework concepts and to show that the conceptual framework is complete and consistent for mapping necessary concepts for the assessment and management of fitness for use of biodiversity data.

As MCZ uses (among other tools, techniques, and practices) a software package based on scientific workflow, called FP-Akka Kurator (DOU et al., 2012), to assist data curation, we have adopted a bottom-up approach to instantiate the framework. That means that instead of starting with the definition of DQ Needs and then defining the appropriate DQ Solutions (a top-down approach), we have started with mapping MCZ's DQ Solutions (the scientific workflow) and then defining which DQ Needs concepts they address. The FP-Akka Kurator can be used following the user documentation presented in the Kurator Project Wiki, available at: <http://wiki.datakurator.net/web/FP-Akka_User_Documentation>.

We initially noted that, within Kurator, some DQ Needs components, such as measures, validations and improvements applied to datasets were not addressed. In order to explore all aspects of the framework, we have implemented a new API that enables us to fill those gaps. The beta version of this API, called BDQ Toolkit, was developed specifically for this framework validation, but it can be used for other purposes. It is available at: <http://toolkit.bdq.biocomp.org.br:3020>.

In order to validate the framework with an appropriate dataset we applied the defined DQ Solutions to a subset of the dataset of Arizona State University Hasbrouck Insect Collection <http://symbiota4.acis.ufl.edu/scan/portal/collections>.

Briefly, the steps performed for the framework validation were:

1. Define a Use Case (scope delimitation);
2. Define DQ Solutions concepts instances based on the FP-Akka Kurator;
3. Define DQ Needs concepts instances based on the defined DQ Solutions concepts instances;
4. Complement DQ Needs with new concepts instances regarding datasets (in order to address all aspects of the Framework);
5. Build the BDQ Toolkit to meet the new DQ Needs concepts instances regarding datasets;
6. Complement the DQ Solutions concepts instances based on the BDQ Toolkit;
7. Run the FP-Akka Kurator and BDQ Toolkit against a real dataset;
8. Define DQ Report concept instances based on those results.
9. Discuss how the DQ Report enables the assessment and management of DQ both at the dataset level and at record level according to the DQ Needs defined for the Use Case.

All the used and generated data in this validation are available in JSON format at: <http://case.bdq.biocomp.org.br:3010/explorer>.

4.2 Results

The results of the validation are organized in three parts, corresponding to the three concepts classes of the framework: DQ Needs, DQ Solutions and DQ Report. Those results are described in details in Appendix B, following the same sequence of presentation of this text in Chapter 3.

4.2.1 DQ Needs

DQ Needs concepts define the meaning of "fitness for use" for a specific context. In summary, in this part of the framework validation we have defined a scope delimitation with a Use Case, a set of IEs that are valuable for this Use Case and a set of Contextualized Dimensions, Contextualized Criteria and Contextualized Enhancements necessary for the assessment and management of the fitness for use for the defined Use Case. See the first section of Appendix B for details.

We have started defining one Use Case for describing "Curation of Biological Collections" as scope delimitation and two related Usages ("General uses of occurrence data" and "Geospatial distribution of species"). The number of Usages assigned to the Use Case and the level of detail of the Usages were purposefully limited, in order to keep the validation simpler to follow and easier to understand, while not affecting the purpose of the framework validation.

We then defined four IE that are valuable for the tasks described by the two mentioned related Usages: Coordinates (Decimal Latitude, Decimal Longitude), Collected Date (Event Date), SciName (Scientific Name, Scientific Name Authorship) and Occurrence (Coordinates, Collected Date and SciName). For those Valuable IEs, it was defined a set of Contextualized Dimensions (DQ Measurement Policy) to measure the quality of each Valuable IE, a set of DQ Criteria (DQ Validation Policy) to validate the quality of each Valuable IE, and a set of DQ Enhancements (DQ Improvement Policy) to improve the quality of each Valuable IE. The three DQ policies have considered both datasets and single records.

The DQ Measurement Policy was composed by 18 Contextualized Dimensions that measure, in a quantitative or qualitative way, the Precision, Completeness, Accuracy and Consistency of the Valuable IEs.

The DQ Validation Policy was composed by 19 Contextualized Criteria that evaluate if a DQ measure has an acceptable value or if data complies with standard format requirements. Those Contextualized Criteria comprise the entire set of Valuable IEs.

The DQ Improvement Policy was composed by five Contextualized Enhancements that may improve the measure of 17 Contextualized Dimensions and make compliant 18 Contextualized Criteria by recommending corrections or applying corrections to both datasets and single records of the Valuable IEs.

The merging of those three policies composes a DQ Profile. This DQ Profile describes what has to be measured, validated and improved in order to enable the assessment and management of the fitness for use of biodiversity data in the context of Curation of Biological Collections. In order to avoid increasing the complexity of the framework validation, we have not addressed some Contextualized Dimensions, Contextualized Criteria and Contextualized Enhancement that may be valuable in a practical context (such as Precision of the Collected Date, or Accuracy of Coordinates), once the way we have defined the DQ Profile is enough for the purpose of this proof of concept.

4.2.2 DQ Solutions

DQ Solutions concepts define the methods and tools for meeting DQ Needs. In summary, in this part of the framework validation we have defined a set of Specifications as methods for measuring, validating and improving all the previously defined Contextualized Dimensions, Contextualized Criteria and Contextualized Enhancements, respectively, and a set of Mechanisms that implement those Specifications. See the second section of Appendix B for details.

We have defined 34 Specifications that describe methods that can be used to perform measures, validations and improvements. For each defined Contextualized Dimensions we have assigned the Specification that describes the method that can perform the correspondent measurement. Accordingly, for each defined Contextualized Criterion we have assigned the Specification that describes the method that can perform the correspondent validation. Finally,

for each defined Contextualized Enhancements we have assigned the Specification that describes the method that can perform the correspondent improvement.

Those 34 Specifications are implemented by two Mechanisms: FP-Akka Kurator, that implements six Specifications; and BDQ Toolkit, that implements 28 Specifications (mostly related to DQ measures, validations and improvements of datasets).

4.2.3 DQ Report

DQ Report concepts define a set of DQ measures, validations and improvements assertions assigned to a dataset or a single record. In summary, in this part of the framework validation we have presented how Data Resources can have their fitness for use assessed and managed with a set of DQ Measures, DQ Validations and DQ Improvements generated according to the previously defined DQ Solutions and DQ Needs. See the third section of Appendix B for details.

In order to better discuss the results, we have chosen one Data Resource of each Resource Type (Dataset and Single Record) to present in section 3 of Appendix B. The dataset is available at: <http://case.bdq.biocomp.org.br:3010/api/v1.0/OrginalData>; and the single record is available at: <http://case.bdq.biocomp.org.br:3010/api/v1.0/OrginalData/555f7b8ed53d8661fd3f53ed>, see Table Z of Appendix B. Although we have discussed only one single record, we have generated DQ Report concepts instances for each of the 52,412 records of the dataset.

We have assigned to the dataset nine DQ Measures and DQ Validations and two DQ Improvements, available at: <http://case.bdq.biocomp.org.br:3010/api/v1.0/DQReports/http\protect\relax\kern.22222em%2F%2Fcase.biocomp.org.br\protect\relax\kern.22222em3010%2Fapi%2Fv1.0%2FOrginalData%2F>, see Tables AA, AB and AC of Appendix B. Those DQ Measures and DQ Validations assertions, associated to the description of the method and tool used to generate them, can assist data users perform the fitness for use assessment of the dataset, according to the DQ Profile defined for the Curation of Biological Collections context.

In order to perform the DQ Management by the DQ Control approach, data users can use the result generated by the first DQ Improvement, which improves the quality of the dataset by accepting all corrections recommendations assigned to each single record of the dataset. This improvement result is available at: <http://case.bdq.biocomp.org.br:3010/api/v1.0/DQReportControls/http\protect\relax\kern.22222em%2F%2Fcase.biocomp.org.br\protect\relax\kern.22222em3010%2Fapi%2Fv1.0%2FOrginalData>, see Tables AF and AG of Appendix B.

However, if the approach desired for performing DQ Management is the DQ Assurance approach, data users can use the result generated by the second DQ Improvement, which improves the quality of the dataset by disregarding all records that are not compliant with all Contextualized Criterion. This improvement result is available at: <http://case.bdq.biocomp.org.br:3010/api/v1.0/DQReportAssurances/http\protect\relax\kern.22222em%2F%2F200.144.182.

24\protect\relax\kern.22222em3010%2Fapi%2Fv1.0%2FDQReportControls>, see Tables AH and AI of Appendix B.

For the mentioned single record, we have assigned to it nine DQ Measures, ten DQ Validations and two DQ Improvements, available at: <http://case.bdq.biocomp.org.br:3010/api/v1.0/DQReports/555f7b8ed53d8661fd3f53ed>, see Tables AA, AB and AC of Appendix B. All the DQ Measures, DQ Validations and DQ Improvements assigned for each single record of the dataset are available at: <http://case.bdq.biocomp.org.br:3010/api/v1.0/DQReports>. Those DQ Measures and DQ Validations assertions, associated to the description of the method and the tool used to generate them, can assist data users perform the fitness for use assessment of each single record according to the DQ Profile defined for Curation of Biological Collections context.

In order to perform the DQ Management by the DQ Control approach, data users can use the result generated by the two DQ Improvements, which recommend a correct SciName value and a correct Collected Date value. This improvement result is available at: <http://case.bdq.biocomp.org.br:3010/api/v1.0/DQReportControls/555f7b8ed53d8661fd3f53ed>, see Tables AL and AM of Appendix B. DQ Management by the DQ Assurance approach was not applied to single records.

4.3 Results Discussion

The presented conceptual framework was designed to be formal, comprehensive and flexible for being adapted and used to countless needs and scenarios.

Depending on the audience and its objectives, the framework implementation can be adapted in order to better fit the desired purposes. For example, from the point of view of the end users of data, the most important thing for them is to interpret the DQ Measures, DQ Validations and DQ Improvements assertions of the DQ Report; however, from the point of view of experts on biology that defines DQ Profiles, the most important part of the framework is identifying and defining the DQ Needs concepts instances.

We can mention a number of audiences, purposes and scenarios to which this framework can be beneficial, such as: (1) developers that want to build reusable, discoverable, comprehensible and standardized DQ tools; (2) national or global initiatives for building biodiversity data portals that want to provide tools for their data users to be enabled to select data based on fitness for use; researchers that want to share new methods on DQ measurement, validation or improvement (e.g. method for measuring the credibility of data provided by citizen scientists or the accuracy of taxonomic identification); (3) institutions that want to create, publish and share the DQ Profile used internally; (4) data users that want to attach to their researches or papers what DQ Profile was used to the DQ assessment and management of the data used in the research; (5) a global collaborative repository of DQ Needs and DQ Solutions concepts instances for easily composing new DQ Profiles based on existing concepts instances and reuse

available methods and tools to meet these DQ Profiles.

Considering those aspects and the complexity of the framework, it is not trivial to present all potential benefits it could address with only one case study. The results of the presented proof of concept address a small part of the potential benefits of the framework. It illustrates how the framework can be used to define the meaning of "fitness for use" in a specific institutional context and represent it in an objective and standardized format, which allows linking it to methods and tools that are able to generate assertions that allow data users to assess the fitness for use of datasets or single records. Besides allowing DQ assessment, the proof of concept also illustrates how the framework enables data users to manage the DQ of a dataset or single record with a DQ control or DQ assurance approach.

The proof of concept has demonstrated that the framework is adequately formalized and flexible, and sufficiently complete for description of DQ needs, solutions and reports in the BI domain.

5 Discussion

A consistent approach to assess and manage DQ is critical for biodiversity data users (HILL et al., 2010; GBIF, 2015b; MALDONADO et al., 2015; ANDERSON et al., 2016; ARNAUD, 2016). However, achieving this goal in the current BI scenario has been particularly difficult because of the idiosyncrasies inherent to the concept of quality (KLOBAS, 1995). DQ assessment and management only can be performed after it has been clearly established the meaning of data "fitness for use" according to data users' standpoint (NICOLAOU; MCKNIGHT, 2006).

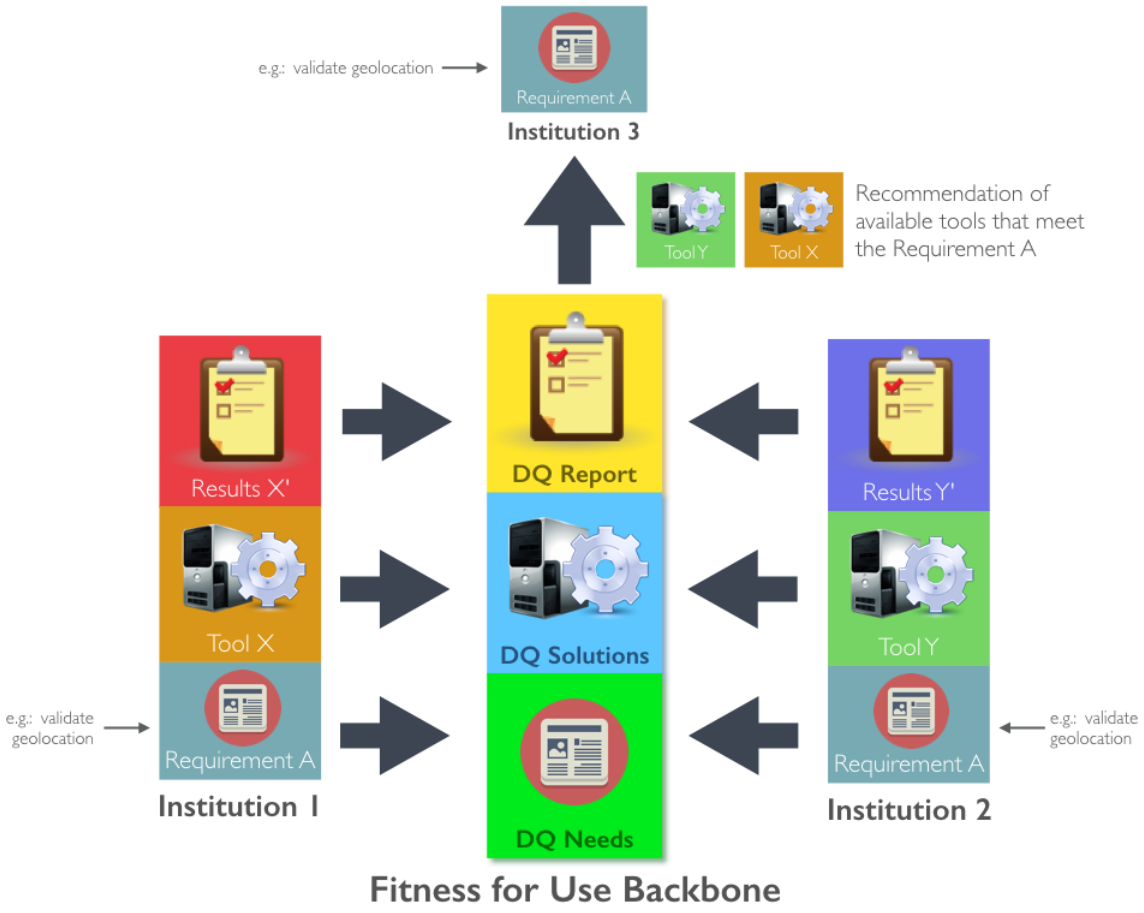
The proposed conceptual framework can support the BI community to describe the meaning of "fitness for use" from a data user's perspective in a formal and standardized way. Based on well-established principles and concepts from DQ literature (GE; HELFERT, 2007; LEE, 2005), the framework allows BI community to join effort to establish a common and standardized way to deal with DQ, in order to support data users in judging the fitness for use of data for specific usages, and in order to support data owners in improving their data, thus making them suitable for a wider range of usages.

Using this "common language" provided by the conceptual framework, the BI community can express and share their understanding of DQ Needs and DQ Solutions, thus improving the synergy of the community and consequently increase the reusability of solutions and decrease the duplication of efforts. With a collaborative model, fostered by the framework, we can generate a searchable repository of common and reusable DQ Profiles, DQ Policies, Use Cases, Dimensions, Criteria, Enhancements, Specifications and Mechanisms for a range of purposes of data usage, enabling institutions to compose their own DQ Needs and Solutions concepts to better suit their goals by reusing the DQ Needs and Solutions concepts shared by other institutions.

For example, if an institution (as Institution 1, in Figure 34) has a specific requirement (e.g. validate geolocation), it can implement a tool (as Tool X, in Figure 34) to meet such requirement, that will generate results (as Results X', in Figure 34). Concomitantly, another different institution (as Institution 2, in Figure 34) may have the same requirement (i.e. validate geolocation) and it also implements a tool (as Tool Y, in Figure 34) to meet the requirement, that will also generate results (as Results Y', in Figure 34). This scenario clearly demonstrates duplication of effort when two different tools are developed to meet the same requirement. In addition, the results generated by the tools, usually can not be compared because they do not share a common standard, generating a classical interoperability problem.

If all the requirements, tools and results were mapped to a common standard, it would enable the interoperability among the metadata about such components, making them comparable, shareable and reusable. It would enable, for example, an institution (as Institution 3, in Figure 34) that has a DQ requirement (e.g. validate geolocation) but that does not have

Figure 34: A "Fitness for Use Backbone".



Source: Author

a tool to meet such requirement, to search in this common “Fitness for Use Backbone" for available tools that have being used by other institutions to meet the specific requirement. In this scenario, the conceptual framework is able to map the requirements, tools and results by using respectively the concepts of DQ Needs, DQ Solutions and DQ Report classes.

Another relevant aspect of the presented conceptual framework is the fact that it splits the concepts into three classes, which increase the potential of mutual collaboration between stakeholders with different backgrounds. Concepts belonging to the DQ Needs class are typically better addressed by biodiversity experts, whereas concepts belonging to DQ Solutions are typically better addressed by informatics experts. The framework can be used to communicate their ideas using a notation and a common foundation that facilitate reaching an agreement. With this division, data users can individually define DQ Needs, and developers can propose DQ Solutions for meeting them, but bearing in mind the same basic conceptual framework. This model also allows us to systematically identify gaps with respect to DQ Needs that do not have appropriate DQ Solutions. For example, it may show that there is no suitable DQ measurement method for measuring a determined Contextualized Dimension (e.g. credibility of species observation from citizen science sources), or may show that there is this measurement

method, but there is no Mechanism that implements it.

DQ Report concepts class enables generating comparable DQ documentation that can be interpreted by data users to judge if a Data Resource is fit for use and how its quality can be improved. When data users or data owners have DQ Report elements (DQ Measures, Validations, and Improvements) assigned to a Data Resource, they can make a coherent assessment concerning the fitness for use of the Data Resource and compare its quality with that of other Data Resources.

