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

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.

Keywords: Quality, Digital libraries, Metadata quality frameworks, 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 et al., 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.

Email address: {alice.tani, leonardo.candela, donatella.castelli}@isti.cnr.it (Alice Tani, Leonardo Candela, Donatella Castelli)

This lack has several implications, including the impossibility of introducing systematic approaches to its automatic measurement and enhancement. Similarly to data quality (Madnick et al., 2009), metadata quality is a complex concept that can intuitively be defined as "fitness for use" (Wang and Strong, 1996; Eppler, 2006). Very often (Strong et al., 1997; Batini and 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 et al., 2011b) 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 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 characterised 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) 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 et al., 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 et al., 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 environments. Some years later, reflecting on the NSDL experience, Lagoze et al. (2006) recognized that the need for expressing context for resources was not taken into account in the NSDL experience. In the meanwhile this need was largely being recognized in the literature (cf. Section 4).

Just starting from Bruce and Hillmann (2004), a number of studies and research efforts have been performed aiming at systematizing metadata quality concepts. All these studies have resulted in proposing frameworks to identify and assess quality parameters and metrics in specific contexts, rather than defining what metadata quality is. And such results could not be different, being "fitness for use" the ultimate, vague meaning of quality. These frameworks are presented in the next section.

3. Quality Assessment Frameworks

Digital Libraries are devoted to manage different typologies of data and corresponding metadata, possibly coming from different contexts. How evaluating metadata quality in such environments has been explored since early 2000 when networked repositories were started to diffuse.

In their analysis based on the Learning Objects and e-Prints communities of practice, Barton et al. (2003) point out that, in an environment where each metadata repository or archive is part of a wider system that aims at interoperability, quality assurance for metadata is a much more difficult issue than in a local context. Similarly, Stvilia et al. (2004) recognize that one of the most ticklish issues in the theory of information quality is how to account for the context-sensitive nature of information quality and value. As observed by most authors, it may happen that in the original context metadata quality is high, but in an aggregated environment the same metadata has low quality. This generally happens because quality parameters that are valid in the original context

Framework	Parameters	Metrics
Bruce and Hillmann (2004) (cf. Sec. 3.1)	7	n.a.
Ochoa and Duval (2009) (cf. Sec. 3.2)	7	13
Stvilia et al. (2007) (cf. Sec. 3.3)	22	41
Hughes (2004) (cf. Sec. 3.4)	7	7
Bethard et al. (2009) (cf. Sec. 3.4)	7	7
DL.org (Candela et al., 2011a) (cf. Sec. 3.5)	20	0
5SQual (Moreira et al., 2009) (cf. Sec. 3.5)	10	10

Table 1: Summary of metadata quality assessment frameworks

may be different from the parameters adopted in the aggregated environment, where metadata will likely be used to reach a purpose different from the one for which they were originally created. Furthermore, in the original context there may be information which is not explicitly specified because it is considered as assumed knowledge, e.g., the controlled vocabularies used in metadata fields, while in the aggregated environment this unspecified information leads to a deterioration in quality because it does not allow a correct data interpretation. Therefore, as digital repositories grow in size, number and diversity, and aggregated environments become increasingly widespread, the problem of ensuring a sufficient general level of quality becomes fundamental.

The remainder of this section presents frameworks to assess metadata quality. In all of them, the definition of the qualities to be assessed, i.e., *parameters* or *dimensions*, and methods to assess them, i.e., *metrics*, is the most critical activity. The first frameworks discussed were devised for repositories aimed at meeting the requirements of finding, identifying, selecting, obtaining information objects through their metadata. Namely: (*i*) the *Bruce and Hillmann framework* (cf. Sec. 3.1) that elaborates on 7 generic metadata quality characteristics ranging from completeness to accessibility; (*ii*) the *Ochoa and Duval framework* (cf. Sec. 3.2) that complements Bruce and Hillmann framework by proposing 13 quality metrics to evaluate the quality of item-level metadata in a collection; (*iii*) the *Stvilia et al. framework* (cf. Sec. 3.3) that identifies 22 quality dimensions and proposes 41 metrics for their calculation, of which 30 are based on object or collection attributes. Then, two studies are briefly cited coming from specialized communities, namely, the Open Language Archives Community and the Community of educational digital libraries (cf. Sec. 3.4). The remaining ones have grown in the context of modern digital libraries, intended as frameworks assessing the quality of the entire service (cf. Sec. 3.5). A comparative summary is presented in Table 1 while concluding remarks on the discussed frameworks are in Section 3.6.

