Biodiversity Data Quality Profiling: A practical guideline

Allan Koch Veiga, Antonio Mauro Saraiva

Departamenteo de Engenharia de Computação e Sistemas Digitais,
Escola Politécnica – Universidade de São Paulo (USP)
Av. Prof. Luciano Gualberto, travessa 3, 158, Cidade Universitária - São Paulo - SP - Brasil - CEP: 05508-900

{allan.kv,saraiva}@usp.br

Abstract. *The increasing availability of biodiversity data worldwide, provided by an in creasing number of institutions, 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 on the outcomes of these uses. To tackle this issues a conceptual framework was defined in the context of the Biodiversity Information Standards (TDWG) to serve as consistent approach to assess and manage data quality (DQ) of biodiversity data. Based on this framework we propose a method to define DQ Profiles that describes the meaning of "fitness for use" in a given context and enable the DQ assessment and improvement.*

Resumo. *A crescente disponibilidade de dados de biodiversidade em todo o mundo, providos por um número crescente de instituições, e o crescente uso desses dados para uma variedade de usos suscitaram preocupações relacionadas a "adequação ao uso" desses dados e o impacto nos resultados desses usos. Para abordar estas questões, definiu-se um framework conceitual no contexto do Biodiversity Information Standards (TDWG) para servir como uma abordagem consistente para avaliar e gerir a Qualidade dos Dados (QD) em dados da biodiversidade. Com base neste quadro, propomos um método para definir Perfis DQ que descrevem o significado de "adequação ao uso" em um dado contexto e consequentemente permitir a avaliação e melhoria da QD.*

1. Introduction

The research field called Biodiversity Informatics (BI), which aims at applying informatics concepts, techniques and tools to research and development on biodiversity, has existed for the past 40 years, and during that time much effort was concentrated 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 into a myriad of usages [17-24 do artigo].

In this context, the community around BI 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, such as the Global Biodiversity Information Facility (GBIF) [GBIF 2017].

However, the increasing amount of freely available data from an also increasing amount of sources, which may have different and unclear level of concerning with the

quality of their data, has risen concerns 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 [Ge and Helfert].

Performing DQ assessment became a critical issue in the BI context, specially because not enough information about the quality of data is provided, making it difficult to split a subset of data that is fit for use for different specific purposes. Furthermore, determining if data are fit for use is an action that is highly dependent of the "data use" and deal with DQ for all potential biodiversity data usages is impractical for the most 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 [Veiga 2017].