The concern for DQ in the BI community has existed for several decades, but more recently it has become more evident and intense (CHAPMAN, 2004; DALCIN, 2005; CHAPMAN, 2005b; CHAPMAN, 2005a; HILL et al., 2010; OTEGUI et al., 2013; GBIF, 2015b; MALDONADO et al., 2015; ANDERSON et al., 2016; ARNAUD, 2016). The I and II Symposium on Biodiversity Data Quality that were held in the 2013 and 2014 TDWG Annual Conference, which were only part of the main meeting, received an impressive number of submissions - a total of up to 16 presentations, and had more attendants than the other parallel sessions, with its room filled beyond its maximum capacity. This is a clear evidence that there is a high interest of part of BI community in topics related to DQ and fitness for use. In fact, fitness for use is one of the core components of the Framework of Global Biodiversity Informatics Outlook (GBIO). This framework was proposed to help focus effort and investment towards better understanding of life on Earth. In this context, a key component is to promote efficient mechanisms to enable amateurs, experts and automated tools to correct and annotate data elements to improve quality and their fitness to be used for particular purposes (HOBERN et al., 2013).

Based on this demand, the Biodiversity Information Standards (TDWG) and the Global Biodiversity Information Facility (GBIF) have joined efforts to create the Biodiversity Data Quality Interest Group (BDQ-IG). In the scope of BDQ-IG, this conceptual framework was used as foundation for the TDWG/GBIF Task Group 1 (TDWG/GBIF TG1). The goal of that task group is to develop a conceptual framework that serves as a common ground for a collaborative development of solutions (encompassing tools, policies, and concepts) for DQ assessment and management based on data fitness for use (BDQIG, 2015).

The BDQ-IG has also created two additional task groups. TDWG/GBIF TG2 addresses DQ Tools, to catalog existing DQ tools and to identify gaps concerning the availability of tools, and TDWG/GBIF TG3 addresses DQ Use Cases, to identify DQ requirements for specific uses of data. In this context, the presented Framework can be used to map the outcomes from the Use Cases Task Group into the DQ Needs concepts, based on data users' expertise/knowledge and to map the outcomes from the DQ Tools Task Group into DQ Solutions concepts based on developers expertise. In this way, both groups can work in parallel, while maintaining the same conceptual foundation.

Improvement in biodiversity data is an essential step to be performed throughout the data life cycle. In particular, this step needs to for many workflows in biodiversity sciences, located between data gathering and data synthesis and analysis. Research satisfaction and trust in the results is often based on loose feelings or strict measurements of the data to be "enough

to answer the research questions". Understanding "what works" and "how much is enough" can be supported by the presented conceptual framework.

6 Final Remarks

Enabling DQ assessment and management is an urgent and serious issue in the current scenario of BI. To achieve this goal, a number of actions must be conducted, such as designing and developing effective methods, guidelines, policies, tools, services, protocols, metadata schemas, controlled vocabularies and training based on well consolidated international standards.

However, to get at this point, the BI community must use a common and solid conceptual base to guide discussions and actions to one single direction. To set a common conceptual background is the first step to efficiently enable the BI community to discuss and conduct proper strategies and actions toward DQ assessment and management. This agreement among the international community in pursuit of a common goal using an agreed-upon conceptual background is critical to avoid duplication of efforts and to increase the synergy among the stakeholders.

The presented conceptual framework was designed to fill this gap. Based on it, the BI community can progress on the development of the necessary components to enable DQ assessment and management in a consistent, formal and collaborative way.

The framework defines concepts for describing DQ needs in a given context, describing and linking solutions to meet such needs and generating standardized DQ reports to communicate the current DQ status. With that information data users are able to judge the data "fitness for use" and data owners/holders are able to perform the enhancement of DQ for making data fitter for a wider range of uses.

This work is supported by the GBIF/TDWG BDQ-IG, who created a task group for defining a framework for BDQ based on the presented conceptual framework. The goal of this task group is to provide a common ground for a collaborative development among the BI community.

Engaged with the TDWG/GBIF community, in the scope of DQ Framework task group, this work will continue to be advanced and improved, and new methods, best practices and tools regarding the use and the application of the framework in organizations will be proposed to foster the spread and adoption of the conceptual framework through the community, and to incorporate its input.

As future work, a new task group could be created to define a metadata schema and a controlled vocabulary based on the framework concepts, in order to enable the data and systems interoperability. In addition, we already have initiated discussions on how the conceptual framework can interact in a workflow context using provenance information. Initiatives on defining a formal ontology based on the conceptual framework would be important contributions to improve and foster the use of the framework.

In addition, as future work, we plan to build a platform that uses the conceptual framework to provide a backbone to register and retrieve information about DQ needs, solutions and report. This tool could enable defining, registering and sharing descriptions of Use Cases, IEs, Dimensions, Criteria, Enhancements, Measurement, Validation and Improvement Policies, DQ Profiles, Measurement, Validation and Improvement Methods, Mechanisms and Measures, Validations and Improvements applied to Data Resources. That repository can be applied, for example, to reuse DQ needs concepts to create new DQ Profiles, search for available DQ methods and Mechanisms and consult the current status of quality of Data Resource based on the associated DQ Measures, DQ Validations and DQ Improvement.

Bibliography

ALA. *ALA Sandbox | Atlas of Living Australia*. 2012. Disponível em: <http://www2014.ala.org.au/blogs-news/ala-sandbox/>.

ALABRI, A. Enhancing the Quality and Trust of Citizen Science Data. 2010.

ANDERSON, R. P. et al. *Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling*. Copenhagen, 2016. 1–27 p.

ARNAUD, E. *Report of the Task Group on GBIF Data Fitness for Use in Agrobiodiversity*. Copenhagen, 2016.

ASKHAM, N. et al. The Six Primary Dimensions for Data Quality Assessment. *Group, DAMA UK Working*, p. 16, 2013.

BALLOU, D.; PAZER, H. Modeling data and process quality multi-input multi-output. information systems. *Management Science*, v. 2, p. 150–162, 1985.

BDQIG. *GBIF Community Site: GBIF/TDWG biodiversity data quality interest group*. 2015. Disponível em: <http://community.gbif.org/pg/groups/21292/gbiftdwg-biodiversity-data-quality-interest-group/>.

BECK, J. et al. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, v. 19, p. 10–15, 2014. Disponível em: <http://daphnia.ecology.uga.edu/ecol8910-spring2015/?p=131>.

BERNDT, D. J. et al. Healthcare data warehousing and quality assurance. *Computer*, v. 34, n. 12, 2001. ISSN 00189162.

BISBY, F. A. The quiet revolution: biodiversity informatics and the internet. *Science (New York, N.Y.)*, v. 289, p. 2309–2312, 2000. ISSN 00368075.

BISHOP, M. *Computer Security: Art and Science*. Addison-Wesley, 2003. ISBN 9780201440997. Disponível em: <https://books.google.com.br/books?id=pfdBiJNfWdMC>.

BOLT, A.; MAZUR, G. H. Jurassic QFD: Integrating service and product quality function deployment. In: . [S.l.: s.n.], 1999.

CAI, L.; ZHU, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, v. 14, p. 2, 5 2015. ISSN 1683-1470. Disponível em: <http://datascience.codata.org/article/10.5334/dsj-2015-002/>.

CANHOS, V. P. et al. Global Biodiversity Informatics: setting the scene for a "new world" of ecological forecasting. *Biodiversity Informatics*, v. 1, p. 1–13, 2004. Disponível em: <http://www.taxondata.org/referencias/pdf/2004_Canhos.pdf>.

Cartolano. Biodiversity Data Digitizer. *The Proceedings of TDWG: Provisional Abstracts of the 2010 Annual Conference of the Taxonomic Databases Working Group*, 2010.

CDB. *A Convenção sobre Diversidade Biológica - CDB*. Disponível em: <http://www.ufrgs.br/patrimoniogenetico/arquivos-e-formularios/convencao-sobre-diversidade-biologica>.

CHAPMAN, A. D. *Environmental Data Quality - Data Cleaning Tools*. Campinas, Brazil, 2004. 56 pp p. Disponível em: <http://splink.cria.org.br/docs/appendix_i.pdf>.

CHAPMAN, A. D. *Principles and methods of data cleaning: primary species and species-occurrence data*. [S.l.], 2005. 72 p.

CHAPMAN, A. D. Principles of Data Quality. *Global Biodiversity Information Facility*, p. 61, 2005. Disponível em: <http://circa.gbif.net/Public/irc/gbif/pr/library?l=/webfiles/digit_documents/dataquality_pdf/_EN_1.0_{&}a=d{\T1\textbackslash}npapers2://publication/uuid/CB891E81-FF55-4CB5>.

CHAPMAN, A. D.; WIECZOREK, J. *Guide to Best Practices for Georeferencing*. Copenhagen, 2006. Disponível em: <http://www.taxondata.org/referencias/pdf/Best_Practices.pdf>.

CHEN, B. et al. Research and Implementation of Information Quality Improvement. In: . [S.l.: s.n.], 2009. p. 225–229.

CHENGALUR-SMITH, I.; BALLOU, D.; PAZER, H. The impact of data quality information on decision making: an exploratory analysis. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 11, n. 6, p. 853–864, 1999. ISSN 10414347. Disponível em: <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=824597>.

CHRISMAN, N. R. The role of quality information in the long-term functioning of a GIS. *APRS*, v. 2, p. 303–321, 1983.

CoL. *User Guide | Catalogue of Life*. 2015. Disponível em: <http://www.catalogueoflife.org/content/user-guide>.

CROSBY, P. B. *Quality without tears : the art of hassle-free management*. [S.l.]: McGraw-Hill, 1995. 205 p. ISBN 9780070145115.

DALCIN, E. C. *Data Quality Concepts and Techniques Applied to Taxonomic Databases*. 266 p. Tese (Doutorado) — University of Southampton, 2005.

DEVLIN, K. Naive Set Theory. In: . [s.n.], 1993. p. 1–28. Disponível em: <https://goo.gl/05c6ZF>.

DOU, L. et al. Kurator: A Kepler Package for Data Curation Workflows. *Procedia Computer Science*, v. 9, p. 1614–1619, 2012. ISSN 18770509. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050912002980>.

EPPLER, M. J. *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*. Springer Science & Business Media, 2006. 398 p. ISBN 3540322256. Disponível em: <https://books.google.com/books?hl=en{&}lr={&}id=nRcCx0riH_4C{}.>

FOX, C.; LEVITIN, A.; REDMAN, T. The notion of data and its quality dimensions. *Information Processing & Management*, v. 30, n. 1, p. 9–19, 1 1994. ISSN 03064573. Disponível em: <http://www.sciencedirect.com/science/article/pii/0306457394900205>.

GBIF. *dwca-validator, GitHub repository*. 2015. Disponível em: <https://github.com/gbif/dwca-validator>.

GBIF. *Terms of reference: Task group on data fitness for use on research into invasive alien species*. Copenhagen, 2015.

GE, M.; HELFERT, M. A review of information quality research-develop a research agenda. In: . [S.l.: s.n.], 2007.

HARPER, J. L.; HAWKSWORTH, D. L. Biodiversity: measurement and estimation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, v. 345, n. 1311, p. 5–12, 7 1994. ISSN 0962-8436. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/7972355>.

HILL, A. W. et al. *GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network*. Copenhagen, 2010. 25 p.

HOBERN, D. et al. *Global Biodiversity Informatics Outlook: Delivering Biodiversity Knowledge in the Information Age*. Copenhagen, 2013. 41 p. Disponível em: <http://imsgbif.gbif.org/CMS_ORC/?doc_id=5353{&}downloa>.

HUANG, K.-T.; LEE, Y.; WANG, R. *Quality information and knowledge*. Upper Saddle River: Prentice Hall, 1999. ISBN 0130101419.

KAPLAN, D. et al. Assessing Data Quality in Accounting Information Systems. *Communications of the ACM*, v. 41, n. 2, p. 72 – 78, 1998. ISSN 00010782.

KLOBAS, J. E. Beyond information quality: fitness for purpose and electronic information resource use. *Journal of Information Science*, v. 21, n. 2, p. 95–114, 1 1995. ISSN 0165-5515. Disponível em: <http://jis.sagepub.com/cgi/doi/10.1177/016555159502100301http://jis.sagepub.com/cgi/doi/10.1177/016555159502100204>.

KNORR-CETINA, K. K. *The manufacture of knowledge : an essay on the constructivist and contextual nature of science*. [S.l.]: Pergamon Press, 1981. 189 p. ISBN 0080257771.

LEE, Y. *Readings in Information Quality*. [S.l.]: MIT Information Quality (MITIQ) Program, 2005.

LEE, Y. et al. AIMQ: a methodology for information quality assessment. *Information&Management*, v. 40, p. 133–146, 2002.

LEE, Y. W. *Journey to data quality*. Cambridge, Mass.: MIT Press, 2006. ISBN 0262122871 (alk. paper) 9780262122870 (alk. paper). Disponível em: <http://www.loc.gov/catdir/toc/fy0705/2006043331.html>.

LEE, Y. W.; STRONG, D. M. Knowing-why about data processes and data quality. *Journal of Management Information Systems*, v. 20, n. 3, p. 13–39, 2003.

LEVENSHTEIN, V. I. *Binary codes capable of correcting deletions, insertions, and reversals*. 1966. 707–710 p.

MACLAURIN, J.; STERELNY, K. *What is biodiversity?* [S.l.]: University of Chicago Press, 2008. 217 p. ISBN 0226500829.

MALDONADO, C. et al. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? *Global Ecology and Biogeography*, v. 24, n. 8, p. n/a–n/a, 6 2015. ISSN 1466822X. Disponível em: <http://doi.wiley.com/10.1111/geb.12326>.

MAZUR, G. H. QFD for service industries. In: . [S.l.: s.n.], 1993.

MCGILVRAY, D. *Executing data quality projects: ten steps to quality data and trusted information*. Morgan Kaufmann, 2008. Disponível em: <http://books.google.com.br/books?hl=en{&}lr={&}id=2h5VnbGdh4cC{&}oi=fnd{&}pg=PP2{&}dq="Executing+data+quality+projects"{&}ots=98w9reg1d8{&}sig=Ae0XTE>.

MCZ. *About the Museum of Comparative Zoology - Harvard University*. 2016. Disponível em: <http://www.mcz.harvard.edu/about/index.html>.

MCZBASE. *MCZbase: The Database of the Zoological Collections - Museum of Comparative Zoology - Harvard University*. 2016. Disponível em: <http://mczbase.mcz.harvard.edu/>.

NICOLAOU, A. I.; MCKNIGHT, D. H. Perceived information quality in data exchanges: Effects on risk, trust, and intention to use. *Information Systems Research*, v. 17, n. 4, p. 332–351, 2006. ISSN 10477047.

NOVAK, J.; CAÑAS, A. *The theory underlying concept maps and how to construct and use them, Technical Report IHMC CmapTools 2006-01 Rev 01-2008, Florida Institute for Human and Machine Cognition*. [S.l.], 2008. 1–57 p. Disponível em: <http://www.swwhs.org/site/wp-content/uploads/2013/06/2013-APPsychologySummerReadings.pdf>.

NOY, N. F.; MCGUINNESS, D. L. *What is an ontology and why we need it*. 2016. Disponível em: <http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html>.

OTEGUI, J. et al. Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). *PloS one*, Public Library of Science, v. 8, n. 1, p. e55144, 1 2013. ISSN 1932-6203. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0055144>.

PETERSON, A. T. *BITC, Data Capture, What is Biodiversity Data - YouTube*. YouTube, 2014. Disponível em: <https://www.youtube.com/watch?v=2oaDV50YImI>.

PIPINO, L.; LEE, Y.; WANG, R. Data Quality Assessment. *Communications of the ACM*, v. 45, 2002. Disponível em: <d://junk/PipinoLeeWangCACMApr02-2212641024/PipinoLeeWangCACMApr02.pdf>.

PostgreSQL. *PostgreSQL: Documentation: 9.1: fuzzystrmatch*. 2015. Disponível em: <http://www.postgresql.org/docs/9.1/static/fuzzystrmatch.html>.

RASHID, H. et al. *Ecosystems and Human Well-Being: A Framework For Assessment (Millennium Ecosystem Assessment Series)*. Washington, DC: Island Press, 2003. v. 5. ISSN 00029513. ISBN 1597260401. Disponível em: <http://www.who.int/entity/globalchange/ecosystems/ecosys.pdf>.

REDMAN, T. *Data quality for the information age*. Boston: Artech House, 1996. ISBN 0890068836 (alk. paper).

REDMAN, T. *Data quality: the field guide*. Boston: Digital Press, 2001.

RICHARDS, K. et al. *A Beginner ' s Guide to Persistent Identifiers*. [s.n.], 2011. 33 p. ISBN 8792020143. Disponível em: <http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf>.

ROSE, P. Quality in services and services in quality. In: . [s.n.], 1994. p. 2–1. Disponível em: <http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=305447{&}con=yes{&}userTyp>.

SARAIVA, A. M.; CANHOS, D. A. L. Sistemas de informação e ferramentas computacionais para pesquisa, educação e disseminação do conhecimento sobre polinizadores. In: *Polinizadores do Brasil - contribuição e perspectivas para a biodiversidade, uso sustentável, conservação e serviços ambientais*. [S.l.]: EDUSP, 2012.

SAUERWEIN, E. et al. THE KANO MODEL: HOW TO DELIGHT YOUR CUSTOMERS. p. 313–327.

SearchCIO. *What is Total Quality Management (TQM)?* 2005. Disponível em: <http://searchcio.techtarget.com/definition/Total-Quality-Management>.

SHIELDS, P.; RANGARJAN, N. *A playbook for research methods: integrating conceptual frameworks and project management*. Stillwater: New Forums Press, 2013.

STEINHAGE, V. Automated Identification of Bee Species in Biodiversity Information Systems. 2003.

STRONG, D. M.; LEE, Y. W.; WANG, R. Y. Data quality in context. *Communications of the ACM*, v. 40, n. 5, p. 103–110, 1997.

STVILIA, B. et al. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, v. 58, n. 12, p. 1720–1733, 10 2007. ISSN 15322882. Disponível em: <http://doi.wiley.com/10.1002/asi.20652>.

VEIGA, A. K.; SARAIVA, A. M.; CARTOLANO, E. A. Data quality concepts and methods applied to biological species occurrence data. In: MILDORF, T.; JR, K. C. (Ed.). *ICT for agriculture, rural development and environment–where we are? Where we will go?* Prague: Czech Centre for Science and Society, 2012. Disponível em: <http://www.enorasis.eu/uploads/files/Modelling/ictbook-120613124719-phpapp02.pdf>.

VEIGA, A. K.; SARAIVA, A. M.; CARTOLANO, E. A. Data Quality control in biodiversity informatics: the case of species occurrence data. *IEEE Latin America Transactions*, v. 12, n. 4, p. 683–693, 2014. Disponível em: <http://www.ewh.ieee.org/reg/9/etrans/ieee/issues/vol12/vol12issue4June2014/20KochVeiga.htm>.

WAND, Y.; WANG, R. Y. *Anchoring data quality dimensions in ontological foundations*. 1996. 86–95 p.

WANG, R. A Product perspective on Total Data Quality Management. *Communications of the ACM*, v. 41, 1998. Disponível em: <d://junk/WangCACMFeb98-0514591048/WangCACMFeb98.pdf>.

WANG, R. et al. An Information Product Approach for Total Information Awareness. In: . [s.n.], 2003. Disponível em: <d://junk/IEEEAero03-3431019848/IEEEAero03.pdf>.

WANG, R.; KON, H.; MADNICK, S. Data Quality Requirements Analysis and Modeling. In: *Data Engineering*. [S.l.]: Citeseer, 1993. p. 670–677.