3.1. Bruce and Hillmann Framework

Among the first attempts to give a general definition of metadata quality dimensions is Bruce and Hillmann (2004). They identify seven general characteristics of metadata quality: *completeness*, *accuracy*, *provenance*, *conformance*

to expectations, *logical consistency and coherence*, *timeliness*, and *accessibility*. These seven dimensions aim to be domain-independent, i.e., "one might think of these characteristics as places to look for quality in collection-specific schemas and implementations, rather than checklists or quantitative systems suitable for direct application".

The authors accurately introduce and describe each dimension and its characteristics, but give no formal definition nor metrics. In addition to the seven dimensions, they identify several levels of quality for metadata, i.e., the element set or *metadata scheme level*, the *syntactic level*, and the *data values* themselves. In fact, they affirm that any definition of quality should evaluate the attributes of metadata at such different levels.

The authors note that it is not possible to state which of the seven dimensions they describe is most important, nor which most urgently needs to be present for a given application, since the importance of each quality criterion is strictly influenced by the nature of the resource to be described, as well as by the environment in which the metadata is to be constructed or derived. Thus great emphasis is put on the fact that perception of quality strictly depends on context.

As an application of the seven quality dimensions composing their framework, these authors propose a "what to ask for and where to look" compliance checklist: in practice, they suggest a series of questions that might be asked in order to check whether metadata under assessment is in conformance with the established criteria, as well as several quality compliance indicators that might be used to answer the questions. Compliance indicators may be automated means, or human techniques, or both, and they include the use of graphical analysis software, as well as the presence of controlled vocabularies, provenance information at several levels of detail, and advanced documentation such as an expression of the metadata intentions. Clearly, the checklist questions imply subjective evaluation and do not provide any way for giving a quantitative measurement of the quality criteria.

About the above framework, Hillmann and Phipps (2007) point out that "Although the criteria provide opportunities to converse about quality, without ways to measure that quality, they remain frustratingly beyond reach". Therefore, they suggest to consider a view where *Application Profiles* (Heery and Patel, 2000) are seen as a "template for expectation", and where metadata under assessment can be compared to that template for obtaining quantitative measurements of the quality parameters. In particular, Hillmann and Phipps see a potential for assessing the following dimensions: (*i*) *completeness*, by relying on the use of "obligation" imposed by the profile it is quite straightforward to verify whether a metadata is complete or not; (*ii*) *conformance to expectations*, by relying on descriptions of conditions that should occur when a value is present it is possible to verify whether a metadata satisfies the expectations or not; (*iii*) *accuracy*, for instance it is possible to quantify the level of invalid vocabulary terms when a vocabulary encoding scheme is specified. With regard to the other dimensions, these authors admit that it may be too difficult or even impossible to assess them by relying on Application Profiles for expressing

expectations.

There are studies proposing to supplement the Bruce and Hillmann Framework dimensions with other related to *shareability*. Grounded in experiences in cultural heritage institutions, Shreeves et al. (2005) observed that metadata may be of high quality within a local database but this quality may be lost when metadata are combined in a federated environment. Thus, "understanding of the criteria for high quality, 'shareable' metadata is crucial to the next step in the development of federated digital libraries". Accordingly, Shreeves et al. (2006) suggest shareability as an adjunctive metadata quality dimension. In the authors' view, shareable metadata is metadata which can be understood and used outside of its local environment by aggregators to provide more advanced services. That is, shareable metadata should be useful and usable to services outside of its local context given the resource described. For this, they suggest the following characteristics specifically conceived to assess shareability in addition to characteristics of quality metadata: (*i*) content is optimized for sharing; (*ii*) metadata within shared collections reflects consistent practices; (*iii*) metadata is coherent; (*iv*) context is provided; (*v*) the metadata provider communicates with aggregators through direct or indirect means; and (*vi*) metadata and sharing mechanisms conform to standards. It is to note here that the concept of "shareability" has become a basic one in the cultural heritage community as a result of increased expectations for making descriptive metadata openly available (Riley and Shepherd, 2009). The centrality of such a concept, as well as that of "interoperability" which is strictly connected to it, has revealed in the field of digital library infrastructures as to become one of their main issues (Candela et al., 2010).