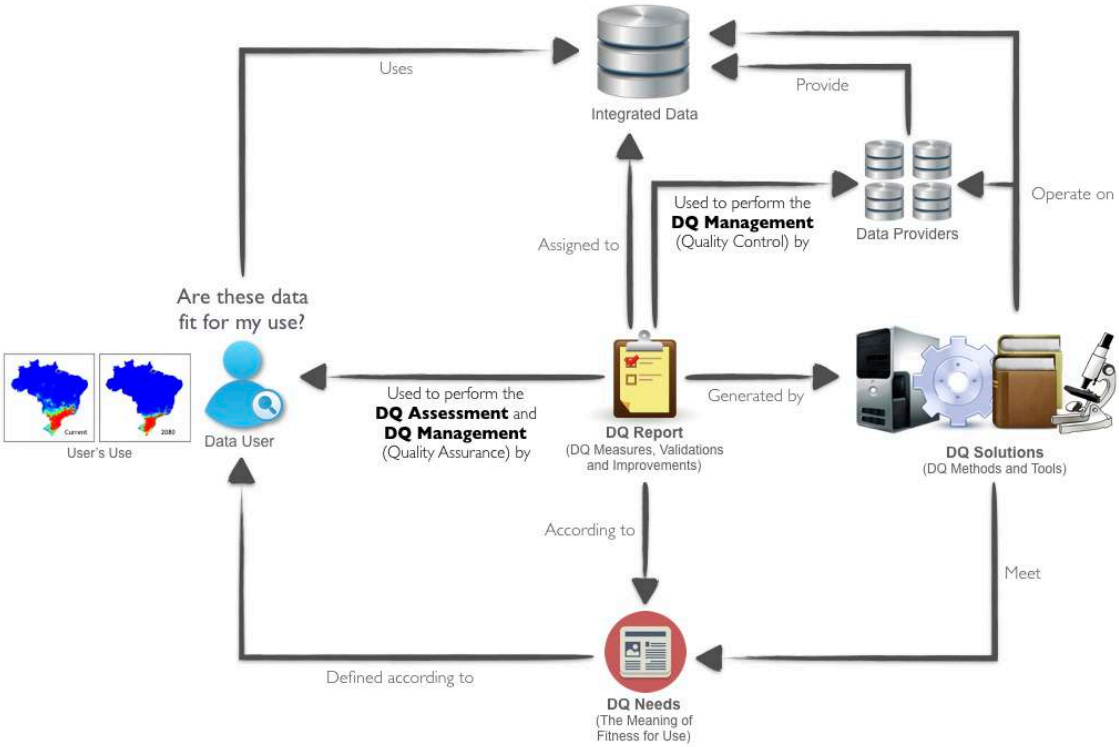

Figure 1. BI scenario and main components regarding DQ assessment and management: DQ Needs, DQ Solutions and DQ Report [Veiga 2017].

Based on a well defined user's DQ needs, DQ solutions must be delivered in order to meet the DQ needs. DQ solutions 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 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.

These three main components were formerly defined in a conceptual framework in the context of the Biodiversity Information Standards (TDWG) [Veiga 2017, Veiga and Saraiva 2016]. The original version of the framework is highly comprehensive and formal, composed by 29 interrelated concepts.

Due the comprehensiveness of the conceptual framework, it allows different interpretations and manner of use it according different stakeholders. To contextualize how different stakeholder can take advantage of the conceptual framework, we selected four stakeholders to describe their role in DQ context: DQ Profilers, Developers, Data Users and DQ Holders.

DQ Profilers are experts on DQ and/or as specific domain that uses biodiversity-related data and are interested into formalizing the way DQ is handling in a specific domain.

Developers are experts on to develop technical solutions for DQ and are interested into formalizing techniques and tools used for DQ and generate standardized and comparable outputs.

Data Users are experts on a specific domain which uses biodiversity-related data and are interested into assessing the quality of data and their fitness for use.

Data Holders are institution or people that holds, manage and curate biodiversity-related data and are interested into improving the quality and the fitness for use of data with efficiency.

For the purpose of this paper, we focus on the first stakeholder, DQ Profilers, proposing a method to define DQ Profiles based on the conceptual framework to formally describing a "meaning of data fitness for use" in a given context.

2. The conceptual framework: a brief practical overview

Formal details on the conceptual framework can be found at [Veiga 2017]. In this section will present a lite view of the framework according to a practical perspective. In this context, the framework will be approached according to three main components: DQ Profile, DQ Solutions and DQ Report, as illustrated at Figure 2.

DQ Profile defines a structure to describe the meaning of data fitness for use in a given context. A DQ Profile describes DQ needs requirements for a given context/scope. In order to implement and apply such requirements on data, it is necessary to use set of

DQ Solutions, that involves methods and mechanisms applied to meet DQ Profiles requirements.

DQ Solutions define a structure to describe methods (technical specifications) and mechanisms (tools that act on data) in order meet the DQ Profile requirements. DQ Solutions operate on Data Resources (both single records as multi records) and generates DQ Assertions assigned to each Data Resource. A set of selected DQ Assertions represents a DQ Report. DQ Report defines a set of selected DQ Assertions according to a DQ Profile requirements assigned to a Data Resources.