WANG, R. Y. et al. Manage Your Information as a Product. *Sloan Management Review*, v. 39, n. 05, p. 95–105, 1998.

WANG, R. Y.; REDDY, M. P.; KON, H. B. Toward quality data: An attribute-based approach. *Decision Support Systems*, v. 13, n. 3-4, p. 349–372, 1995.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, v. 12, n. 4, p. 5–33, 1996.

WANG, Y. R. Y.-Y. R.; ZIAD, M.; LEE, Y. W. *Data quality*. [S.l.]: Kluwer Academic Publishers, 2001. 167 p. ISBN 0792372158.

WIECZOREK, J. et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, Public Library of Science, v. 7, n. 1, p. e29715, 1 2012. ISSN 1932-6203. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0029715>.

WILSON, D.; REEDER, D. *Mammal species of the World. A taxonomic and geographic reference*. 1992. Disponível em: <http://www.nmnh.si.edu/msw/>.

Appendix

APPENDIX A – Mathematical Formalization

This document presents a mathematical representation to describe the concepts using a sets theory notation. Following the constructs developed in the conceptual model, we present a formal definition of the conceptual framework.

Mathematical Formalization Introduction

To define the conceptual framework, we present a mathematical representation to describe the concepts using a sets theory notation (Devlin 1993). Following the constructs developed in the conceptual model, we define a "domain" as a set of instances of the same concept (Wang *et al*. 1995). Thus, each fundamental or derived concept can be formally represented by a domain, represented by uppercase letters. An instance of a domain or a subset of instances (within "{}") of a domain can be formally represented by the same letter of the domain, but in lowercase.

To mathematically represent the derived concepts, we first need to represent the fundamental concepts as domains. Thus, let U be the domain for representing the concept Use Case, IE be the domain for the concept IE, D be the domain for the concept Dimension, C be the domain for the concept Criterion, E be the domain for the concept Enhancement, S be the domain for the concept Specification, M be the domain for the concept Mechanism and DR be the domain for Data Resource.

The relationship between fundamental concepts can be defined by functions or tuples (Devlin 1993). We use functions when the concept depends on other concepts, e.g. the function $VIE(u) = \{ie \mid ie \subset IE \wedge u \in U\}$ represent a subset of instances of the domain IE (which represent the fundamental concept IE) as a function of the parameter u (dependency), which is an instance of the domain U (which represent the fundamental concept Use Case). We use $\in$ when the variable is an unique instance of a particular domain and $\subset$ when the variable is a subset of instances of a particular domain. When the concept does not depend on another concept, we can define a new concept by using a tuple representation, e.g.: the domain $CD = \{cd \mid cd =< ie, d, rt > \wedge\ ie \in IE \wedge d \in D \wedge rt \in RT\}$ represents a set of instances of cd, where each instance of cd is a tuple composed by an instance of ie, d, and rt, where ie represents an instance of the domain IE, d represents an instance of the domain D and rt represents an instance of the domain RT (which will be described next).

Beyond those fundamental domains which represent the fundamental concepts of the conceptual model, it is necessary to use other complementary domains which are not fundamental concepts, but they can be seen as properties of them, used in order to compose more complete fundamental and derived concepts. These complementary domains are: Usage, Persistent Identifier, Resource Type, Value and Assertion. Thus, let US be the

domain for the concept Usage, which are encompassed by Use Cases; let ID be the domain for Persistent Identifier of Data Resources, which can be a DOI, LSID, URL, for instance (GBIF, 2011); let RT be the domain for the Resource Type (http://purl.org/dc/terms/type) which can assume only one of two values in the context of this paper: "Dataset" (http://purl.org/dc/dcmitype/Dataset) or "Single Record"; let V be the domain for Value of Data Resources, which represent the content of Data Resources; and let R be the domain for Assertions, which is obtained by Mechanisms when a validation, measurement or improvement is performed upon a Data Resource.

Using this notation we can represent the creation of instances of fundamental and derived concepts in a concise way, according to the definitions of derived concepts - described in the following subsections. To exemplify it, an instance of Dimension could be represented by this notation: $d_1 \in D$ where d_1 denote "completeness"; or an instance of Valuable IE for a particular Use Case could be represented by the notation: $vie(u_1) = \{ie_1, ie_2\}$ where ie_1 represents "coordinates", ie_2 represents "scientific name" and both are valuable to the u_1, which represents an Use Case to deal with Niche Modeling, for instance.

Mathematical Formalization A
USE CASE COVERAGE

Let UC be the domain for the Use Case Coverage, which define the scope delimitation of the Use Case. Each Use Case U encompass a subset of Usages US. Then, let UC domain be a subset of Usages us coveraged by an Use Case u. Thereby, UC domain is defined by:

$$UC(u) = \{ us \mid u \in U \wedge us \subset US\}$$

For example: Let u_1 denote a Use Case for dealing with Niche Modeling, us_1 denotes a Usage related to "Niche Modeling with MAXENT" (Merow *et al*. 2013) and us_2 a Usage related to "Niche Modeling with GARP" (Townsend Peterson *et al*. 2007). Thus, the Use Case Coverage for the u_1 can be defined as:

$$uc(u_1) = \{us_1, us_2\}$$

This instance means that there is a Use Case for dealing with Niche Modeling which encompasses MAXENT and GARP modeling techniques.

Mathematical Formalization B

DATA RESOURCE

Let DR be the domain for the Data Resource. Then, let DR domain be a set of tuples composed by a Persistent Identifier id, a Resource Type rt and a Value v. Notice that Resource Type can assume only one of two possible values, sr for Single Records or ds for Dataset. Thereby, DR domain is defined by:

$$DR = \{ dr \mid dr = < id, rt, v >, id \in ID, rt \in RT, (rt = sr \vee rt = ds) \wedge v \in V \}$$

For example: Let dr_1 denote a plant collection dataset (http://ipt.inpa.gov.br/resource.do?r=inpa-herbarium-collection) which has an id_1 equal "3cc6171e-8c52-4f65-ad7a-32c74e395f29", rt_1 equal "Dataset" and the v_1 equal a Darwin Core Archive with 251.744 records. Thus, the Data Resource dr_1 can be defined as:

$$dr_1 =< id_1, rt_1, v_1 >$$

This instance means that dr_1 is a Data Resource which represents the Dataset "3cc6171e-8c52-4f65-ad7a-32c74e395f29" which contains 251.744 records in Darwin Core Archive format.

Mathematical Formalization C

VALUABLE IE

Let VIE be the domain of valuable IE which represent a subset of instances of the domain IE which are valuable for a particular Use Case. Then, let $VIE(u)$ denote the subset of IE that are valuable for the purposes of the Use Case u. Thereby, VIE is defined by:

$$VIE(u) = \{ie \mid ie \subset IE \wedge u \in U\}$$

For example: Let u_1 be a Use Case for dealing with DQ for distribution modeling purposes and ie_1 be the IE for “coordinates” and ie_2 the IE for “taxon name”. Thus, the $vie(u_1)$ could be defined as:

$$vie(u_1) = \{ie_1, ie_2\}$$

This instance means that “coordinates” and “taxon name” are Valuable IE for the Use Case that deals with distribution modeling.

Mathematical Formalization D
CONTEXTUALIZED DIMENSION

Let CD be the domain for the derived concept Contextualized Dimension. Then, let CD be the set of tuples which denote a Dimension d in the context of an IE ie and the resource type rt, which the Dimension measures. Thereby, CD domain is defined by:

$$CD = \{ cd \mid cd =< ie, d, rt >,\ ie \in IE,\ d \in D \wedge rt \in RT \}$$

For example: Let ie_1 denote the IE “coordinates”, d_1 denote the Dimension “precision” and rt_1 denote the “Single Record” Resource Type. Thus, an element of Contextualized Dimension can be defined by:

$$cd_1 = < ie_1, d_1, rt_1 >$$

This instance means that the Contextualized Dimension cd_1 measures “coordinates precision of single records”.

Mathematical Formalization E
CONTEXTUALIZED CRITERION

Let CC be the domain for the derived concept Contextualized Criterion. Then, let CC be the set of tuples which denotes a Criterion c in the context of an IE ie and the Resource Type rt which this Criterion validates. Thereby, CC domain is defined by:

$$CC = \{ cc \mid cc = \ < ie, c, rt >,\ ie \in IE,\ c \in C \wedge rt \in RT \}$$

For example: Let ie_1 denote the IE “basis of record”, c_1 denote the Criterion “Value be in the controlled vocabulary” and rt_1 denote “Single Record” Resource Type. Thus, an element of CC could be defined as:

$$cc_1 = \ < ie_1, c_1, rt_1 >$$

This instance means that the Contextualized Criterion cc_1 validates the data compliance with the statement “The value of Basis of Records of single records must be in the controlled vocabulary”.

Mathematical Formalization F
ACCEPTABLE DQ MEASURE

Let AM be a subset of Contextualized Criteria which recommend a minimum measurement for a particular Contextualized Dimension. Then, let $AM(cd)$ denote a subset of Contextualized Criteria which establishes a minimum value for the Contextualized Dimension cd. Thereby, AM domain is defined by:

$$AM(cd) = \{cc \mid cd \in CD \wedge cc \subset CC\}$$

For example: Let cd_1 be a Contextualized Dimension “Coordinates Completeness for Dataset”, cc_1 denote the Contextualized Criteria “All of species occurrences records must have coordinates completed” and cc_2 denote the Contextualized Criteria “Dataset must have at least 90% of records with completed coordinates”. Thus, the $am(cd_1)$ could be defined as:

$$am(cd_1) = \{cc_1, cc_2\}$$

This instance means that the Contextualized Criteria “All of species occurrences records must have coordinates completed” and “Dataset must have at least 90% of records with completed coordinates” establish acceptable values for the Contextualized Dimension "Coordinates Completeness for Dataset". Thus the validation of these two Contextualized Criteria, cc_1 and cc_2, depend on the measurement value of Contextualized Dimension cd_1.

Mathematical Formalization G
CONTEXTUALIZED ENHANCEMENT

Let CE be the domain for the Contextualized Enhancement. Then, let CE be the set of tuples which denote an Enhancement e in the context of an IE ie and the Resource Type rt which this Enhancement improves. Thereby, CE domain is defined by:

$$CE = \{ ce \mid ce = \ < ie, e, rt >,\ ie \in IE,\ e \in E \wedge rt \in RT \}$$

For example: Let ie_1 denote the IE "taxon name", e_1 denote the Enhancement "Recommend valid and similar values" and rt_1 denote "Single Record" Resource Type. Thus, an element of CE could be defined as:

$$ce_1 = \ < ie_1, e_1, rt_1 >\}$$

This instance means that the Contextualized Enhancement ce_1 improves the quality of "Taxon names of single record by recommending valid and similar names".

Mathematical Formalization H
IMPROVEMENT TARGET

Let IT be a subset of Contextualized Dimension which may be improved and a subset of Contextualized Criteria which may be satisfied by a particular Contextualized Enhancement. Then, let $IT(ce)$ denote the union of a subset of Contextualized Dimensions which can be improved by Contextualized Enhancement ce with a subset of Contextualized Criteria which can be satisfied by the Contextualized Enhancement ce. Thereby, IT domain is defined by:

$$IT(ce) = \{cd \cup cc \mid cd \in CD, cc \in CC \wedge ce \in CE\}$$

For example: Let ce_1 be a Contextualized Enhancement "Recommending a coordinates based on the location description", cd_1 denote the Contextualized Dimension "Coordinates Completeness of Single Records" and cc_2 denote the Contextualized Criteria

“Dataset must have at least 90% of records georeferenced”. Thus, the $it(ce_1)$ could be defined as:

$$it(ce_1) = \{cd_1, cc_2\}$$

This instance means that the Contextualized Enhancement which recommend coordinates based on local description improves the "Coordinates Completeness of Single Records" and may comply with the Contextualized Criteria “Dataset must have at least 90% of records georeferenced”.

Mathematical Formalization I
DQ MEASUREMENT POLICY

Let MP be the domain for the DQ Measurement Policy for a particular Use Case. Then, let $MP(u)$ denote the set of Contextualized Dimensions used for measuring DQ in the context of the Use Case u. Thereby, MP domain is defined by:

$$MP(u) = \{cd \mid cd \subset CD \wedge u \in U\}$$

For example: Let u_1 denote a Use Case for dealing with DQ for distribution modeling purposes, $vie(u_1) = \{ie_1, ie_2\}$ where ie_1 denotes the IE for “coordinates” and ie_2 the IE for “taxon name”, d_1 denotes the Dimension “completeness”, d_2 denotes the Dimension “accuracy”, rt_1 denotes a Resource Type “Dataset” and rt_2 denotes a Resource Type “Single Record”. Thus, the $mp(u_1)$ could be defined as:

$$mp(u_1) = \{< ie_1, d_1, rt_2 >, < ie_1, d_1, rt_1 >, < ie_2, d_1, rt_1 >, < ie_2, d_2, rt_2 >\}$$

or

$$mp(u_1) = \{cd_1, cd_2, cd_3, cd_4\}$$

This instance means that the policy for measuring DQ in the context of distribution modeling, is comprised of the measurement of: “Coordinates Completeness of Single

Records", "Coordinates Completeness of Datasets", "Taxon Name Completeness of Datasets" and "Taxon Name Accuracy of Single Records".

Mathematical Formalization J
DQ VALIDATION POLICY

Let VP be the domain for the DQ Validation Policy for a particular Use Case. Then, let $VP(u)$ denotes the set of Contextualized Criteria used to validate DQ in the context of the Use Case u. Thereby, VP domain is defined by:

$$VP(u) = \{cc \mid cc \subset CC \wedge u \in U\}$$

For example: Let u_1 denote a Use Case for dealing with DQ for distribution modeling, cc_1 denote the Contextualized Criterion "Taxon nomenclature must be valid according a taxonomic authority" and cc_2 denote the Contextualized Criteria "At least 90% of records must be georeferenced". Thus, the $vp(u_1)$ can be defined as:

$$vp(u_1) = \{cc_1, cc_2\}$$

This instance means that the policy for validating DQ in the context of distribution modeling, is comprised of the following Contextualized Criteria: "Taxon nomenclature must be valid according a taxonomic authority" and "At least 90% of records must be georeferenced".

Mathematical Formalization K
DQ IMPROVEMENT POLICY

Let IP be the domain for the DQ Improvement Policy for a particular Use Case. Then, let $IP(u)$ denote the set of Contextualized Enhancements used for improving DQ in the context of the Use Case u. Thereby, IP domain is defined by:

$$IP(u) = \{ce \mid ce \subset CE \wedge u \in U\}$$

For example: Let u_1 denote a Use Case for dealing with DQ for distribution modeling, ce_1 denote the Contextualized Enhancement “Recommending similar and valid scientific name” and ce_2 denote the Contextualized Enhancement “Georeferencing using location description”. Thus, the $ip(u_1)$ can be defined as:

$$ip(u_1) = \{ce_1, ce_2\}$$

This instance means that the policy for improving DQ in the context of distribution modeling, is comprised of the Contextualized Enhancements: “Recommending similar and valid scientific name” and “Georeferencing using location description”.

Mathematical Formalization L

DQ Profile

Let DQP be the domain for the DQ Profile for a particular Use Case. Then, let $DQP(u)$ be the union of the DQ Measurement Policy $mp(u)$, DQ Validation Policy $vp(u)$ and DQ Improvement Policy $ip(u)$ in the context the Use Case u. Thereby, DQP domain is defined by:

$$DQP(u) = \{dqp \mid dqp = mp(u) \cup vp(u) \cup ip(u),\ mp \in MP,\ vp \in VP,\ ip \in IP \ \wedge\ u \in U\}$$

For example: Let u_1 denote a Use Case for dealing with DQ for distribution modeling, cd_1 denote the Contextualized Dimension “Coordinates Completeness of Single Records”, cd_2 denote the Contextualized Dimension “Coordinates Completeness of Datasets”, cd_3 denote the Contextualized Dimension “Taxon Name Completeness of Datasets”, cd_4 denote the Contextualized Dimension “Taxon Name Accuracy of Single Records”, cc_1 denote the Contextualized Criterion “Taxon nomenclature must be valid according a taxonomic authority”, cc_2 denote the Contextualized Criteria “At least 90% of records must be georeferenced”, ce_1 denote the Contextualized Enhancement “Recommending similar and valid scientific name” and ce_2 denote the Contextualized Enhancement “Georeferencing using location description”.

Let the DQ Measurement Policy for u_1 be $mp(u_1) = \{cd_1, cd_2, cd_3, cd_4\}$, DQ Validation Policy be $vp(u_1) = \{cc_1, cc_2\}$ and DQ Improvement Policy be $ip(u_1) = \{ce_1, ce_2\}$. Thus, the $dqp(u_1)$ can be defined as:

$$dqp(u_1) = \{mp(u_1),\ vp(u_1),\ ip(u_1)\}$$

or

$$dqp(u_1) = \{cd_1, cd_2, cd_3, cd_4, cc_1, cc_2, ce_1, ce_2\}$$

This instance means that the DQ Profile in the context of distribution modeling is comprised of the union of the DQ Measurement Policy, DQ Validation Policy and DQ Improvement Policy of the same Use Case. It represents all effort related DQ for assessing and managing the fitness for use in a specific context.

Mathematical Formalization M

MEASUREMENT METHOD

Let MM be the domain of Measurement Method which represents a subset of instances of the domain Specifications which are able to measure a particular Contextualized Dimension. Then, let $MM(cd)$ denote the subset of Specifications that describe methods to measure the Contextualized Dimension cd. Thereby, MM domain is defined by:

$$MM(cd) = \{s \mid s \subset S \wedge cd \in CD\}$$

For example: Let cd_1 be a Contextualized Dimension for "Completeness of Scientific Names of Datasets", s_1 denotes the Specification "`FOREACH(Dataset.records as r) IF(r.scientificName!= NULL) THEN completeness++; RETURN Dataset.records.size/completeness;` " and s_2 denotes the Specification "`SELECT count(*) AS completeness FROM records WHERE records.scientificName <> "";`". Thus, the $mm(cd_1)$ can be defined as:

$$mm(cd_1) = \{s_1, s_2\}$$

This instance means that the Contextualized Dimension "Completeness of Scientific Names of Datasets" can be measured by two different Specifications: "Iterate records and

calculate the proportion of records with scientific name different to null" and "Count the number of records with scientific name different to empty".

Mathematical Formalization N
VALIDATION METHOD

Let VM be the domain of Validation Method which represents a subset of instances of the domain Specifications which are able to validate a particular Contextualized Criterion. Then, let $VM(cc)$ denote the subset of Specifications that describe methods to validate the Contextualized Criterion cc. Thereby, VM domain is defined by:

$$VM(cc) = \{s \mid s \subset S \wedge cc \in CC\}$$

For example: Let cc_1 be a Contextualized Criterion "Basis of record of single records must be well-formed", s_1 the Specification "$BasisOfRecordValue \subset \{$FossilSpecimen, HumanObservation, PreservedSpecimen$\}$" and s_2 the Specification "GBIFControlledVocabularyList.contain(record.basisOfRecord)". Thus, the $vm(cc_1)$ can be defined as:

$$vm(cc_1) = \{s_1, s_2\}$$

This instance means that the criteria "Basis of record must be well-formed" can be validated by "checking if basis of record is in a specific list of values" or "checking if basis of record is in the list of GBIF controlled vocabulary".