3.2. Ochoa and Duval Framework

The framework defined by Ochoa and Duval (2009) is strictly related to the Bruce and Hillmann one (cf. Sec. 3.1). While Bruce and Hillman devised their framework to guide human reviewers, Ochoa and Duval work aims at proposing a framework that comprises meaningful quality parameters, i.e., quality parameters that might be used by human reviewers, associated with automatic calculable measures of quality. In particular, they complement the Bruce and Hillmann framework by proposing automatic measurement methods for the seven parameters of such a framework.

For each quality parameter one or more metrics are proposed with a rationale, the calculation formulas and some guidelines.

The authors point out that the proposed metrics are not intended to be a comprehensive or definite set, but should be considered as "a first step for the automatic evaluation of metadata quality". In fact, the following characteristics and limitations are observed: (*i*) they are easy to implement in real environments and can be used for a wide range of digital repositories; (*ii*) they are are standard- and community-of-practice- agnostic, even though the parameters needed to initialize the calculations are context-dependent; (*iii*) they are mainly designed to work over metadata in the form of text and numbers. For metadata containing non alphanumeric information new approaches are needed;

(*iv*) they are mainly conceived for metadata instances conforming with a relatively stable metadata schema; (*v*) the normalization of the metrics may not always be possible; and (*vi*) the mix of the quality parameters is the general quality of the metadata instance, although no proposal is made on how to mix them since there may be several tradeoffs between the characteristics.

Ochoa and Duval (2009) conducted three validation studies to evaluate the proposed metrics with respect to: the correlation with human-made quality assessment; the effectiveness in discriminating key properties of diverse metadata sets; and the effectiveness as automatic low-quality filters. The following results have been obtained: (*i*) some metrics correlated well with human reviewers while others seems completely unrelated. In particular, the *Qtinfo*, i.e., the metric measuring the information content of the text fields, seems to be a very good approximation of human perceived quality. Thus human reviewers tend to evaluate metadata as content, i.e., longer and specialized text receive an higher score than a shorter one; (*ii*) the metrics are effective in discriminating manually generated metadata (expected to have an high quality) from automatically generated ones. In particular, the metrics point out that human experts filled more fields than the automatic approach as well as they tend to select a richer set of categorical values; (*iii*) some of the metrics (*Qcomp*, *Qwcomp*, *Qtinfo*) as well as the average of all the proposed ones are an effective approach for building automatic quality filters; and (*iv*) some quality parameters are very difficult to be evaluated by humans, e.g., the variability of categorical values, while the metrics were able to capture them.

3.3. Stvilia et al. Framework

Stvilia et al. (2007) proposal has been driven by the need to define a general framework and an effective measurement model, which is a pre-requisite for information quality (IQ). By revising previously defined IQ assessment frameworks in the data management field (Eppler, 2006; Wang and Strong, 1996) they observe that most of them are "ad hoc, intuitive, and incomplete and may not produce robust and systematic measurement models". Therefore, in contrast to context-specific quality assessment models depending on variables determined by local needs, these authors focus on studying the causes of quality changes, and present a framework consisting of typologies of *IQ problems*, related activities, and a systematically organized taxonomy of *IQ dimensions*. In this framework, an IQ problem is said to be occurring when the IQ of an information entity does not meet the IQ requirements of an activity on one or more IQ dimensions, and an IQ dimension is defined as any aspect characterizing the IQ concept. In any case, the authors clearly recognize that information quality is contextual, and state that their framework could be used as "a knowledge resource and guide for developing IQ measurement models for many different settings".

Four major sources of IQ problems are identified: *mapping*, *changes to the information entity*, *changes to the underlying entity or condition*, and *context changes*.

