Information Processing and Management 49 (2013) 1194–1205

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Dealing with metadata quality: The legacy of digital library efforts

Alice Tani, Leonardo Candela *, Donatella Castelli

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Consiglio Nazionale delle Ricerche, Via G. Moruzzi, 1-56124 Pisa, Italy

ARTICLE INFO

Article history:
Received 19 July 2012
Received in revised form 5 April 2013
Accepted 28 May 2013
Available online 2 July 2013

Keywords:
Quality
Digital Libraries
Metadata quality frameworks
Data infrastructures

ABSTRACT

In this work, we elaborate on the meaning of metadata quality by surveying efforts and experiences matured in the digital library domain. In particular, an overview of the frameworks developed to characterize such a multi-faceted concept is presented. Moreover, the most common quality-related problems affecting metadata both during the creation and the aggregation phase are discussed together with the approaches, technologies and tools developed to mitigate them. This survey on digital library developments is expected to contribute to the ongoing discussion on data and metadata quality occurring in the emerging yet more general framework of data infrastructures.

1. Introduction

Data and metadata represent a key element in our knowledge-based society. In the light of the critical role they play in domains including business, government and science (Nature, 2008; Hanson, Sugden, & Alberts, 2011; Hey et al., 2009; Borgman, 2010, 2011), dealing with their *quality* is fundamental. Being conscious of data and metadata quality aspects is a primary need in environments supporting and promoting sharing and reuse of data and metadata like modern data infrastructures do (Thanos, 2012; Ashley et al., 2012; Boulton et al., 2012). In particular, *metadata* – being data that give information about other data – cover a fundamental function in enabling any form of data management, and their "quality" deeply influences the overall quality of the services offered by relying on the data they characterize.

Despite that the relevance and impact of metadata quality is universally recognized in the literature, there is no agreement yet on what metadata quality is. This lack has several implications, including the impossibility of introducing systematic approaches to its automatic measurement and enhancement. Similarly to data quality (Madnick, Wang, Lee, & Zhu, 2009), metadata quality is a complex concept that can intuitively be defined as "fitness for use" (Wang & Strong, 1996; Eppler, 2006). Very often (Strong, Lee, & Wang, 1997; Batini & Scannapieco, 2006), (a) its understanding and assessment change from one community of practice to another, (b) its notion depends on the actual use of the data, and (c) an actual characterization can only be built by taking into account its multiple facets, and, therefore, by defining it in terms of a number of specific quality dimensions.

Digital Libraries (Candela, Castelli, & Pagano, 2011, chap. 1) have been conceived since the beginning as tools aiming at supporting and revolutionizing the practices through which citizens have access to human knowledge and produce new artefacts (Ioannidis, 2005). The typology of data they offer is not limited to texts, images and music only. Rather, a Digital Library is nowadays called to make available the rich array of data that is needed by the community of practice it is serving. Very often such data are borrowed from other Digital Libraries or repositories thus data are expected to be (re-)used in

* Corresponding author.
E-mail addresses: alice.tani@isti.cnr.it (A. Tani), leonardo.candela@isti.cnr.it (L. Candela), donatella.castelli@isti.cnr.it (D. Castelli).

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.05.003

domains different from their initial one. All this is achieved by heavily relying on metadata. Digital Libraries have faced a plethora of metadata quality issues and have developed solutions aiming at mitigating the effects of such issues.

The paper surveys how metadata quality issues have been addressed until now in the digital library domain. Such a survey investigates two diverse yet complementary elements: (i) the quality frameworks introduced to characterize "metadata quality" as to lay its foundations and promote a systematic approach to methods for the automatic evaluation and improvement of metadata quality; (ii) the approaches presented in the literature to actually deal with metadata quality issues, both to evaluate and to improve metadata quality. Through the analysis of the work done and of the lessons learned in the addressed context, we expect to contribute to solution of similar issues faced in other contexts such as the emerging yet more general framework of data infrastructures. This contribution range from ready to use solutions and approaches to typologies of strategies and methodologies to be eventually adapted and exploited in context different from the Digital Library one.