Mathematical Formalization O
IMPROVEMENT METHOD

Let IM be the domain of Improvement Method which represents a subset of instances of the domain Specifications which are able to improve DQ according a particular Contextualized Enhancement. Then, let $IM(ce)$ denote the subset of Specifications that describe methods to improve according to the Contextualized Enhancement ce. Thereby, IM domain is defined by:

$$IM(ce) = \{s \mid s \subset S \wedge ce \in CE\}$$

For example: Let ce_1 be a Contextualized Enhancement "Recommending valid and similar names for taxon names of single records", s_1 denotes the Specification "`SELECT scientific_name FROM col_names order by levenshtein('value', scientific_name) limit 3;`" and s_2 denote the Specification "`SELECT scientific_name FROM col_names order by soundex('value', scientific_name) limit 2;`" (Postgresql 2015, Levenshtein 1965). Thus, the $im(ce_1)$ can be defined as:

$$im(ce_1) = \{s_1, s_2\}$$

This instance means that the enhancement "Recommending valid and similar names for taxon names of single records" can be performed by "recommending the three most similar names of *Catalog of Life* according the levenshtein algorithm (Levenshtein 1965)" or "recommending the two most similar names of *Catalog of Life* (CoL 2015) according the soundex algorithm" (Postgresql 2015).

Mathematical Formalization P

IMPLEMENTATION

Let I be the domain of Implementation which represents a subset of instances of the domain Mechanisms which implements a particular Specification. Then, let $I(s)$ denote the subset of Mechanisms that implement the Specification s. Thereby, I domain is defined by:

$$I(s) = \{m \mid m \subset M \wedge s \in S\}$$

For example: Let s_1 be a Specification "`SELECT scientific_name FROM col_names order by levenshtein('value', scientific_name) limit 3;`", m_1 the Mechanism "Web service of Biodiversity Data Digitizer (BDD) (Cartolano *et al*. 2010) for suggesting valid scientific names" and m_2 the Mechanism "DwC-A Validator 2.0 (GBIF 2015)". Thus, the $i(s_1)$ can be defined as:

$$i(s_1) = \{m_1, m_2\}$$

This instance means that the Specification “Recommend the three most similar names of Catalog of Life according the levenshtein algorithm” was implemented by a “Web service of the BDD” and by the “software DwC-A Validator 2.0”.

Mathematical Formalization Q

MECHANISM COVERAGE

Let MC be the domain of Mechanism Coverage which represents a subset of instances of the domain Specification which are implemented by a particular Mechanism. Then, let $MC(m)$ denote the subset of Specifications that are implemented by the Mechanism m. Thereby, MC domain is defined by:

$$MC(m) = \{s \mid s \subset S \wedge m \in M\}$$

For example: Let s_1 be the Specification “`SELECT scientific_name FROM col_names order by levenshtein('value', scientific_name) limit 3;`”, s_2 be the Specification "`SELECT COUNT(*) FROM dataset WHERE decimal_latitude<>'' AND decimal_longitude<>'';` " and m_1 denote a fictional Mechanism denominated “Biodiversity Data Quality Tool (BDQ-Tool)” which implements those Specifications. Thus, the $mc(m_1)$ can be defined as:

$$mc(m_1) = \{s_1, s_2\}$$

This instance means that the Mechanism “Biodiversity Data Quality Tool (BDQ-Tool)" implements the Specifications "Recommend the three most similar names of Catalog of Life according the levenshtein algorithm” and "Counting georeferenced records”.

Mathematical Formalization R

DQ MEASURE

Let DQM be the domain for the DQ Measure of a particular Data Resource. Then, let $DQM(dr)$ denote a set of DQ Measure concerning the Data Resource dr. Thereby, DQM domain is defined by:

$$DQM(dr) = \{dqm \mid dqm =< cd, s, m, r >, cd \in CD, s \in S, m \in M, r \in R \wedge dr \in DR\}$$

For example: Let dr_1 be a Data Resource which represents a dataset of species occurrences. Let cd_1 the Contextualized Dimension "Coordinates Numerical Precision of Datasets", s_1 the Specification "`FOREACH(dataset.records as r) {numericalPrecisionPerRecord += lower(split(r.coordinates.lat,".").size, split(r.coordinates.lat,".").size);} RETURN numericalPrecisionPerRecord/dataset.records.size;` ", m_1 the Mechanism "DwC-A Validator 2.0" and r_1 the Assertion "6.16352" . Thus, the $dqm(dr_1)$ can be defined as:

$$dqm(dr_1) = \{< cd_1, s_1, m_1, r_1 >\}$$

This instance means that the coordinates numerical precision of the dataset is "6.16352" and this value was assigned by the software DwC-A Validator 2.0 which calculated the value by the average of significant digits of each record of the dataset.

Mathematical Formalization S
DQ VALIDATION

Let DQV be the domain for the DQ Validation of a particular Data Resource. Then, let $DQV(dr)$ denote a set of DQ Validations concerning the Data Resource dr. Thereby, DQV domain is defined by:

$$DQV(dr) = \{dqv \mid dqv = < cc, s, m, r >, cc \in CC, s \in S, m \in M, r \in R \wedge dr \in DR\}$$

For example: Let dr_1 be a Data Resource which represents a species occurrence record. Let cc_1 denote the Contextualized Criterion "Geodetic Datum must be supplied", s_1

denote the Specification "`IF(record.geodetic_datum!="") RETURN "valid" ELSE RETURN "not valid";` ", m_1 denotes the Mechanism "Darwin Test" and r_1 denotes Assertion "Valid" . Thus, the $dqv(dr_1)$ can be defined as:

$$dqv(dr_1) = \{< cc_1 s_1, m_1, r_1 >\}$$

This instance means that there is a DQ Validation which declares that the Contextualized Criterion "Geodetic Datum must be supplied" is "Valid" for the specific species occurrence and this validation was performed by the software Darwin Test by checking if the field Geodetic Datum of the record was not empty.

Mathematical Formalization T
DQ IMPROVEMENT

Let DQI be the domain for the DQ Improvement of a particular Data Resource. Then, let $DQI(dr)$ denote a set of DQ Improvements concerning the Data Resource dr. Thereby, DQI domain is defined by:

$$DQI(dr) = \{dqi \mid dqi = < ce, s, m, r >, ce \in CE, s \in S, m \in M, r \in R \wedge dr \in DR\}$$

For example: Let dr_1 be a Data Resource which represents a species occurrence record. Let ce_1 denote the Contextualized Enhancement "Suggest a similar and valid scientific name", s_1 denote the Specification "`SELECT scientific_name FROM col_db WHERE col_db.valid=TRUE ORDER BY levenshtein(col_bd.scientific_name,record.scientific_name) LIMIT 2;`", m_1 denote the Mechanism "DwC-A Validator 2.0" and r_1 denote the Assertion ["Apis","Iris"] . Thus, the $dqi(dr_1)$ can be defined as:

$$dqi(dr_1) = \{< ce_1, s_1, m_1, r_1 >\}$$

This instance means that there is a correction recommendation for exchanging the current value of the scientific name by the Assertion "Apis" or "Iris". This Assertion was obtained by selecting the two most similar valid names based on the Levenshtein distance in the Catalog of Life database using the software DwC-A Validator 2.0.

Mathematical Formalization U

DQ ASSESSMENT

Let A be the domain for the DQ Assessment for a particular Data Resource. Then, let $A(dr)$ denote a set of DQ Measures, DQ Validations and DQ Improvements which assist data users assess the fitness for use of the Data Resource dr. Thereby, A domain is defined by:

$$A(dr) = \{dqm(dr) \cup dqv(dr) \cup dqi(dr) \mid dqm \in DQM,\ dqv \in DQV,\ dqi \in DQI \wedge dr \in DR\}$$

For example: Let dr_1 be a Data Resource which represents a record of species occurrence. Let dqm_1 denote a DQ Measure related to the “Coordinates Numerical Precision of Single Record” with Assertion equal “6.7” for dr_1, dqm_2 denote a DQ Measure related to the “Coordinates Completeness of Single Record” with Assertion equal “Complete” for dr_1, dqm_3 denote a DQ Measure related to the “Accuracy of Scientific Names of Single Record” with Assertion equal “Valid Name” for dr_1, dqv_1 denote a DQ Validation related to “Basis of record must be well-formed” with Assertion equal “Valid” for dr_1 and dqi_1 denote the DQ Improvement related to the procedure “Double check taxon identification by at least two different taxonomists” with Assertion “Checked by one regional expert in the taxon with high certainty and one non-expert in the taxon with some doubt” for dr_1. Thus, the $a(dr_1)$ can be defined as:

$$a(dr_1) = \{dqm_1, dqm_2, dqm_3, dqv_1, dqi_1\}$$

This instance represents five DQ Assertions related to three measures, one validation and one improvement, associated to a record of species occurrence. Based on the measures of the Contextualized Dimensions "Coordinates Precision", "Coordinates Completeness" and "Scientific Name Accuracy", the validation of the Contextualized Criterion “Basis of record is well-formed” and the improvement related to the Contextualized Enhancement "Double check taxon identification by at least two different taxonomists", data users are able to assess if the record of species occurrence is fit for use. In addition, this judgement can be refined if users take into account information about the Specifications and Mechanisms used to obtain those measures, validations and improvement.

Mathematical Formalization V
DQ MANAGEMENT BY QUALITY CONTROL

Let QC be the domain for the DQ Management by Quality Control for a particular Data Resource. Then, let $QC(dr)$ denote a set of DQ Validations and Improvements which assist the decrease of error and improvement of the fitness for use of the Data Resource dr. Thereby, QC domain is defined by:

$$QC(dr) = \{dqv(dr) \cup dqi(dr) \mid dqv \in DQV,\ dqi \in DQI \wedge dr \in DR\}$$

For example: Let dr_1 be a Data Resource which represents a record of species occurrence. Let dqv_1 denote the DQ Validation related to the Contextualized Criterion “Basis of record must be well-formed” with Assertion equal “Not valid” for dr_1. Let dqi_1 denote the DQ Improvement related to Contextualized Enhancement “Recommend a valid and similar taxon name” with Assertion “*Apis mellifera*” for dr_1. Thus, the $QC(dr_1)$ can be defined as:

$$qc(dr_1) = \{dqv_1, dqi_1\}$$

This instance means that the quality of the record of species occurrence can be improved by double checking the invalid Basis of Record and accepting the correction recommendation with the value “*Apis mellifera*” for the taxon name.

Mathematical Formalization W
DQ MANAGEMENT BY QUALITY ASSURANCE

Let QA be the domain for the DQ Management by Quality Assurance for a particular Data Resource. Then, let $QA(dr)$ denotes a set of DQ Validations which have satisfactory Assertion to assure that the Data Resource dr is fit for use. Thereby, QA domain is defined by:

$$QA(dr) = \{dqv(dr) \mid dqv \in DQV \wedge dr \in DR\}$$

For example: Let dr_1 be a Data Resource which represents a record of species occurrence. Let dqv_1 denote the DQ Validation related to the Contextualized Criterion “Scientific name must be valid” with Assertion equal “Valid” for dr_1 and dqv_2 denote the DQ Validation related to the Contextualized Criterion “Coordinates must be valid” with Assertion equal “Valid” for dr_1. Thus, the $QA(dr_1)$ can be defined as:

$$qa(dr_1) = \{dqv_1, dqv_2\}$$

This instance means that the record of species occurrence satisfies all of the requirements defined by qa, then the quality of the record is assured.

Bibliographic Reference

Cartolano EA, Saraiva AM, Veiga AK, Krobath DB, Saraiva LGP, Tavares G. Biodiversity Data Digitizer. In The Proceedings of TDWG: Provisional Abstracts of the 2010 Annual Conference of the Taxonomic Databases Working Group. 2010.

CoL. User Guide | Catalogue of Life [Internet]. 2015 [cited 2015 Aug 21]. Available from: http://www.catalogueoflife.org/content/user-guide

Devlin KJ. The Joy of Sets: Fundamentals of Contemporary Set Theory, 2nd edition, Springer-Verlag, New York, NY, 1993.

GBIF. dwca-validator. GitHub repository. 2015 [cited 2015 June 14]. Available from: https://github.com/gbif/dwca-validator

GBIF. A Beginner’s Guide to Persistent Identifiers. Authors: Richards K, White R, Nicolson N, Pyle R. Copenhagen: Global Biodiversity Information Facility. 2011; ISBN: 87-92020-14-3. Available from: http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf

Levenshtein V. Binary codes cable of correcting spurious insertions and deletions of ones. Problems of Information Transmission. 1965; 1:8-17.

Merow C, Smith MJ, Silander JA. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography (Cop) [Internet]. 2013 Oct 18 [cited 2014 Jul 11];36(10):1058–69. Available from: http://doi.wiley.com/10.1111/j.1600-0587.2013.07872.x

PostgreSQL. PostgreSQL: Documentation: 9.1: fuzzystrmatch [Internet]. 2015 [cited 2015 Aug 21]. Available from: http://www.postgresql.org/docs/9.1/static/fuzzystrmatch.html

Wang R, Reddy M, Kon H. Toward quality data: An attribute- based approach. Journal of Decision Support Systems. 1995; vol. 13, no. 3-4. pp. 349-372.

Townsend Peterson A, Papeş M, Eaton M. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography (Cop) [Internet]. 2007 Aug 31 [cited 2015 Jun 9];30(4):550–60. Available from: http://doi.wiley.com/10.1111/j.0906-7590.2007.05102.x

APPENDIX B – Study Case

The purpose of this document is to present a proof of concept that allow us to stress all the concepts of the framework, and show how it can be used to assess the fitness for use of biodiversity data.

FRAMEWORK VALIDATION - PROOF OF CONCEPT

The purpose of this document is to present a proof of concept that allows us to stress all the concepts of the framework, and show how it can be used to assess the fitness for use of biodiversity data.

This document is organized into three sections: DQ Needs, DQ Solutions and DQ Report, and each section is organized into two subsections: Fundamental Concepts and Derived Concepts. In summary, this document presents:

1. **DQ Needs**: Definition of the meaning of "fitness for use" for a specific context. In this section we define a scope delimitation with a Use Case, a set of IEs that are valuable for this Use Case and a set of Contextualized Dimensions, Contextualized Criteria and Contextualized Enhancements necessary for the assessment and management of the fitness for use for the defined Use Case.
    1.1. **Fundamental Concepts**: Definition of instances of the fundamental concepts Use Case, IE, Dimension, Criterion and Enhancement and properties Usages and Resource Type. No relationship between concepts and properties are defined in this subsection. It serves as a fundamental concepts repository that provides concepts instances to be combined for defining new derived concepts instances.
    1.2. **Derived Concepts**: Definition of instances of the derived concepts Valuable IE, Contextualized Dimension, Contextualized Criterion, Contextualized Enhancement, Acceptable DQ Measures, Improvement Target, DQ Measurement Policy, DQ Validation Policy, DQ Improvement Policy and DQ Profile. Those concepts instances are defined by combining fundamental and other derived concepts instances.
2. **DQ Solutions**: Definition of methods and tools for meeting the defined DQ Needs. In this section we define a set of Specifications as methods for measuring, validating and improving all the previously defined Contextualized Dimensions, Contextualized Criteria and Contextualized Enhancements, respectively, and a set of Mechanisms that implement those Specifications.
    2.1. **Fundamental Concepts**: Definition of instances of the fundamental concepts Specification and Mechanism. No relationship between concepts are defined in this subsection. It serves as a fundamental concepts repository that provides concepts instances to be combined for defining new derived concepts instances.
    2.2. **Derived Concepts**: Definition of instances of the derived concepts Measurement Methods, Validation Methods, Improvement Methods, Implementations and Mechanisms Coverage. Those concepts instances are defined by combining fundamental and other derived concepts instances from DQ Solutions and DQ Needs concept classes.
3. **DQ Report**: Definition of a set of DQ measures, validations and improvements assertions assigned to a dataset or to a single record. In this section we presents how two Data Resources can have their fitness for use

assessed and managed with a set of DQ Measures, DQ Validations and DQ Improvements generated according to the DQ Solutions and DQ Needs defined previously.

3.1. **Fundamental Concept**: Definition of instances of the fundamental concept Data Resource. No relationship between concepts are defined in this subsection. It serves as a fundamental concepts repository that provides concepts instances to be combined for defining new derived concepts instances.

3.2. **Derived Concepts**: Definition of instances of the derived concepts DQ Measure, DQ Validation, DQ Improvement, DQ Assessment, DQ Management by Quality Control and DQ Management by Quality Assurance. Those concepts instances are defined by combining fundamental and derived concepts instances from DQ Report, DQ Solutions and DQ Needs concept classes.

1. DQ Needs

This section covers the definition of fundamental and derived concepts related to DQ requirements in a context delimited by the concept Use Case described in Table A. The DQ Needs concepts describe how data must be addressed to enable the assessment and management of fitness for use of data according to a particular Use Case context. In this case, we will address DQ Needs related to curation of biological collections.

1.1. Fundamental Concepts

This subsection describes DQ Needs concepts instances that are independent of context and of the relationships with other concepts. In order to allow the concepts instances to be reused for different purposes and contexts, they are listed here (like a repository of fundamental concepts instances) without any relationship with other concept instances. Such relationships will be presented later in Subsection 1.2.

Table A presents a Use Case that defines a scope delimitation concerning DQ. In the case of this proof of concept a single use case was defined for the sake of brevity of the text and ease of understanding of the process and its results.

Table A - Use Case

ID	Label	Description
u_1	`Curation of Biological Collections`	This Use Case is related to the curation tasks required to keep biological collections data complete, consistent, accurate and precise to ensure species occurrence data is fit for general use.

Table B presents the Usages, that are specific tasks performed with data by data users.

Table B - Usages

ID	Label	Description

us_1	General uses of occurrence data	General uses of data that comprise the location and date of a taxon event.
us_2	Geospatial distribution of species	General uses of species occurrences data for species distribution visualization and geospatial analysis.

Table C presents a set of IEs that represent elements of the real world.

Table C - Information Elements

ID	Label	Description
ie_1	Coordinates	A geospatial location on Earth. **Composed of**: dwc:decimalLatitude, dwc:decimalLongitude
ie_2	Collected Date	A day, month and year. **Composed of**: dwc:eventDate
ie_3	SciName	The scientific name of a species. **Composed of**: dwc:scientificName, dwc:authorshipName
ie_4	Occurrence	An occurrence comprises the location and date where a species sample was collected or observed and its identification with its scientific name. **Composed of:** ie_1 , ie_2 , ie_3

Table D presents a set of fundamental measurable quality aspects of data.

Table D - Dimensions

ID	Label	Description
d_1	Precision	Measure of the exactness, resolution or granularity of data.
d_2	Completeness	Measure of the extent to which data is present and suitable for use for a specific task.
d_3	Accuracy	Measure of the correctness or absence of errors in the data.
d_4	Consistency	Measure of the absence of contradictions in the data.

Table E presents a set of statements that describe fundamental principles or standards by which data are judged regarding quality.

Table E - Criteria

ID	Statement
c_1	DQ measure must have acceptable value.
c_2	Data must comply with standard format requirement.