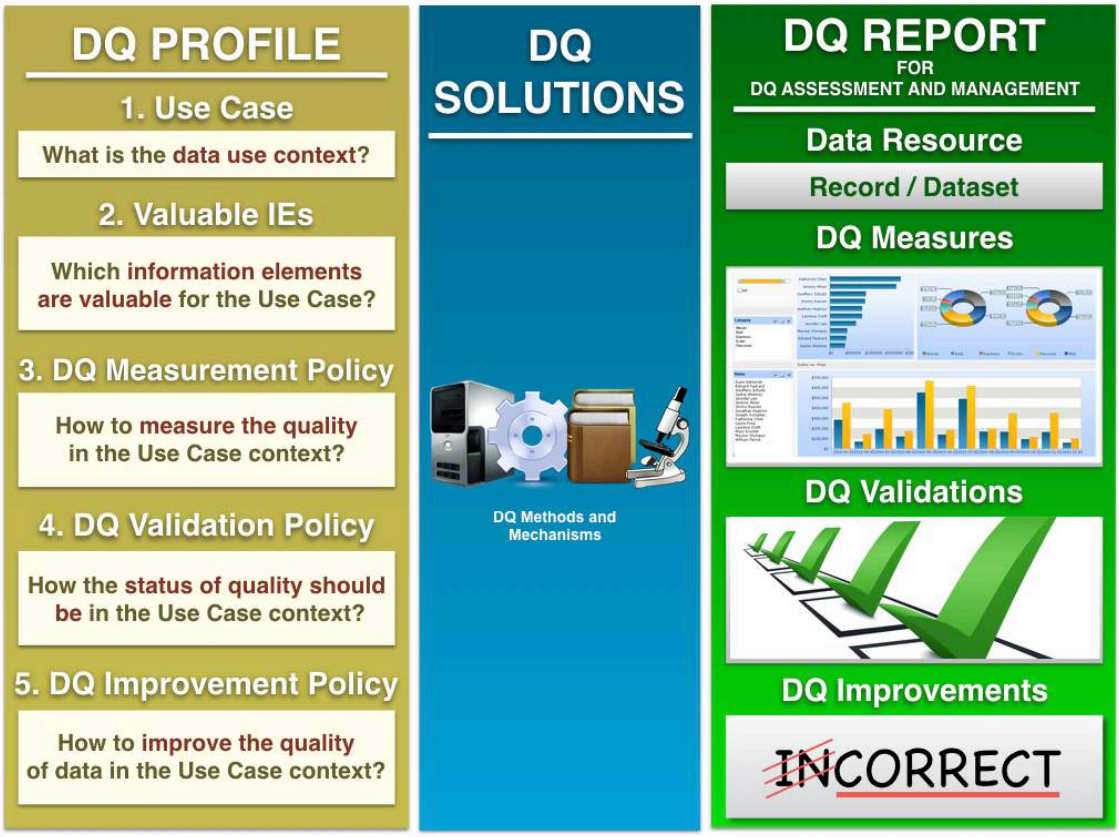

Figure 2. Structure of the Conceptual Framework

With a DQ Report assigned to a Data Resource, data users, holders, aggregators and custodians are enable to assess and improve the quality of the Data Resource according to the related DQ Profile definition.

Next section presents a practical method to define a DQ Profile in a given context.

3. Biodiversity DQ Profiling

Due to the idiosyncratic nature of the concept of “quality”, it is essential to understand what "data fitness for use" means according to the data user’s perspective in order to enable the DQ assessment and management.

In this contexts, defining “data fitness for use” involve to define three elements: use, data and fitness. Accordingly, DQ Profile encompasses these elements by five main

components: Use Case (use), Information Elements (IE) (data), DQ Measurement Policy (fitness), DQ Validation Policy (fitness) and DQ Improvement Policy (fitness).

In this context we propose a method to define a DQ Profile composed by five steps: (1) Define a Use Case; (2) Define the valuable IE in the context of the Use Case; (3) Define a DQ Measurement Policy in the Use Case context; (4) Define a DQ Validation Policy in the Use Case context and; (5) Define a DQ Improvement Policy in the Use Case context. Next we present a brief description of each step.

3.1. Defining a Use Case

By definition, it is necessary to clearly define what is the "data use context" to define the meaning of "fitness for use". The concept Use Case defines a context/scope delimitation for a DQ Profile.

A Use Case can represent a specific data use context, e.g., distribution model for the wild bee *Tetragonisca angustula s.l.* in Brazil; a generic data use or domain context, e.g. species distribution modeling, national species checklist definition, agrobiodiveristy etc; institutional context, e.g. Museum of Comparative Zoology of Harvard University [Veiga 2017], Botanical Garden of Rio de Janeiro; or an aggregator context, e.g. GBIF, Sistema de Informação sobre a Biodiversidade Brasileira (SIBBr) or Atlas of Living Australia (ALA).