The rest of this paper is organized as follows. Section 2 introduces the concept of "metadata quality". Section 3 presents a number of quality frameworks that have been proposed to identify an effective way to define and measure metadata quality. Section 4 reviews metadata quality problems analyzed in the recent literature in the field of Digital Libraries and digital repositories, and also describes proposed possible solutions for specific quality problems, namely strategies for quality assurance in the metadata creation phase, quality evaluation, and cleaning. Section 5 concludes by highlighting research directions for data and metadata quality issue in data infrastructures.

2. Metadata quality in Digital Libraries

Metadata is a key element in the digital library domain. Actually, such a kind of data has characterized this domain since the beginning and for a long time it has been – in some cases this is still the case – the sole data digital library and repository systems have been requested to manage since they act as placeholders for real resources. Because of this core role, metadata quality is a characteristic that is directly associated with the digital library value and effectiveness, e.g., if metadata quality is poor so is the discovery of digital library information objects.

However, defining "what metadata quality is" is a very challenging task. It can be affirmed that no consensus has been reached on this concept until now, apart from the shared understanding that the difficulties in defining it come from its intrinsic characteristic of being a multidimensional and context specific concept. Bruce and Hillmann (2004, chap. 15) stated that "Like pornography, metadata quality is difficult to define. We know it when we see it, but conveying the full bundle of assumptions and experience that allow us to identify it is a different matter". In the rest of this section a brief survey of the evolution of the "metadata quality" concept understanding is presented.

Early discussion on quality of metadata – actually, bibliographic records since the term "metadata" was not largely diffused – mainly concerned the rising costs of making bibliographic descriptions and the need to provide access to the increasing volume of library materials in the context of the Library of Congress as well as large OPACs. To solve such issues Graham (1990) urged catalogers to distinguish truly important and necessary aspects of cataloging from those elements that were nonessential for the average user. Thus, in Graham's view, the conception of quality seems to be made independent of conformance to traditional cataloging rules rather be seen as related to the "fitness for use" understanding.

The theme of quality of metadata for networked resources remained a relatively unexplored research area until it was discussed within a study to assess metadata records from 42 Federal agencies' implementation of the Government Information Locator Service (Moen, Stewart, & McClure, 1997). The study concluded that "no consensus has been reached on operational and conceptual definitions of quality; likewise, validated procedures for assessing metadata are lacking". Actually, great interest in these results rose when they were presented at the IEEE International Forum on Research and Technology Advances in Digital Libraries, ADL '98 (Moen, Stewart, & McClure, 1998) as there were emerging environments characterized by increasing diversity of resources, data formats and application-specific functions, thus requiring quality criteria to consider contextual requirements – e.g., the specific functionality needed by the application, the nature of the described resources, the particular metadata formats conveying the information. Similar considerations had already been made by Moen et al. (1997) that concluded saying that "... given the force of user perspective on the representation of volatile information, and the lack of proven standards, systems of metadata ... may require uniquely tailored approaches to quality assessment"; however, "the results of this analysis of metadata content will contribute to a developing dialog about assessing the quality of metadata".

A stronger debate about metadata quality issues in networked domains emerged around 2003, possibly moved by the pioneering work performed by Dushay and Hillmann (2003) in creating the National Science Digital Library (NSDL) as an aggregator gathering, through the OAI-PMH protocol, large amounts of metadata from repositories of resources in the fields of science, technology and mathematics. In that context, the strict relation between quality and compliance to bibliographic description praxis is still present. As a matter of fact authors state that most quality problems arose in that context because "increasingly complex array of resources were being described by untrained people instead of well trained librarians, or by automated means with ill-documented methods". This statement assimilating "quality" with conformance to bibliographic principles, could apply to the quite uniform context characterizing how NSDL was being created (i.e., Dublin Core records describing scientific literature, aggregated through the OAI-PMH protocol). However, short later, Bruce and Hillmann (2004) recognized that metadata quality issues deriving from dependence on context are particularly evident in aggregated