Table F presents a set of statements that describe activities required to improve DQ.

Table F - Enhancements

ID	Statement
e_1	Recommend correction based on associated data.
e_2	Recommend correction based on standard(s).
e_3	Recommend correction based on authorities' data source(s).
e_4	Correct data based on correction recommendations.
e_5	Disregard unsuitable data.

Although the Resource Type is a property inherent to the concept Data Resource, it is used in several contextualized concepts of DQ Needs class. Therefore, we present the Resource Type instances in this Subsection, in Table G.

Table G - Resource Types

ID	Label	Description
rt_1	Single Record	Data is a single record.
rt_2	Dataset	Data is a set of records.

1.2. Derived Concepts

This subsection describes DQ Needs concepts instances that are defined by the relationships between fundamental concepts instances or other derived concepts instances. These relationships define new concepts instances with their own meaning.

Table H presents a definition of the relationship between a Use Cases concept and a set of Usages (properties).

Table H - Use Case Coverage

Definition	Description

$uc(u_1) = \{us_1, us_1\}$	The coverage of the Use Case `Curation of Biological Collections` encompasses the Usages `General uses of occurrence data` and `Geospatial distribution of species`.

Table I presents a definition of the relationship between a Use Case and a set of IEs. It defines a set of IEs that are valuable for the purposes of a particular Use Case.

Table I - Valuable IE

Definition	Description
$vie(u_1) = \{ie_1, ie_2, ie_3, ie_4\}$	`Coordinates`, `Collected Date`, `SciName` and `Occurrence` are considered valuable for `Curation of Biological Collections`. These IE must have quality in order for a data (`Single Record` or `Dataset`) to be fit for use; accordingly those IE must be the target of DQ Assessment and DQ Management in order to measure, validate and improve the quality of data.

The same fundamental Dimension can have different meanings when it is related to different IEs and Resource Types. Table J presents a set of measurable quality aspects where each Dimension measures the quality of a specific IE and Resource Type. Each instance in the table defines a quantitative or qualitative quality measurement for a specific IE and Resource Type.

Table J - Contextualized Dimensions

Definition	Label	Description
$cd_1 = < d_1, ie_1, rt_1 >$	`Single Record Coordinates Numerical Precision`	Measure the numerical `Precision` (d_1) of `Coordinates` (ie_1) based on the number of decimals of `Coordinates` of a `Single Record` (rt_1). Fewer digits mean worse `Precision` of `Coordinates`, but many digits does not necessarily mean better real `Precision` of `Coordinates`.
$cd_2 = < d_2, ie_1, rt_1 >$	`Single Record Coordinates Completeness`	Measure the `Completeness` based on the presence of all values for `Coordinates`.
$cd_3 = < d_4, ie_1, rt_1 >$	`Single Record Coordinates Consistency`	Measure the `Consistency` of `Coordinates` with the associated country, state/province, county and locality of record.

$cd_4 = < d_2, ie_1, rt_2 >$	Dataset Coordinates Completeness	Measure the Completeness based on the proportion of records that have all values of Coordinates supplied.
$cd_5 = < d_4, ie_1, rt_2 >$	Dataset Coordinates Consistency	Measure the Consistency based on the proportion of records that have Coordinates values consistent with associated country, state/province, county and locality of record.
$cd_6 = < d_1, ie_1, rt_2 >$	Dataset Coordinates Numerical Precision	Measure the Precision based on the average number of decimal digits of Coordinates of records of Dataset. Fewer digits mean worse Precision, but many digits does not necessarily mean better true Coordinates Precision.
$cd_7 = < d_2, ie_3, rt_1 >$	Single Record SciName Completeness	Measure the Completeness based on the presence of all values for SciName.
$cd_8 = < d_3, ie_3, rt_1 >$	Single Record SciName Accuracy	Measure the Accuracy based on the presence of correct values of SciName according to nomenclature authorities.
$cd_9 = < d_2, ie_3, rt_2 >$	Dataset SciName Completeness	Measure the Completeness based on the proportion of records that have all values supplied for SciName.
$cd_{10} = < d_3, ie_3, rt_2 >$	Dataset SciName Accuracy	Measure the Accuracy based on the proportion of records that have presence of correct values of SciName according to nomenclature authorities.
$cd_{11} = < d_2, ie_2, rt_1 >$	Single Record Collected Date Completeness	Measure the Completeness based on the presence of value for Collected Date.
$cd_{12} = < d_4, ie_2, rt_1 >$	Single Record Collected Date Consistency	Measure the Consistency of Collected Date value with modified date and with the life span of the collector.
$cd_{13} = < d_2, ie_2, rt_2 >$	Dataset Collected Date Completeness	Measure the Completeness based on the proportion of records that have a value supplied for Collected Date.

$cd_{14} = < d_4, ie_2, rt_2 >$	Dataset Collected Date Consistency	Measure the Consistency based on the proportion of records which Collected Date value is consistent with modified date and with the life span of the collector.
$cd_{15} = < d_2, ie_4, rt_1 >$	Single Record Occurrence Completeness	Measure the Completeness based on the presence of all values for Coordinates, SciName and Collected Date.
$cd_{16} = < d_2, ie_4, rt_2 >$	Dataset Occurrence Completeness	Measure the Completeness based on the proportion of records in a Dataset that have Coordinates, SciName and Collected Date values supplied.
$cd_{17} = < d_3, ie_4, rt_1 >$	Single Record Occurrence Accuracy	Measure the Accuracy based on the Consistency measures of Coordinates and Collected Date and the Accuracy measure of SciName.
$cd_{18} = < d_3, ie_4, rt_2 >$	Dataset Occurrence Accuracy	Measure the Accuracy based on the proportion of records in a Dataset that have consistent Coordinates and Collected Date and accurate SciName .

The same fundamental Criterion can have different meanings when it is related to different IEs and Resource Types. Table K presents a set of statements that describe principles or standards by which a specific IE of data, in the context of a determined Resource Type, is judged regarding quality.

Table K - Contextualized Criteria

Definition	Statement
$cc_1 = < c_1, ie_1, rt_1 >$	Coordinates Numerical Precision must be higher than 4
$cc_2 = < c_1, ie_1, rt_1 >$	Coordinates must be complete
$cc_3 = < c_1, ie_1, rt_1 >$	Coordinates must be consistent
$cc_4 = < c_1, ie_1, rt_2 >$	All records in a Dataset must have complete Coordinates
$cc_5 = < c_1, ie_1, rt_2 >$	All records in a Dataset must have consistent Coordinates
$cc_6 = < c_1, ie_1, rt_2 >$	Average value of Coordinates Numerical Precision within a Dataset must be higher than 4
$cc_7 = < c_1, ie_3, rt_1 >$	SciName must be complete

$cc_8 = < c_1, ie_3, rt_1 >$	SciName must be accurate
$cc_9 = < c_1, ie_3, rt_2 >$	All records in a Dataset must have complete SciName
$cc_{10} = < c_1, ie_3, rt_2 >$	All records in a Dataset must have accurate SciName
$cc_{11} = < c_1, ie_2, rt_1 >$	Collected Date must be complete
$cc_{12} = < c_1, ie_2, rt_1 >$	Collected Date must be consistent
$cc_{13} = < c_1, ie_2, rt_2 >$	All records in a Dataset must have complete Collected Date
$cc_{14} = < c_1, ie_2, rt_2 >$	All records of Dataset must have consistent Collected Date
$cc_{15} = < c_1, ie_4, rt_1 >$	Occurrence must be complete
$cc_{16} = < c_1, ie_4, rt_2 >$	All records in a Dataset must have Occurrence complete
$cc_{17} = < c_2, ie_2, rt_1 >$	Collected Date must follow ISO 8601 standard format
$cc_{18} = < c_1, ie_4, rt_1 >$	Occurrence must be accurate
$cc_{19} = < c_1, ie_4, rt_2 >$	All records in a Dataset must have accurate Occurrence

A Contextualized Criterion may define an acceptable measure for a Contextualized Dimension. Table L presents a set of definitions of Contextualized Criteria that define an acceptable measure for a specific Contextualized Dimension.

Table L - Acceptable DQ Measures

Definition	Description
$am(cd_1) = \{cc_1\}$	The Contextualized Criterion cc_1 defines an acceptable measure for the Contextualized Dimension cd_1
$am(cd_2) = \{cc_2\}$	The Contextualized Criterion cc_2 defines an acceptable measure for the Contextualized Dimension cd_2
$am(cd_3) = \{cc_3\}$	The Contextualized Criterion cc_3 defines an acceptable measure for the Contextualized Dimension cd_3
$am(cd_4) = \{cc_4\}$	The Contextualized Criterion cc_4 defines an acceptable measure for the Contextualized Dimension cd_4
$am(cd_5) = \{cc_5\}$	The Contextualized Criterion cc_5 defines an acceptable measure for

	the Contextualized Dimension cd_5
$am(cd_6) = \{cc_6\}$	The Contextualized Criterion cc_6 defines an acceptable measure for the Contextualized Dimension cd_6
$am(cd_7) = \{cc_7\}$	The Contextualized Criterion cc_7 defines an acceptable measure for the Contextualized Dimension cd_7
$am(cd_8) = \{cc_8\}$	The Contextualized Criterion cc_8 defines an acceptable measure for the Contextualized Dimension cd_8
$am(cd_9) = \{cc_9\}$	The Contextualized Criterion cc_9 defines an acceptable measure for the Contextualized Dimension cd_9
$am(cd_{10}) = \{cc_{10}\}$	The Contextualized Criterion cc_{10} defines an acceptable measure for the Contextualized Dimension cd_{10}
$am(cd_{11}) = \{cc_{11}\}$	The Contextualized Criterion cc_{11} defines an acceptable measure for the Contextualized Dimension cd_{11}
$am(cd_{12}) = \{cc_{12}\}$	The Contextualized Criterion cc_{12} defines an acceptable measure for the Contextualized Dimension cd_{12}
$am(cd_{13}) = \{cc_{13}\}$	The Contextualized Criterion cc_{13} defines an acceptable measure for the Contextualized Dimension cd_{13}
$am(cd_{14}) = \{cc_{14}\}$	The Contextualized Criterion cc_{14} defines an acceptable measure for the Contextualized Dimension cd_{14}
$am(cd_{15}) = \{cc_{15}\}$	The Contextualized Criterion cc_{15} defines an acceptable measure for the Contextualized Dimension cd_{15}
$am(cd_{16}) = \{cc_{16}\}$	The Contextualized Criterion cc_{16} defines an acceptable measure for the Contextualized Dimension cd_{16}
$am(cd_{17}) = \{cc_{18}\}$	The Contextualized Criterion cc_{18} defines an acceptable measure for the Contextualized Dimension cd_{17}
$am(cd_{18}) = \{cc_{19}\}$	The Contextualized Criterion cc_{19} defines an acceptable measure for the Contextualized Dimension cd_{18}

The same fundamental Enhancement can have different meanings when it is related to different IEs and Resource Types. Table M presents a set of statements that describe activities required to improve DQ of an IE in the context of a determined Resource Type.

Table M - Contextualized Enhancements

Definition	Statement
$ce_1 = < e_3, ie_3, rt_1 >$	Recommendation of SciName based on nomenclature

	authorities
$ce_2 = < e_2, ie_2, rt_1 >$	Recommendation of Collected Date based on ISO 8601 standard
$ce_3 = < e_1, ie_1, rt_1 >$	Recommendation of Coordinates based on the associated country, state/province, county and locality
$ce_4 = < e_4, ie_1, rt_2 >$	Accept all recommendations for each Single Record in Dataset
$ce_5 = < e_5, ie_1, rt_2 >$	Disregard all records that are not compliant with all DQ Criteria

A Contextualized Enhancement might improve the measures of a set of Contextualized Dimensions and might make the set of Contextualized Criteria compliant. Table N presents a set of definitions of Contextualized Dimensions that might be improved, and a set of Contextualized Criteria that might be made compliant with the application of a specific Contextualized Enhancement.

Table N - Improvement Target

Definition	Description
$it(ce_1) = \{cd_8, cd_{10}, cd_7, cd_9, cd_{17}, cd_{18}, cd_{15}, cd_{16}, cc_8, cc_{10}, cc_7, cc_9, cc_{18}, cc_{19}, cc_{15}, cc_{16}\}$	The Contextualized Enhancement ce_1 might enable the improvement of the measures of the Contextualized Dimensions cd_7, $cd_8, cd_9, cd_{10}, cd_{15}, cd_{16}, cd_{17}$ and cd_{18} and might make the Contextualized Criteria cc_7, cc_8, cc_9, cc_{10}, cc_{15}, cc_{16}, cc_{18} and cc_{19} compliant.
$it(ce_2) = \{cd_{11}, cd_{12}, cd_{13}, cd_{14}, cd_{15}, cd_{17}, cd_{16}, cd_{18}, cc_{11}, cc_{12}, cc_{17}, cc_{13}, cc_{14}, cc_{15}, cc_{18}, cc_{16}, cc_{19}\}$	The Contextualized Enhancement ce_2 might enable the improvement of the measures of the Contextualized Dimensions cd_{11}, $cd_{12}, cd_{13}, cd_{14}, cd_{15}, cd_{16}, cd_{17}$ and might make the Contextualized Criteria cc_{11}, cc_{12}, cc_{13}, cc_{14}, cc_{15}, cc_{16}, cc_{18} and cc_{19} compliant.
$it(ce_3) = \{cd_2, cd_3, cd_4, cd_5, cd_{15}, cd_{17}, cd_{16}, cd_{18}, cc_2, cc_3, cc_4, cc_5, cc_{15}, cc_{18}, cc_{16}, cc_{19}\}$	The Contextualized Enhancement ce_3 might enable the improvement of the measures of the Contextualized Dimensions cd_2, $cd_3, cd_4, cd_5, cd_{15}, cd_{16}, cd_{17}$ and cd_{18} and might make the Contextualized Criteria cc_2, cc_3, cc_4, cc_5, cc_{15}, cc_{16}, cc_{18} and cc_{19} compliant.
$it(ce_4) = \{cd_2, cd_3, cd_4, cd_5, cd_7, cd_8, cd_9, cd_{10}, cd_{11}, cd_{12}, cd_{13}, cd_{14}, cd_{15}, cd_{16}, cd_{17}, cd_{18}, cc_2, cc_3, cc_4, cc_5, cc_7, cc_8, cc_9, cc_{10}, cc_{11}, cc_{12}, cc_{13}, cc_{14}, cc_{15}, cc_{16}, cc_{17}, cc_{18}, cc_{19}\}$	The Contextualized Enhancement ce_4 might enable the improvement of the measures of the Contextualized Dimensions cd_2, $cd_3, cd_4, cd_5, cd_7, cd_8$, cd_9, cd_{10}, cd_{11}, cd_{12}, cd_{13}, cd_{14}, cd_{15}, cd_{16}, cd_{17} and cd_{18} and might make the Contextualized Criteria cc_2, cc_3, cc_4, cc_5, cc_7, cc_8, cc_9, cc_{10}, cc_{11}, cc_{12}, cc_{13}, cc_{14}, cc_{15}, cc_{16}, cc_{17}, cc_{18} and cc_{19} compliant.

$it(ce_5) = \{cd_4, cd_5, cd_6, cd_{13}, cd_{14}, cd_9, cd_{10}, cd_{16}, cd_{18}, cc_4, cc_5, cc_6, cc_9, cc_{10}, cc_{13}, cc_{14}, cc_{16}, cc_{19}\}$	The Contextualized Enhancement ce_5 might enable the improvement of the measures of the Contextualized Dimensions cd_4, $cd_5, cd_6, cd_9, cd_{10}, cd_{13}, cd_{14}$, cd_{16} and might make the Contextualized Criteria cc_4, cc_5, cc_6, cc_9, cc_{10}, cc_{13}, cc_{14}, cc_{16} and cc_{19} compliant.

Table O presents a set of Contextualized Dimensions that are used to measure the DQ in the context of a specific Use Case.

Table O - DQ Measurement Policy

Definition	Description
$mp(u_1) = \{cd_1, ..., cd_{18}\}$	The set $\{cd_1, ..., cd_{18}\}$ of Contextualized Dimensions defines what should be measured in order to assist fitness for use assessment in the context of `Curation of Biological Collections` (u_1).

Table P presents a set of Contextualized Criteria that are used to validate the DQ in the context of a specific Use Case.

Table P - DQ Validation Policy

Definition	Description
$vp(u_1) = \{cc_1, ..., cc_{19}\}$	The set $\{cc_1, ..., cc_{19}\}$ of Contextualized Criteria defines what should be validated in order to assist the fitness for use assessment and management in the context of `Curation of Biological Collections` (u_1).

Table Q presents a set of Contextualized Enhancements that are used to improve the DQ in the context of a specific Use Case.

Table Q - DQ Improvement Policy

Definition	Description
$ip(u_1) = \{ce_1, ..., ce_5\}$	The set $\{ce_1, ..., ce_5\}$ of Contextualized Enhancement defines what should be improved in order to assist the fitness for use assessment and management in the context of `Curation of Biological Collections` (u_1).

Table R presents a set of Contextualized Dimensions, Contextualized Criteria and Contextualized Enhancements that are used for the assessment and management of the fitness for use of data in the context of a specific Use Case.

Table R - DQ Profile

Definition	Description

$dqp(u_1) = \{cd_1, ..., cd_{18}, cc_1, ..., cc_{19}, ce_1, ..., ce_5\}$	The set $\{cd_1, ..., cd_{18}, cc_1, ..., cc_{19}, ce_1, ..., ce_5\}$ of Contextualized Dimensions, Contextualized Criteria and Contextualized Enhancement, respectively, defines what should be measured, validated and improved in order to assist the fitness for use assessment and management in the context of `Curation of Biological Collections` (u_1).

2. DQ Solutions

This section covers fundamental and derived concepts related to methods, tools and any other resource used or effort made to meet DQ Needs. The DQ Solution concepts describe how DQ Needs can be addressed in terms of implementation to enable the assessment and management of fitness for use of data. In this case, we will specifically address DQ Solutions to meet DQ Needs described in Section 1.

2.1. Fundamental Concepts

This subsection describes DQ Solutions concepts instances that are independent of context and of relationships with other concepts. In order to allow the concepts instances to be reused for different purposes and contexts, they are listed here (like a repository of fundamental concepts instances) without any relationship with other concept instances, such relationships will be presented later in the Subsection 2.2.

Table S presents brief descriptions in natural language of methods of how to perform some deed of measurement, validation and improvement of data.