3.2. Defining valuable IE

An Information Element (IE) is a representation of an element or set of elements in a formal representation of the data. An IE is a single element or set of elements present in the data 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 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 and Wieczorek 2006).

Defining valuable IE is performed by selecting a subset of IE that is required or valuable for the purposes in the Use Case context; therefore, this subset should be the target of DQ efforts, either for quality measurement, validation or improvement.

3.2. Defining a DQ Measurement Policy

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 et al. 1995; Strong et al. 1997).

Dimensions are measurable quality aspects of data (Wang and 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 and Wang 1996; Fox et al. 1994; Cai and 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 and Zhu 2015).

To define a DQ Measurement Policy, a set of relevant DQ Dimensions must be selected and defined according to Use Case context. To define a DQ Dimension in a Use Case context, it is necessary to describe the Dimension (e.g. completeness, consistency, conformity) in the context of an IE (e.g. coordinates, event date, country, scientific name) and Data Resource type (i.e. single record or dataset). For example, in a given Use Case context, ”Coordinates Completeness of Datasets” may represent the proportion of records with values supplied for decimal latitude and decimal longitude, in another Use Case context, ”Coordinates Completeness of Datasets” may represent the total number of records with values supplied for decimal latitude, decimal longitude and geodetic *datum*.

3.2. Defining a DQ Validation Policy

A DQ Criterion is a statement that describes acceptable DQ measures by which data are judged regarding their fitness for some use. DQ Criteria are used to validate if the quality of data are satisfactory to be used in a specific Use Case context. Data compliant with the Criterion means the data are fit for use according to the related DQ Dimension.

For example, "coordinate completeness of a dataset must be equal to 100%” is a Criterion used to validate if the measure of the DQ Dimension completeness in the context of IE coordinates and resource type dataset has the measure equals 100%. If data has a measure equals 100%, the data is compliant with the Criterion, else the data is not compliant with the Criterion, and consequently unfit for use.

DQ Measurement Policy is defined by selecting a set of Criteria to split data that is fit for use from data which is not fit for use for a particular Use Case context.

3.2. Defining a DQ Improvement Policy

DQ Enhancements are statements that describe activities required to improve DQ. An Enhancement can be a description of a procedure, 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.: "Fill taxon hierarchy based on the most specific name."
- **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 DQ Enhancement simultaneously.

A DQ Improvement Policy is defined by selecting a set of relevant DQ Enhancements for improving the measure of quality according to the DQ Measurement Policy and consequently make data more compliant to the DQ Validation Policy.

4. Final Remarks

The presented method enables to define the meaning of fitness (through the DQ Measurement, Validation and Improvement policies) of a set of valuable IE in the context of a specific Use Case context. All these components putting together defines a DQ Profile.

Based on a DQ Profile, a set of method and mechanisms (usually softwares) for measure, validate and improve the quality of Data Resources, generating customized DQ Reports suitable for the assessment and improvement of DQ in the Use Case context.

This work has been developed in the context of TDWG/GBIF Biodiversity DQ Interest Group [BDQ-IG 2017], more specifically, in the context of the Task Group 1 - Framework on DQ [TG1 2017].

References

Boulic, R. and Renault, O. (1991) “3D Hierarchies for Animation”, In: New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons ltd., England.

Dyer, S., Martin, J. and Zulauf, J. (1995) “Motion Capture White Paper”, http://reality.sgi.com/employees/jam_sb/mocap/MoCapWP_v2.0.html, December.

Holton, M. and Alexander, S. (1995) “Soft Cellular Modeling: A Technique for the Simulation of Non-rigid Materials”, Computer Graphics: Developments in Virtual Environments, R. A. Earnshaw and J. A. Vince, England, Academic Press Ltd., p. 449-460.

Knuth, D. E. (1984), The TeXbook, Addison Wesley, 15th edition.

Smith, A. and Jones, B. (1999). On the complexity of computing. In *Advances in Computer Science*, pages 555–566. Publishing Press.