Table S - Specifications

ID	Description
s_1	**Method description**: Check if values (disregarding extra spaces) for both dwc:decimalLatitude and dwc:decimalLongitude are different to zero and different to empty. Applicable to Darwin Core based single records.
s_2	**Method description**: Check if dwc:decimalLatitude and dwc:decimalLongitude values are consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality. Applicable to Darwin Core based single records. **Georeference sources**: GeoLocate | Land data from Natural Earth | Country boundary data from GeoCommunity.
s_3	**Method description**:

	Calculate the average of the number of characters after "." of dwc:decimalLatitude and dwc:decimalLongitude. Applicable to Darwin Core based single records.
s_4	**Method description**: Calculate the proportion of records in dataset with values (disregarding extra spaces) different to zero and different to empty for both dwc:decimalLatitude and dwc:decimalLongitude. Applicable to datasets of Darwin Core based records.
s_5	**Method description**: Calculate the proportion of records dwc:decimalLatitude and dwc:decimalLongitude values consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality. Applicable to datasets of Darwin Core based records. **Georeference sources**: GeoLocate | Land data from Natural Earth | Country boundary data from GeoCommunity.
s_6	**Method description**: Calculate the mean of the average number of characters after "." of dwc:decimalLatitude and dwc:decimalLongitude of records of dataset. Applicable to datasets of Darwin Core based records.
s_7	**Method description**: Check if values (disregarding extra spaces) for both dwc:scientificName and dwc:scientificNameAuthorship are different to empty. Applicable to Darwin Core based single records.
s_8	**Method description**: Check against nomenclature authorities if there is an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values. Applicable to Darwin Core based single records. **Nomenclatural authority sources**: Catalog of Life | Global Name Resolver | Global Name Index | GBIF CheckListBank Backbone | IPNI | Index Fungorum | WoRMS.
s_9	**Method description**: Calculate the proportion of records in dataset with values (disregarding extra spaces) different to empty for both dwc:scientificName and dwc:scientificNameAuthorship. Applicable to datasets of Darwin Core based records.
s_{10}	**Method description**: Calculate the proportion of records with an exact match of both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclatural authorities. Applicable to datasets of Darwin Core based records.

	Nomenclatural authority sources: Catalog of Life | Global Name Resolver | Global Name Index | GBIF CheckListBank Backbone | IPNI | Index Fungorum | WoRMS.
s_{11}	**Method description**: Check if value (disregarding extra spaces) for dwc:eventDate is different to empty. Applicable to Darwin Core based single records.
s_{12}	**Method description**: Check if dwc:eventDate value is correctly formatted according to ISO 8601 standard and is before dwc:modified and within life span of dwc:collectedBy. Applicable to Darwin Core based single records. **Life span of collectors sources**: Harvard List of Botanists | FilteredPush Entomologists List.
s_{13}	**Method description**: Calculate the proportion of records in dataset that have the dwc:eventDate value (disregarding extra spaces) different to empty. Applicable to datasets of Darwin Core based records.
s_{14}	**Method description**: Calculate the proportion of dwc:eventDate value records that are correctly formatted according to ISO 8601 standard and that are before dwc:modified and within life span of collector. Applicable to datasets of Darwin Core based records. **Life span of collectors sources**: Harvard List of Botanists | FilteredPush Entomologists List.
s_{15}	**Method description**: Check if values (disregarding extra spaces) for both dwc:decimalLatitude and dwc:decimalLongitude are different to zero or empty and values for dwc:eventDate, dwc:scientificName and dwc:scientificNameAuthorship are different to empty. Applicable to Darwin Core based single records.
s_{16}	**Method description**: Change signs (positive and negative) of dwc:decimalLatitude and dwc:decimalLongitude and recommend the dwc:decimalLatitude and dwc:decimalLongitude that is consistent with dwc:country or consistent with the georeference of dwc:country, dwc:stateProvince, dwc:county and dwc:locality. Applicable to Darwin Core based single records. **Georeference sources**: GeoLocate | Land data from Natural Earth | Country boundary data from GeoCommunity.
s_{17}	**Method description**: Check if dwc:decimalLatitude and dwc:decimalLongitude values are consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close

	enough (currently defined as 200 km) to the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality and if dwc:eventDate value is correctly formatted according to ISO 8601 standard and is before dwc:modified and within life span of dwc:collectedBy and if there is an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclature authorities. Applicable to Darwin Core based single records. **Georeference sources**: GeoLocate | Land data from Natural Earth | Country boundary data from GeoCommunity. **Life span of collectors sources**: Harvard List of Botanists | FilteredPush Entomologists List. **Nomenclatural authority sources**: Catalog of Life | Global Name Resolver | Global Name Index | GBIF CheckListBank Backbone | IPNI | Index Fungorum | WoRMS.
s_{18}	**Method description**: Recommend the most similar and valid Scientific Name and Scientific Name Authorship according to nomenclature authorities based on string similarity algorithms and in nomenclature rules and conventions. Details can be found at: http://sourceforge.net/p/filteredpush/svn/HEAD/tree/trunk/FP-Tools/FP-CurationServices/src/main/java/edu/harvard/mcz/nametools/. Applicable to Darwin Core based single records. **Nomencaltural authority sources**: Catalog of Life | Global Name Resolver | Global Name Index | GBIF CheckListBank Backbone | IPNI | Index Fungorum | WoRMS.
s_{19}	**Method description**: Calculate the proportion of records in dataset with values for dwc:decimalLatitude and dwc:decimalLongitude that are consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality and value for dwc:eventDate is correctly formatted according to ISO 8601 standard and is before dwc:modified and within life span of dwc:collectedBy and there is an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclature authorities. Applicable to Darwin Core based single records. **Georeference sources**: GeoLocate | Land data from Natural Earth | Country boundary data from GeoCommunity. **Life span of collectors sources**: Harvard List of Botanists | FilteredPush Entomologists List. **Nomenclatural authority sources**: Catalog of Life | Global Name Resolver | Global Name Index | GBIF CheckListBank Backbone | IPNI | Index Fungorum | WoRMS.

s_{20}	**Method description**: Recommend ISO 8601 standardized value for dwc:eventDate based on its own value, changing the order of year, month and day if they can be parsed or try to use dwc:year, dwc:month, dwc:day and dwc:startDayOfYear values to infer the corresponding dwc:evendDate value. Applicable to Darwin Core based single records.
s_{21}	**Method description**: Check if the proportion of records in dataset with values for dwc:decimalLatitude and dwc:decimalLongitude that are consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality and value for dwc:eventDate is correctly formatted according to ISO 8601 standard and is before dwc:modified and within life span of dwc:collectedBy and there is an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclature authorities is equal 100%. Applicable to Darwin Core based single records. **Georeference sources**: GeoLocate | Land data from Natural Earth | Country boundary data from GeoCommunity. **Life span of collectors sources**: Harvard List of Botanists | FilteredPush Entomologists List. **Nomenclatural authority sources**: Catalog of Life | Global Name Resolver | Global Name Index | GBIF CheckListBank Backbone | IPNI | Index Fungorum | WoRMS.
s_{22}	**Method description**: Calculate the proportion of records with values (disregarding extra spaces) for both dwc:decimalLatitude and dwc:decimalLongitude that are different to zero or empty and values for dwc:eventDate, dwc:scientificName and dwc:scientificNameAuthorship that are different to empty. Applicable to datasets of Darwin Core based records.
s_{23}	**Method description**: Check if the average of the number of characters after "." of dwc:decimalLatitude and dwc:decimalLongitude is higher than 4. Applicable to Darwin Core based single records.
s_{24}	**Method description**: Check if dwc:eventDate is according to ISO 8601 standard specification. See details in: http://momentjs.com/docs/. Applicable to Darwin Core based single records.
s_{25}	**Method description**: Check if the proportion of records in dataset with values (disregarding extra spaces) different to zero and different to empty for both dwc:decimalLatitude

	and dwc:decimalLongitude is equal 100%. Applicable to datasets of Darwin Core based records.
s_{26}	**Method description**: Check if the mean of the average of the number of characters after "." of dwc:decimalLatitude and dwc:decimalLongitude of records of dataset is higher than 4. Applicable to datasets of Darwin Core based records.
s_{27}	**Method description**: Check if the proportion of records in dataset with values (disregarding extra spaces) different to empty for both dwc:scientificName and dwc:scientificNameAuthorship is equal to 100%. Applicable to datasets of Darwin Core based records.
s_{28}	**Method description**: Check if the proportion of records in dataset with values (disregarding extra spaces) different to empty for dwc:eventDate is equal to 100%. Applicable to datasets of Darwin Core based records.
s_{29}	**Method description**: Check if the proportion of records that have values (disregarding extra spaces characters) for both dwc:decimalLatitude and dwc:decimalLongitude that are different to zero or empty and values for dwc:eventDate, dwc:scientificName and dwc:scientificNameAuthorship that are different to empty is equal 100%. Applicable to datasets of Darwin Core based records.
s_{30}	**Method description**: Check if the proportion of records dwc:decimalLatitude and dwc:decimalLongitude values that are consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality is equal 100%. Applicable to datasets of Darwin Core based records.
s_{31}	**Method description**: Check if the proportion of records with an exact match of both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclature authorities is equal 100%. Applicable to datasets of Darwin Core based records.
s_{32}	**Method description**: Check if the proportion of records of dwc:eventDate is correctly formatted according to ISO 8601 standard and is before dwc:modified and within life span of collector is equal 100%. Applicable to a dataset of Darwin Core records.
s_{33}	**Method description**: For all records in dataset, change values of dwc:scientificName and dwc:scientificNameAuthorship to the most similar and valid dwc:scientificName and dwc:scientificNameAuthorship according to nomenclature authorities based on string similarity algorithms and in nomenclature rules and conventions; change value of dwc:eventDate to an ISO 8601 standardized value based on

	its own value when it is possible to parse value into year, month and day or change dwc:eventDate to a date obtained from dwc:year, dwc:month, dwc:day and dwc:startDayOfYear values; and change values of dwc:decimalLatitude and dwc:decimalLongitude values with different signs (positive and negative) that is consistent with dwc:country or consistent with the georeference of dwc:country, dwc:stateProvince, dwc:county and dwc:locality. Applicable to datasets of Darwin Core based records. **Georeference sources**: GeoLocate | Land data from Natural Earth | Country boundary data from GeoCommunity. **Life span of collectors sources**: Harvard List of Botanists | FilteredPush Entomologists List. **Nomenclatural authority sources**: Catalog of Life | Global Name Resolver | Global Name Index | GBIF CheckListBank Backbone | IPNI | Index Fungorum | WoRMS.
s_{34}	**Method description**: Disregard records in dataset with dwc:decimalLatitude and dwc:decimalLongitude values that are both equal to zero or one of them is equal to empty or they are not consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are not consistent with the bounds of dwc:country or not close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality; or dwc:eventDate value is equal to empty or is not correctly formatted according to ISO 8601 standard or is not before dwc:modified or is not within life span of dwc:collectedBy; or there is not an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclature authorities or one of their values is empty. Applicable to datasets of Darwin Core based records. **Georeference sources**: GeoLocate | Land data from Natural Earth | Country boundary data from GeoCommunity. **Life span of collectors sources**: Harvard List of Botanists | FilteredPush Entomologists List. **Nomenclatural authority sources**: Catalog of Life | Global Name Resolver | Global Name Index | GBIF CheckListBank Backbone | IPNI | Index Fungorum | WoRMS.

Table T describes two software packages that are capable of performing several kinds of measurements, validations and improvements.

Table T - Mechanisms

ID	Label	Description

m_1	FP-AKKA Kurator	**Name**: FP-AKKA Kurator **Description**: This software produces a data quality report on a set of natural science collections data. The software examines the data records for internal consistency, checks them against external services (such as Geolocate), identifies potential problems, and where possible, proposes corrections that may be applied to the data. **Institution**: Kurator Project **URL**: http://wiki.datakurator.net/web/Kurator **User Documentation**: http://wiki.datakurator.net/web/User_Documentation **Developer Documentation**: http://wiki.datakurator.net/web/Developer_Documentation **Source Code:** https://github.com/kurator-org
m_2	BDQ Toolkit	**Name**: Biodiversity Data Quality Toolkit **Description:** This toolkit provides an Application Programming Interface (API) for ad hoc use by applications that are able to consume Representational State Transfer (REST) web services. This toolkit provides mechanisms to measure and validate DQ of both datasets and single records based on the Darwin Core standard. **Institution**: Biocomp **URL**: http://toolkit.bdq.biocomp.org.br:3020 **SOURCE CODE**: https://github.com/allankv/bdq-toolkit-case.git

2.2. Derived Concepts

This Subsection describes DQ Solutions concepts instances that are defined by the relationships between fundamental concepts instances or other derived concepts instances. These relationships define new concepts instances with their own meaning.

Table U defines the relationship between Specifications and Contextualized Dimensions. Each row defines the Specifications that can be used as method to measure a specific Contextualized Dimension. Due to the scope delimitation of this proof of concept, the

cardinality of all of the listed measurement methods is one to one, but in a broader context the cardinality may be one to many.

Table U - Measurement Methods

Definition	Description
$mm(cd_1) = \{s_3\}$	The Contextualized Dimension `Single Record Coordinates Numerical Precision` (cd_1) can be measured using the Specification s_3.
$mm(cd_2) = \{s_1\}$	The Contextualized Dimension `Single Record Coordinates Completeness` can be measured using the Specification s_1.
$mm(cd_3) = \{s_2\}$	The Contextualized Dimension `Single Record Coordinates Consistency` can be measured using the Specification s_2.
$mm(cd_4) = \{s_4\}$	The Contextualized Dimension `Dataset Coordinates Completeness` can be measured using the Specification s_4.
$mm(cd_5) = \{s_5\}$	The Contextualized Dimension `Dataset Coordinates Consistency` can be measured using the Specification s_5.
$mm(cd_6) = \{s_6\}$	The Contextualized Dimension `Dataset Coordinates Numerical Precision` can be measured using the Specification s_6.
$mm(cd_7) = \{s_7\}$	The Contextualized Dimension `Single Record SciName Completeness` can be measured using the Specification s_7.
$mm(cd_8) = \{s_8\}$	The Contextualized Dimension `Single Record SciName Accuracy` can be measured using the Specification s_8.
$mm(cd_9) = \{s_9\}$	The Contextualized Dimension `Dataset SciName Completeness` can be measured using the Specification s_9.
$mm(cd_{10}) = \{s_{10}\}$	The Contextualized Dimension `Dataset SciName Accuracy` can be measured using the Specification s_{10}.
$mm(cd_{11}) = \{s_{11}\}$	The Contextualized Dimension `Single Record Collected Date Completeness` can be measured using the Specification s_{11}.
$mm(cd_{12}) = \{s_{12}\}$	The Contextualized Dimension `Single Record Collected Date Consistency` can be measured using the Specification s_{12}.
$mm(cd_{13}) = \{s_{13}\}$	The Contextualized Dimension `Dataset Collected Date Completeness` can be measured using the Specification s_{13}.
$mm(cd_{14}) = \{s_{14}\}$	The Contextualized Dimension `Dataset Collected Date Consistency` can be measured using the Specification s_{14}.
$mm(cd_{15}) = \{s_{15}\}$	The Contextualized Dimension `Single Record Occurrence Completeness` can be measured using the Specification s_{15}.

$mm(cd_{16}) = \{s_{22}\}$	The Contextualized Dimension `Dataset Occurrence Completeness` can be measured using the Specification s_{22}.
$mm(cd_{17}) = \{s_{17}\}$	The Contextualized Dimension `Single Record Occurrence Accuracy` can be measured using the Specification s_{17}.
$mm(cd_{18}) = \{s_{19}\}$	The Contextualized Dimension `Dataset Occurrence Accuracy` can be measured using the Specification s_{19}.

Table V defines the relationship between Specifications and Contextualized Criteria. Each row defines the Specifications that can be used as a method to validate the compliance of a specific Contextualized Criterion. Due to the scope delimitation of this proof of concept, the cardinality of all of the listed measurement methods is one to one, but in a broader context the cardinality may be one to many.

Table V - Validation Methods

Definition	Description
$vm(cc_1) = \{s_{23}\}$	The Contextualized Criterion `Coordinates Numerical Precision must be higher than 4` (cc_1) can be validated using the Specification s_{23}.
$vm(cc_2) = \{s_1\}$	The Contextualized Criterion `Coordinates must be complete` can be validated using the Specification s_1.
$vm(cc_3) = \{s_2\}$	The Contextualized Criterion `Coordinates must be consistent` can be validated using the Specification s_2.
$vm(cc_4) = \{s_{25}\}$	The Contextualized Criterion `All records in a Dataset must have complete Coordinates` can be validated using the Specification s_{25}.
$vm(cc_5) = \{s_{30}\}$	The Contextualized Criterion `All records in a Dataset must have consistent Coordinates` can be validated using the Specification s_{30}.
$vm(cc_6) = \{s_{26}\}$	The Contextualized Criterion `Average value of Coordinates Numerical Precision within a Dataset must be higher than 4` can be validated using the Specification s_{26}.
$vm(cc_7) = \{s_7\}$	The Contextualized Criterion `SciName must be complete` can be validated using the Specification s_7.
$vm(cc_8) = \{s_8\}$	The Contextualized Criterion `SciName must be accurate` can be validated using the Specification s_8
$vm(cc_9) = \{s_{27}\}$	The Contextualized Criterion `All records in a Dataset must have complete SciName` can be validated using the Specification s_{27}.

$vm(cc_{10}) = \{s_{31}\}$	The Contextualized Criterion `All records in a Dataset must have accurate SciName` can be measured using the Specification s_{31}
$vm(cc_{11}) = \{s_{11}\}$	The Contextualized Criterion `Collected Date must be complete` can be validated using the Specification s_{11}.
$vm(cc_{12}) = \{s_{12}\}$	The Contextualized Criterion `Collected Date must be consistent` can be validated using the Specification s_{12}.
$vm(cc_{13}) = \{s_{28}\}$	The Contextualized Criterion `All records in a Dataset must have complete Collected Date` can be validated using the Specification s_{28}.
$vm(cc_{14}) = \{s_{32}\}$	The Contextualized Criterion `All records of Dataset must have consistent Collected Date` can be validated using the Specification s_{32}.
$vm(cc_{15}) = \{s_{15}\}$	The Contextualized Criterion `Occurrence must be complete` can be validated using the Specification s_{15}.
$vm(cc_{16}) = \{s_{29}\}$	The Contextualized Criterion `All records in a Dataset must have Occurrence complete` can be validated using the Specification s_{29}.
$vm(cc_{17}) = \{s_{24}\}$	The Contextualized Criterion `Collected Date must follow ISO 8601 standard format` can be validated using the Specification s_{24}.
$vm(cc_{18}) = \{s_{17}\}$	The Contextualized Criterion `Occurrence must be accurate` can be validated using the Specification s_{17}.
$vm(cc_{19}) = \{s_{21}\}$	The Contextualized Criterion `All records in a Dataset must have accurate Occurrence` can be validated using the Specification s_{21}.

Table W defines the relationship between Specifications and Contextualized Enhancements. Each row defines the Specifications that can be used as methods to improve DQ according to a specific Contextualized Enhancements. Due to the scope delimitation of this proof of concept, the cardinality of all of the listed measurement methods is one to one, but in a broader context the cardinality may be one to many.

Table W - Improvement Methods

Definition	Description
$im(ce_1) = \{s_{18}\}$	The Contextualized Enhancement `Recommendation of SciName based on nomenclature authorities` (ce_1) can be obtained using the Specification s_{18}.

$im(ce_2) = \{s_{20}\}$	The Contextualized Enhancement `Recommendation of Collected Date based on ISO 8601 standard` can be obtained using the Specification s_{20}.
$im(ce_3) = \{s_{16}\}$	The Contextualized Enhancement `Recommendation of Coordinates based on the associated country, state/province, county and locality` can be obtained using the Specification s_{16}.
$im(ce_4) = \{s_{33}\}$	The Contextualized Enhancement `Accept all recommendations for each Single Record in Dataset` can be performed using the Specification s_{33}.
$im(ce_5) = \{s_{34}\}$	The Contextualized Enhancement `Disregard all records that are not compliant with all DQ Criteria` can be performed using the Specification s_{34}.

Table X defines the relationship between Specifications and Mechanisms. Each row defines the Mechanisms that implement a specific Specification. Due to the scope delimitation of this proof of concept, the cardinality of all of the listed implementations is one to one, but in a broader context the cardinality may be one to many.

Table X - Implementations

Definition	Description
$i(s_1) = \{m_2\}$	`BDQ Toolkit` (m_2) implements the Specification s_1
$i(s_2) = \{m_1\}$	`FP-Akka Kurator` implements the Specification s_2
$i(s_3) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_3
$i(s_4) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_4
$i(s_5) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_5
$i(s_6) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_6
$i(s_7) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_7
$i(s_8) = \{m_1\}$	`FP-Akka Kurator` implements the Specification s_8
$i(s_9) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_9
$i(s_{10}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{10}
$i(s_{11}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{11}
$i(s_{12}) = \{m_1\}$	`FP-Akka Kurator` implements the Specification s_{12}
$i(s_{13}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{13}

$i(s_{14}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{14}
$i(s_{15}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{15}
$i(s_{16}) = \{m_1\}$	`FP-Akka Kurator` implements the Specification s_{16}
$i(s_{17}) = \{m_1\}$	`FP-Akka Kurator` implements the Specification s_{17}
$i(s_{18}) = \{m_1\}$	`FP-Akka Kurator` implements the Specification s_{18}
$i(s_{19}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{19}
$i(s_{20}) = \{m_1\}$	`FP-Akka Kurator` implements the Specification s_{20}
$i(s_{21}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{21}
$i(s_{22}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{22}
$i(s_{23}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{23}
$i(s_{24}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{24}
$i(s_{25}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{25}
$i(s_{26}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{26}
$i(s_{27}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{27}
$i(s_{28}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{28}
$i(s_{29}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{29}
$i(s_{30}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{30}
$i(s_{31}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{31}
$i(s_{32}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{32}
$i(s_{33}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{33}
$i(s_{34}) = \{m_2\}$	`BDQ Toolkit` implements the Specification s_{34}

Table Y also defines the relationship between Specifications and Mechanisms. Each row defines the Specifications implemented by a specific Mechanism.

Table Y - Mechanism Coverage

Definition	Description
$mc(m_1) = \{s_2, s_8, s_{12}, s_{16}, s_{18}, s_{20}\}$	`FP-Akka Kurator` (m_1) implements the Specifications s_2, s_8, s_{12}, s_{16}, s_{18} and s_{20}.

$mc(m_2) = \{s_1, s_3, s_4, s_5, s_6, s_7, s_9, s_{10}, s_{11}, s_{13}, s_{14}, s_{15}, s_{17}, s_{19}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32}\}$	`BDQ Toolkit` implements the Specifications s_1, s_3, s_4, s_5, s_6, s_7, s_9, s_{10}, s_{11}, s_{13}, s_{14}, s_{15}, s_{17}, s_{19}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27} s_{28}, s_{29}, s_{30}, s_{31} and s_{32}.

3. DQ Report

DQ Report defines the DQ Measures, DQ Validations and DQ Improvements which enable the DQ Assessment and DQ Management of fitness for use of Data Resources (datasets or single record). This section presents the DQ Report concepts instances generated according to the DQ Solutions and DQ Needs defined in Sections 1 and 2, respectively, in order to enable the assessment and management of the fitness for use of a dataset and its single records according to the `Curation of Biological Collections` context.

For this proof of concept we used a dataset from Arizona State University Hasbrouck Insect Collection (http://symbiota4.acis.ufl.edu/scan/portal/collections) to generate the DQ Measures, DQ Validations and DQ Improvements, and therefore, perform the DQ Assessment and DQ Management by Quality Control and Quality Assurance of the entire dataset and its single records. These Data Resources are available in JSON format (machine readable) through an API at: http://case.bdq.biocomp.org.br:3010/explorer/#!/OrginalData.

We used two Mechanisms to operate on the Data Resources and generate the Assertions assigned to them. The results generated by the Mechanism FP-Akka Kurator are available at: http://case.bdq.biocomp.org.br:3010/explorer/#!/FPAkkaOutput. The results generated by the Mechanism BDQ Toolkit are available at: http://case.bdq.biocomp.org.br:3010/explorer/#!/BDQToolkitOutput.

All results generated by this experiment can be consulted through the Loopback API Explorer (StrongLoop, 2015) avaliable at: http://case.bdq.biocomp.org.br:3010/explorer.

3.1. Fundamental Concepts

This subsection describes the instances of the fundamental concept of the DQ Report class, the Data Resource.

We used a dataset and a single record as examples of Data Resources for illustrating the fitness for use assessment and management. They are presented in Table Z.

Table Z - Data Resources

Definition	Variable	Value
$dr_1 =< id_1, rt_2, v_1 >$	id_1	http://case.bdq.biocomp.org.br:3010/api/v1.0/OrginalData
	rt_2	Dataset
	v_1	Download at: http://case.bdq.biocomp.org.br:3010/api/v1.0/OrginalData

$dr_2 =< id_2, rt_1, v_2 >$	id_2	555f7b8ed53d8661fd3f53ed
	rt_1	Single Record
	v_2	**eventDate**: 0000-00-00 **municipality**: Tijuana Municipality **oaiid**: SCAN.occurrence.925670 **identifiedBy**: R.M. Bohart **georeferenceSources**: Google Earth **geodeticDatum**: WGS84 **family**: Vespidae **catalogNumber**: ASUHIC0032157 **recordedBy**: C. Saheitlin **stateProvince**: Baja California **year**: 1968 **dateIdentified**: 1972 **startDayOfYear**: 176 **scientificName**: Vespula pensylvanica **georeferenceVerificationStatus**: requires verification **scientificNameAuthorship**: Rohwer, 1857 **ownerInstitutionCode**: ASU **taxonID**: http://api.gbif.org/v1/species/1311698 **collectionCode**: ASUHIC **modified**: 2013-12-05 20:02:40 **country**: Mexico **occurrenceRemarks**: no collection date record **decimalLatitude**: 32.507822 **basisOfRecord**: PreservedSpecimen **institutionCode**: ASU **decimalLongitude**: -116.975289 **month**: 6 **locality**: Los Angeles **day**: 24 **georeferencedBy**: David Fleming **id**: 555f7b8ed53d8661fd3f53ed

3.2. Derived Concepts

This subsection presents the instances of derived concepts of the DQ Report class, the DQ Measures, DQ Validations, DQ Improvement, DQ Assessment, DQ Management by Quality Control and Quality Assurance related to the the two Data Resources defined in the Table Z.

Table AA presents all the DQ Measures assigned to each Data Resource defined in Table Z.

Table AA - DQ Measures

Definition	DQ Measures
$dqm(dr_1) = \{< cd_4, s_4, m_2, r >,$	**Contextualized Dimension**: Coordinates Completeness

$< cd_6, s_6, m_2, r >,$ $< cd_9, s_9, m_2, r >,$ $< cd_{13}, s_{13}, m_2, r >,$ $< cd_{16}, s_{22}, m_2, r >,$ $< cd_5, s_5, m_2, r >,$ $< cd_{10}, s_{10}, m_2, r >,$ $< cd_{14}, s_{14}, m_2, r >,$ $< cd_{18}, s_{19}, m_2, r >\}$	**Specification**: Calculate the proportion of records in dataset with values (disregarding extra spaces) different to zero and different to empty for both dwc:decimalLatitude and dwc:decimalLongitude. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: 0.9999236815996336 **Contextualized Dimension**: Coordinates Numerical Precision **Specification**: Calculate the mean of the average number of characters after "." of dwc:decimalLatitude and dwc:decimalLongitude of records of dataset. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: 5.546315729222315 **Contextualized Dimension**: SciName Completeness **Specification**: Calculate the proportion of records in dataset with values (disregarding extra spaces) different to empty for both dwc:scientificName and dwc:scientificNameAuthorship. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: 0.9728115698694956 **Contextualized Dimension**: Collected Date Completeness **Specification**: Calculate the proportion of records in dataset that have the dwc:eventDate value (disregarding extra spaces) different to empty. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: 1 **Contextualized Dimension**: Occurrence Completeness **Specification**: Calculate the proportion of records with values (disregarding extra spaces) for both dwc:decimalLatitude and dwc:decimalLongitude that are different to zero or empty and values for dwc:eventDate, dwc:scientificName and dwc:scientificNameAuthorship that are different to empty. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: 0.9727352514691292

	Contextualized Dimension: Coordinates Consistency **Specification**: Calculate the proportion of records dwc:decimalLatitude and dwc:decimalLongitude values consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: 0.8898153094711135 **Contextualized Dimension**: SciName Accuracy **Specification**: Calculate the proportion of records with an exact match of both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclatural authorities. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: 0.49282607036556514 **Contextualized Dimension**: Collected Date Consistency **Specification**: Calculate the proportion of dwc:eventDate value records that are correctly formatted according to ISO 8601 standard and that are before dwc:modified and within life span of collector. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: 0.9670113714416546 **Contextualized Dimension**: Occurrence Accuracy **Specification**: Calculate the proportion of records in dataset with values for dwc:decimalLatitude and dwc:decimalLongitude that are consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality and value for dwc:eventDate is correctly formatted according to ISO 8601 standard and is before dwc:modified and within life span of dwc:collectedBy and there is an exact match with both dwc:scientificName and

	dwc:scientificNameAuthorship values with nomenclature authorities. Applicable to Darwin Core based single records. **Mechanism**: BDQ Toolkit **Assertion**: 0.4286613752575746
$dqm(dr_2) = \{< cd_1, s_3, m_2, r >,$ $< cd_2, s_1, m_2, r >,$ $< cd_7, s_7, m_2, r >,$ $< cd_{11}, s_{11}, m_2, r >,$ $< cd_{15}, s_{15}, m_2, r >,$ $< cd_3, s_2, m_1, r >,$ $< cd_8, s_8, m_2, r >,$ $< cd_{12}, s_{12}, m_1, r >,$ $< cd_{17}, s_{17}, m_1, r >\}$	**Contextualized Dimension**: Coordinates Numerical Precision **Specification**: Calculate the average of the number of characters after "." of dwc:decimalLatitude and dwc:decimalLongitude. **Mechanism**: BDQ Toolkit **Assertion**: 6 --- **Contextualized Dimension**: Coordinates Completeness **Specification**: Check whether values (disregarding extra spaces characters) for both dwc:decimalLatitude and dwc:decimalLongitude are different of zero and different of empty. **Mechanism**: BDQ Toolkit **Assertion**: Complete --- **Contextualized Dimension**: SciName Completeness **Specification**: Check whether values (disregarding extra spaces characters) for both dwc:scientificName and dwc:scientificNameAuthorship are different of empty. **Mechanism**: BDQ Toolkit **Assertion**: Complete --- **Contextualized Dimension**: Collected Date Completeness **Specification**: Check whether value (disregarding extra spaces characters) for dwc:eventDate is different of empty. **Mechanism**: BDQ Toolkit **Assertion**: Complete --- **Contextualized Dimension**: Occurrence Completeness **Specification**: Check whether values (disregarding extra spaces characters) for both dwc:decimalLatitude and dwc:decimalLongitude are different of zero or empty and values for dwc:eventDate, dwc:scientificName and dwc:scientificNameAuthorship are different of empty. **Mechanism**: BDQ Toolkit **Assertion**: Complete --- **Contextualized Dimension**: Coordinates Consistency

	Specification: Check whether dwc:decimalLatitude and dwc:decimalLongitude values are consistent with coordinates range (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality. **Mechanism**: FP-Akka Kurator **Assertion**: Not Consistent --- **Contextualized Dimension**: SciName Accuracy **Specification**: Check against taxonomic authorities whether there is an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values. **Mechanism**: FP-Akka Kurator **Assertion**: Not Accurate --- **Contextualized Dimension**: Collected Date Consistency **Specification**: Check whether dwc:eventDate value is correctly formatted according to ISO 8601 standard and is before dwc:modified and between life span of collector. **Mechanism**: FP-Akka Kurator **Assertion**: Not Consistent --- **Contextualized Dimension**: Coordinates Accuracy Consistency **Specification**: Check whether dwc:decimalLatitude and dwc:decimalLongitude values are consistent with coordinates range (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality and whether dwc:eventDate value is correctly formatted according to ISO 8601 standard and is before dwc:modified and between life span of dwc:colletedBy and whether there is an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values with taxonomic authorities. **Mechanism**: FP-Akka Kurator **Assertion**: Not Accurate

Table AB presents all the DQ Validations assigned to each Data Resource defined in the Table Z.

Table AB - DQ Validations

Definition	DQ Validations
$dqv(dr_1) = \{< cc_4, s_{25}, m_2, r >, < cc_6, s_{26}, m_2, r >, < cc_9, s_{27}, m_2, r >, < cc_{13}, s_{28}, m_2, r >, < cc_{16}, s_{29}, m_2, r >, < cc_5, s_{30}, m_2, r >, < cc_{10}, s_{31}, m_2, r >, < cc_{14}, s_{32}, m_2, r >, < cc_{19}, s_{21}, m_2, r >\}$	**Contextualized Criterion**: All records in a Dataset must have complete Coordinates **Specification**: Check if the proportion of records in dataset with values (disregarding extra spaces) different to zero and different to empty for both dwc:decimalLatitude and dwc:decimalLongitude is equal 100%. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: Not Compliant
	Contextualized Criterion: Average value of Coordinates Numerical Precision within a Dataset must be higher than 4 **Specification**: Check if the mean of the average of the number of characters after "." of dwc:decimalLatitude and dwc:decimalLongitude of records of dataset is higher than 4. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: Compliant
	Contextualized Criterion: All records in a Dataset must have complete SciName **Specification**: Check if the proportion of records in dataset with values (disregarding extra spaces) different to empty for both dwc:scientificName and dwc:scientificNameAuthorship is equal to 100%. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: Not Compliant
	Contextualized Criterion: All records in a Dataset must have complete Collected Date **Specification**: Check if the proportion of records in dataset with values (disregarding extra spaces) different to empty for dwc:eventDate is equal to 100%. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: Compliant
	Contextualized Criterion: All records in a Dataset must have Occurrence complete **Specification**: Check if the proportion of records that have values (disregarding extra spaces characters) for both dwc:decimalLatitude and dwc:decimalLongitude that are different to zero or empty and values for dwc:eventDate,

	dwc:scientificName and dwc:scientificNameAuthorship that are different to empty is equal 100%. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: Not Compliant --- **Contextualized Criterion**: All records in a Dataset must have consistent Coordinates **Specification**: Check if the proportion of records dwc:decimalLatitude and dwc:decimalLongitude values that are consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality is equal 100%. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: Not Compliant --- **Contextualized Criterion**: All records in a Dataset must have accurate SciName **Specification**: Check if the proportion of records with an exact match of both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclature authorities is equal 100%. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: Not Compliant --- **Contextualized Criterion**: All records of Dataset must have consistent Collected Date **Specification**: Check if the proportion of records of dwc:eventDate is correctly formatted according to ISO 8601 standard and is before dwc:modified and within life span of collector is equal 100%. Applicable to a dataset of Darwin Core records. **Mechanism**: BDQ Toolkit **Assertion**: Not Compliant --- **Contextualized Criterion**: All records in a Dataset must have accurate Occurrence **Specification**: Check if the proportion of records in dataset with values for dwc:decimalLatitude and dwc:decimalLongitude that are consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes

	range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality and value for dwc:eventDate is correctly formatted according to ISO 8601 standard and is before dwc:modified and within life span of dwc:collectedBy and there is an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclature authorities is equal 100%. Applicable to Darwin Core based single records. **Mechanism**: BDQ Toolkit **Assertion**: Not Compliant
$dqv(dr_2) = \{< cc_1, s_3, m_2, r >,$ $< cc_2, s_1, m_2, r >,$ $< cc_7, s_7, m_2, r >,$ $< cc_{11}, s_{11}, m_2, r >,$ $< cc_{15}, s_{15}, m_2, r >,$ $< cc_{17}, s_{24}, m_2, r >,$ $< cc_3, s_2, m_1, r >,$ $< cc_8, s_8, m_2, r >,$ $< cc_{12}, s_{12}, m_1, r >,$ $< cc_{18}, s_{17}, m_1, r >\}$	**Contextualized Criterion**: Coordinates Numerical Precision must be higher than 4 **Specification**: Check whether the average of the number of characters after "." of dwc:decimalLatitude and dwc:decimalLongitude is higher than 4. **Mechanism**: BDQ Toolkit **Assertion**: Compliant --- **Contextualized Criterion**: Coordinates must be complete **Specification**: Check whether values (disregarding extra spaces characters) for both dwc:decimalLatitude and dwc:decimalLongitude are different of zero and different of empty. **Mechanism**: BDQ Toolkit **Assertion**: Compliant --- **Contextualized Criterion**: SciName must be complete **Specification**: Check whether values (disregarding extra spaces characters) for both dwc:scientificName and dwc:scientificNameAuthorship are different of empty. **Mechanism**: BDQ Toolkit **Assertion**: Compliant --- **Contextualized Criterion**: Collected Date must be complete **Specification**: Check whether value (disregarding extra spaces characters) for dwc:eventDate is different of empty. **Mechanism**: BDQ Toolkit **Assertion**: Compliant --- **Contextualized Criterion**: Occurrence must be complete **Specification**: Check whether values (disregarding extra spaces characters) for both dwc:decimalLatitude and

	dwc:decimalLongitude are different of zero or empty and values for dwc:eventDate, dwc:scientificName and dwc:scientificNameAuthorship are different of empty. **Mechanism**: BDQ Toolkit **Assertion**: Compliant --- **Contextualized Criterion**: Collected Date must follow ISO 8601 standard format **Specification**: Check whether dwc:eventDate is according to ISO 8601 standard specification. **Mechanism**: BDQ Toolkit **Assertion**: Not Compliant --- **Contextualized Criterion**: Coordinates must be consistent **Specification**: Check whether dwc:decimalLatitude and dwc:decimalLongitude values are consistent with coordinates range (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality. **Mechanism**: FP-Akka Kurator **Assertion**: Not Compliant --- **Contextualized Criterion**: SciName must be accurate **Specification**: Check against taxonomic authorities whether there is an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values. **Mechanism**: FP-Akka Kurator **Assertion**: Not Compliant --- **Contextualized Criterion**: Collected Date must be consistent **Specification**: Check whether dwc:eventDate value is correctly formatted according to ISO 8601 standard and is before dwc:modified and between life span of collector. **Mechanism**: FP-Akka Kurator **Assertion**: Not Compliant --- **Contextualized Criterion**: Occurrence must be accurate **Specification**: Check whether dwc:decimalLatitude and dwc:decimalLongitude values are consistent with coordinates range (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are consistent with the bounds of dwc:country or close enough (currently

	defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality and whether dwc:eventDate value is correctly formatted according to ISO 8601 standard and is before dwc:modified and between life span of dwc:colletedBy and whether there is an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values with taxonomic authorities. **Mechanism**: FP-Akka Kurator **Assertion**: Not Compliant

This table presents all the DQ Improvements assigned to each Data Resource defined in the Table Z.

Table AC - DQ Improvement

Definition	DQ Improvements
$dqi(dr_1) = \{< ce_4, s_{33}, m_2, r >, < ce_5, s_{34}, m_2, r >\}$	**Contextualized Enhancement**: Accept all recommendations for each Single Record in Dataset **Specification**: For all records in dataset, change values of dwc:scientificName and dwc:scientificNameAuthorship to the most similar and valid dwc:scientificName and dwc:scientificNameAuthorship according to nomenclature authorities based on string similarity algorithms and in nomenclature rules and conventions; change value of dwc:eventDate to an ISO 8601 standardized value based on its own value when it is possible to parse value into year, month and day or change dwc:eventDate to a date obtained from dwc:year, dwc:month, dwc:day and dwc:startDayOfYear values; and change values of dwc:decimalLatitude and dwc:decimalLongitude values with different signs (positive and negative) that is consistent with dwc:country or consistent with the georeference of dwc:country, dwc:stateProvince, dwc:county and dwc:locality. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: http://case.bdq.biocomp.org.br:3010/api/v1.0/DQReportControls **Contextualized Enhancement**: Disregard all records that are not compliant with all DQ Criteria **Specification**: Disregard records in dataset with dwc:decimalLatitude and dwc:decimalLongitude values that

	are both equal to zero or one of them is equal to empty or they are not consistent with coordinates ranges (latitudes range from -90 to 90 and longitudes range from -180 to 180) and are not consistent with the bounds of dwc:country or not close enough (currently defined as 200 km) from the georeference for dwc:country, dwc:stateProvince, dwc:county and dwc:locality; or dwc:eventDate value is equal to empty or is not correctly formatted according to ISO 8601 standard or is not before dwc:modified or is not within life span of dwc:collectedBy; or there is not an exact match with both dwc:scientificName and dwc:scientificNameAuthorship values with nomenclature authorities or one of their values is empty. Applicable to datasets of Darwin Core based records. **Mechanism**: BDQ Toolkit **Assertion**: http://case.bdq.biocomp.org.br:3010/api/v1.0/DQReportAssurances
$dqi(dr_2) = \{< ce_1, s_{18}, m_1, r >, < ce_2, s_{20}, m_1, r >\}$	**Contextualized Enhancement**: Recommendation of SciName based on nomenclature authorities **Specification**: Recommend the most similar and valid Scientific Name and Scientific Name Authorship according to taxonomic authorities based on string similarity algorithms and in taxonomic rules and conventions. Details can be found at: http://sourceforge.net/p/filteredpush/svn/HEAD/tree/trunk/FP-Tools/FP-CurationServices/src/main/java/edu/harvard/mcz/nametools/. **Mechanism**: FP-Akka Kurator **Assertion**: scientificName: "Vespula pensylvanica", scientificNameAuthorship: "(de Saussure, 1857)" --- **Contextualized Enhancement**: Recommendation of Collected Date based on ISO 8601 standard **Specification**: Recommend ISO 8601 standardized value for dwc:eventDate based on its own value, changing the order of year, month and day whether they can be parsed or try to use dwc:year, dwc:month, dwc:day and dwc:startDayOfYear values to infer the corresponding dwc:evendDate value. **Mechanism**: FP-Akka Kurator **Assertion**: eventDate: "1968-06-24"

3.2.1. DQ Assessment of a Dataset

In order to assess the fitness for use of the dataset dr_1 according to the DQ Needs defined for `Biological Collection Curation`, one can use all DQ Measures and DQ Validations represented as $a(dr_1) = \{dqm(dr_1), dqv(dr_1)\}$, presented in the first row of Table AA and Table AB, summarized in Table AD and Table AE. With those Assertions, associated to the Contextualized Dimension or Contextualized Criterion and the Specification and Mechanism used to obtain the DQ Measure or DQ Validation, data users are able to assess if the dataset is fit for use in the context of `Biological Collection Curation`.

Table AD presents the Contextualized Dimensions with their respective Assertions assigned to the Data Resource dr_1.

Table AD - Summary of DQ Measures for the Dataset

Contextualized Dimension	**DQ Measures for the dr_1**
Coordinates Completeness	99.99%
Coordinates Consistency	88.98%
Coordinates Precision	5.55 decimals
Collected Date Completeness	100%
Collected Date Consistency	96.70%
SciName Completeness	97.28%
SciName Accuracy	49.28%
Occurrence Completeness	97.27%
Occurrence Accuracy	42.86%

Table AE presents the Contextualized Criteria with their respective Assertions assigned to the Data Resource dr_1.

Table AE - Summary of DQ Validations for the Dataset

Contextualized Criterion	**DQ Validations for the dr_1**
All records in a Dataset must have complete Coordinates	Not Compliant
All records in a Dataset must have consistent Coordinates	Not Compliant

Average value of Coordinates Numerical Precision within a Dataset must be higher than 4	Compliant
All records in a Dataset must have complete Collected Date	Compliant
All records of Dataset must have consistent Collected Date	Not Compliant
All records in a Dataset must have complete SciName	Not Compliant
All records in a Dataset must have accurate SciName	Not Compliant
All records in a Dataset must have Occurrence complete	Not Compliant
All records in a Dataset must have accurate Occurrence	Not Compliant

3.2.2. DQ Management by Quality Control of a Dataset

In order to manage the fitness for use of the dataset dr_1 by Quality Control approach according to the DQ Needs defined for `Biological Collection Curation`, one can use the DQ Improvement `Accept all recommendations for each Single Record in Dataset`. The Assertion of this DQ Improvement is a modified dr_1 with their single records corrected, where there are correction recommendations.

For performing the Quality Control of the dataset we can use the DQ Improvement $dqi(dr_1) = \{< ce_4, s_{33}, m_2, r >\}$. This improvement accepts all recommendations for each single record in the dataset. The result of this improvement is available at: http://case.bdq.biocomp.org.br:3010/explorer/#!/DQReportControl.

Table AF presents a comparison of the DQ Measures of the original dr_1 and the improved dr_1.

Table AF - DQ Measures for the Dataset with the Quality Control

Contextualized Dimension	**DQ Measure for the original** dr_1	**With the Quality Control**

Coordinates Completeness	99.99%	99.99%
Coordinates Consistency	88.98%	88.98%
Coordinates Precision	5.55 decimals	5.55 decimals
Collected Date Completeness	100%	100%
Collected Date Consistency	96.70%	**97.08%**
SciName Completeness	97.28%	**99.60%**
SciName Accuracy	49.28%	**70.23%**
Occurrence Completeness	97.27%	**99.59%**
Occurrence Accuracy	42.86%	**61.84%**

Table AG presents a comparison of the DQ Validations of original dr_1 and the improved dr_1.

Table AG - DQ Validations for the Dataset with the Quality Control

Contextualized Criterion	**DQ Validations for the original** dr_1	**With the Quality Control**
All records in a Dataset must have complete Coordinates	Not Compliant	Not Compliant
All records in a Dataset must have consistent Coordinates	Not Compliant	Not Compliant
Average value of Coordinates Numerical Precision within a Dataset must be higher than 4	Compliant	Compliant
All records in a Dataset must have complete Collected Date	Compliant	Compliant
All records of Dataset must have consistent Collected Date	Not Compliant	Not Compliant
All records in a Dataset must have complete SciName	Not Compliant	Not Compliant

All records in a Dataset must have accurate SciName	Not Compliant	Not Compliant
All records in a Dataset must have Occurrence complete	Not Compliant	Not Compliant
All records in a Dataset must have accurate Occurrence	Not Compliant	Not Compliant

3.2.3. DQ Management by Quality Assurance of Dataset

In order to manage the fitness for use of the dataset dr_1 by Quality Assurance approach according to the DQ Needs defined for `Biological Collection Curation`, one can use the DQ Improvement `Disregard all records that are not compliant with all DQ Criteria`. The Assertion of this DQ Improvement is a modified dr_1 with only compliant single records.

The result generated by the Quality Assurance improvement is a subset of records of dr_1 that is fit for use in `Biological Collection Curation` according to the defined DQ Needs. From 52,412 records of the original dr_1, the DQ Improvement selected a subset of 32,385 (61.79%) records that is fit for use in the `Biological Collection Curation` context.

For performing the Quality Control of the dataset we can use the DQ Improvement $dqi(dr_1) = \{< ce_5, s_{34}, m_2, r >\}$. This improvement disregards all records that have not complied with all DQ Criteria. The result of this improvement is available at: http://case.bdq.biocomp.org.br:3010/explorer/#!/DQReportAssurance.

Table AH presents a comparison of the DQ Measures of the original dr_1, the improved dr_1 with the Quality Control and the improved dr_1 with the Quality Assurance.

Table AH - DQ Measures for the Dataset with the Quality Assurance

Contextualized Dimension	**DQ Measure for the original dr_1**	**With the Quality Control**	**With the Quality Assurance**
Coordinates Completeness	99.99%	99.99%	**100%**
Coordinates Consistency	88.98%	88.98%	**100%**
Coordinates Precision	5.55 decimals	5.55 decimals	**5.89 decimals**
Collected Date Completeness	100%	100%	**100%**
Collected Date	96.70%	**97.08%**	**100%**

Consistency			
SciName Completeness	97.28%	**99.60%**	**100%**
SciName Accuracy	49.28%	**70.23%**	**100%**
Occurrence Completeness	97.27%	**99.59%**	**100%**
Occurrence Accuracy	42.86%	**61.84%**	**100%**

Table AI presents a comparison of the DQ Validations of the original dr_1, the improved dr_1 with the Quality Control and the improved dr_1 with the Quality Assurance.

Table AI - DQ Validations for the Dataset with the Quality Assurance

Contextualized Criterion	**DQ Validations for the original dr_1**	**With the Quality Control**	**With the Quality Assurance**
All records in a Dataset must have complete Coordinates	Not Compliant	Not Compliant	**Compliant**
All records in a Dataset must have consistent Coordinates	Not Compliant	Not Compliant	**Compliant**
Average value of Coordinates Numerical Precision within a Dataset must be higher than 4	Compliant	Compliant	Compliant
All records in a Dataset must have complete Collected Date	Compliant	Compliant	Compliant
All records of Dataset must have consistent Collected Date	Not Compliant	Not Compliant	**Compliant**
All records in a Dataset must have complete SciName	Not Compliant	Not Compliant	**Compliant**
All records in a Dataset must have accurate SciName	Not Compliant	Not Compliant	**Compliant**

All records in a Dataset must have Occurrence complete	Not Compliant	Not Compliant	**Compliant**
All records in a Dataset must have accurate Occurrence	Not Compliant	Not Compliant	**Compliant**

3.2.4. DQ Assessment of Single Record

In order to assess the fitness for use of the single record dr_2 according to the DQ Needs defined for `Biological Collection Curation`, one can use all DQ Measures and DQ Validations represented as $a(dr_2) = \{dqm(dr_2), dqv(dr_2)\}$, presented in the second row of Table AA and Table AB, summarized in Table AL and Table AM. With those Assertions, associated to the Contextualized Dimension or Contextualized Criterion and the Specification and Mechanism used to obtain the DQ Measure or DQ Validation, data users are able to assess if the single record is fit for use in the context of `Biological Collection Curation`.

Table AJ presents the Contextualized Dimensions with their respective Assertions assigned to the Data Resource dr_2.

Table AJ - Summary of DQ Measures for the Single Record

Contextualized Dimension	**DQ Measure for the** dr_2
Coordinates Completeness	Complete
Coordinates Consistency	Not Consistent
Coordinates Precision	6 decimals
Collected Date Completeness	Complete
Collected Date Consistency	Not Consistent
SciName Completeness	Complete
SciName Accuracy	Not Accurate
Occurrence Completeness	Complete
Occurrence Accuracy	Not Accurate

Table AK presents the Contextualized Criteria with their respective Assertions assigned to the Data Resource dr_2.

Table AK - Summary of DQ Validations for the Single Record

Contextualized Criterion	**DQ Validations for the dr_2**
Coordinates must be complete	Compliant
Coordinates must be consistent	Not Compliant
Coordinates Numerical Precision must be higher than 4	Compliant
Collected Date must be complete	Compliant
Collected Date must be consistent	Not Compliant
SciName must be complete	Compliant
SciName must be accurate	Not Compliant
Occurrence must be complete	Compliant
Occurrence must be accurate	Not Compliant
Collected Date must follow ISO 8601 standard format	Not Compliant

3.2.5. DQ Management by Quality Control of a Single Record

In order to manage the fitness for use of the single record dr_2 by Quality Control approach according to the DQ Needs defined for `Biological Collection Curation`, one can use the DQ Improvements. The Assertion of this DQ Improvement is a modified dr_1 with their single records corrected, where there are correction recommendations.

For performing the Quality Control of the single record dr_1 we can use the DQ Improvements `Recommendation of SciName based on nomenclature authorities` and `Recommendation of Collected Date based on ISO 8601 standard`. These DQ Improvements recommend a `SciName` based on nomenclatural authorities and a `Collected Date` based on ISO 8601 standard for the single record dr_2. The result of this DQ Improvement is available at: http://case.bdq.biocomp.org.br:3010/api/v1.0/DQReportControls/555f7b8ed53d8661fd3f53ed.

Table AL presents a comparison of the DQ Measures of the original dr_2 and the improved dr_2.

Table AL - DQ Measures for the Single Record with the Quality Control

Dimension	**Measure for the original** dr_2	**With the Quality Control**
Coordinates Completeness	Complete	Complete
Coordinates Consistency	Not Consistent	Not Consistent
Coordinates Precision	6 decimals	6 decimals
Collected Date Completeness	Complete	Complete
Collected Date Consistency	Not Consistent	**Consistent**
SciName Completeness	Complete	Complete
SciName Accuracy	Not Accurate	**Accurate**
Occurrence Completeness	Complete	Complete
Occurrence Accuracy	Not Accurate	Not Accurate

Table AM presents a comparison of the DQ Validations of the original dr_2 and the improved dr_2.

Table AM - DQ Validations for the Single Record with the Quality Control

Contextualized Criterion	**Validations of original** dr_2	**Quality Control**
Coordinates must be complete	Compliant	Compliant
Coordinates must be consistent	Not Compliant	Not Compliant
Coordinates Numerical Precision must be higher than 4	Compliant	Compliant
Collected Date must be complete	Compliant	Compliant
Collected Date must be consistent	Not Compliant	**Compliant**
SciName must be complete	Compliant	Compliant
SciName must be accurate	Not Compliant	**Compliant**
Occurrence must be	Compliant	Compliant

complete		
Occurrence must be accurate	Not Compliant	Not Compliant
Collected Date must follow ISO 8601 standard format	Not Compliant	**Compliant**

References

Hamer, M., Victor, J., Smith, G.F. (2012). Best Practice Guide for Compiling, Maintaining and Disseminating National Species Checklists, version 1.0, released in October 2012. Copenhagen: Global Biodiversity Information Facility, 40 pp, ISBN: 87-92020-48-8, Accessible at http://www.gbif.org/orc/?doc_id=4752.

Rees, T. (2008). Applications of fuzzy (approximate string) matching in taxonomic database searches, with an example multi-tiered approach. [Extended abstract]. Pp. 12-14 in Worcester, T., Bajona, L. & Branton, B. (eds): Proceedings of a Conference on Ocean Biodiversity Informatics, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, 2-4 October 2007. Bedford Institute of Oceanography, 2008 (CSAS/SCCS Proceedings Series 2008/024). Accessible at http://www.dfo-mpo.gc.ca/CSAS/Csas/Publications/Pro-CR/2008/2008_024_e.pdf

Chamberlain, S. (2014). rtaxamatch. GitHub repository. Accessible at https://github.com/sckott/rtaxamatch

GBIF (2015). dwca-validator. GitHub repository. [cited 2015 June 14]. Accessible at https://github.com/gbif/dwca-validator

StrongLoop (2015). Use API Explorer [cited 2015 November 21]. Accessible at https://docs.strongloop.com/display/public/LB/Use+API+Explorer

APPENDIX C – List of Activities Related to the Problem Domain Study

Next is presented a short list of activities that helped to better understand the problem domain.

- Participation and talks in the **TDWG (Biodiversity Information Standards) conferences**, one of the most important forum which aims to discuss standards related to biodiversity data; partial results of this research were presented and discussed in the 2012 (Beijing, China), 2013 (Florence, Italy) and 2014 (Jönköping, Sweden) annual editions;

- Participation and invited talk in the **III Conferência Internacional em Qualidade da Informação** in 2012 (São Paulo);

- Participation at the **Quality Information Brasil (QIBRAS) Training**, given by Dr. Richard Wang, professor expert in DQ of Massachusetts Institute of Technology (MIT) in 2012 (São Paulo, POLI/USP);

- Participation and talk in the **8 th International Conference on Ecological Informatics** in 2012 (Brasília);

- Participation and invited talk in the **Programa de Tutoria de Global Biodiversity Information Facility (GBIF) - Proyecto Australia** - Costa Rica. II Taller: Herramientas para la administración, divulgación y uso de información en biodiversidad in 2013 (San José, Costa Rica);

- Organization and invited talk in the **Workshop de Integração de Sistemas e Banco de Dados and Capacitação para Integração de Sistemas de Informação e Bancos de Dados de Biodiversidade** do Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) and Ministério do Meio Ambiente (MMA) in 2013;

- Participation in the **Meta-Analysis of Pollination Deficits Workshop** of the Global Environment Facility (GEF), United Nations Environment Programme (UNEP) and Food and Agriculture Organization (FAO) of the United Nations (UN) in 2013 (São Paulo, POLI/USP);

- Participation and invited talk in the **I and II Workshop sobre ferramentas computacionais para estudos palinológicos of the Redes de Catálogos Polínicos online (RCPol)** in 2013 (São Paulo, POLI/USP) and 2016 (São Paulo, Bayer: Brasil), respectively;

- Participation and talk in the **IV Workshop do Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr)** in 2014 (Petropolis, Laboratório Nacional de Computação Científica);
- Participation and talk in the **Taller Calidad de Datos: Mejorando los primarios sobre biodiversidad** in 2014 (Bogota, Colombia);
- Talk and discussion of the partial result of this research in a **meeting with staff of GBIF Secretariat** in 2014 (Copenhagen, Denmark);
- Creation and participation of the **TDWG/GBIF Data Quality Interest Group** in 2014;
- Creation and leading the **Task Group on Framework on Data Quality of the TDWG/GBIF Data Quality Interest Group** in 2015;
- Interchange in the **Museum of Comparative Zoology of Harvard University** between August 2015 and February 2016, leading a study case to evaluate the conceptual framework (Cambridge, USA);
- Talk in the **Regional capacity enhancement by setting up Uruguay's data portal of GBIF Capacity Enhancement Support Programme** in 2016 (Montevideo, Uruguay);
- Organization and talk in the **Biodiversity Data Quality Symposium: Developing a Common Framework to Improve Fitness for Use of Biodiversity Data** in 2016 (São Paulo, FAPESP);
- Organization and talks in the **Biodiversity Data Quality Workshop: Developing a Common Framework to Improve Fitness for Use of Biodiversity Data** during three days in 2016 (São Paulo, POLI/USP).