From the analysis of these sources, the authors develop a taxonomy of 22 IQ dimensions[1], systematically organized into three categories: (*i*) *intrinsic*, i.e., dimensions that can be assessed by measuring information aspects in relation to reference standard (e.g., spelling mistakes); (*ii*) *relational*, i.e., dimensions that measure relationships between the information and some aspects of its usage (e.g., accuracy); and (*iii*) *reputational*, i.e., IQ dimensions that measure the position of an information entity in a given structure (e.g., authority).

In addition to the taxonomy, a set of 41 general metric functions (30 of them are object- or collection-based, 11 are process-based) are proposed. The authors also provide a Java implementation of these functions.

Moreover, from the analysis of the sources of IQ problems the authors propose a clustering of information activities that are affected by such problems: (*i*) *representation dependent*, i.e., activities depending on how well an information represents an entity or a condition; (*ii*) *de-contextualizing*, i.e., activities depending on the use of information in contexts different from the one the information is produced for; (*iii*) *stability dependent*, i.e., activities depending on the level of stability of information; and (*iv*) *provenance dependent* i.e., activities depending on the quality of information provenance.

Such a framework can be used for analyzing a specific context and developing an appropriate IQ measurement model. System activities are analyzed to identify those prone to quality problems and to select the relative IQ dimensions and metrics. Two concrete use cases (assessing the quality of aggregated metadata records, assessing the quality of English Wikipedia articles) are discussed in Stvilia et al. (2007).

Many authors, e.g., Park (2009); Ochoa and Duval (2009), assert that it is interesting to note the overlap between this framework and the Bruce and Hillmann one (cf. Sec. 3.1). In particular, the relation between the frameworks is highlighted by Shreeves et al. (2005) that graphically depict the mapping between the parameters of the two.

3.4. Other Approaches to Quality Assessment

The frameworks presented in this section diverge from the way traced by Bruce and Hillmann. In particular, in the framework presented by Hughes (2004) metadata quality assessment is related to context.

Hughes discusses the metadata quality assessment issue in the context of Open Language Archives Community (OLAC). He posits that "any measure of metadata quality benefits from both contextual and referential assessment – metadata on a per record and per collection basis is legitimately assessed against the baseline of broader community of practice, as well as for compliance to any external standard". Accordingly, he proposes 7 metrics (*Archive Diversity*, *Metadata Quality Score*, *Core Elements Per Record*, *Core Element Usage*, *Code Usage*, *Code and Element Usage*, *"Star Rating"*) and an algorithm aiming at

[1]The 22 dimensions are taken from a taxonomy previously derived from an analysis of 32 representative quality assessment frameworks from the IQ literature (Stvilia et al., 2004).

giving a metadata record a score between 0 and 10 representing the adherence of the metadata to best practices for Dublin Core and domain-specific controlled vocabularies.

Bethard et al. (2009) presents a different path toward automatic characterization of resource quality in the realm of educational digital libraries to help the identification of resources to use for learning goals. The quality indicators considered as most predictive are: *has prestigious sponsor*, *content is appropriate for age range*, *has sponsor*, *identifies learning goals*, *has instructions*, *identifies age range*, *organized for learning goals*. Such indicators are "boolean", i.e., each of them can be present or not when assessing a given resource. Moreover, the authors propose an approach relying on machine learning models for assessing the presence of such indicators.

3.5. Digital Library Quality Frameworks

In the Digital Library domain, frameworks aiming at assessing the quality of the entire Digital Library service have been proposed.

The DL.org Digital Library Reference Model (Candela et al., 2011a) is a comprehensive framework aiming at laying the foundations of the whole Digital Library domain. Among its core concepts there is the quality domain, i.e., the set of concepts characterizing Digital Libraries from the quality point of view. Such a domain is quite extent yet basic, it formalizes the following characteristics: (*i*) quality aspects can be associated with any "resource" contributing to form a Digital Library; (*ii*) quality aspects worth to be captured cannot be identified a-priory and are described via a number of "quality parameters"; (*iii*) quality parameters can be assessed by any "actor" (human or inanimate entity such as a software program); (*iv*) quality parameters are associated with a "measure" that is assessed according to a "measurement" which can be subjective or objective as well as quantitative or qualitative. In addition to that, the model presents a number of quality parameters (more than 40), clustered according to the domain they are primarily conceived for (e.g., generic, functionality) by clearly stating that this list is open, i.e., a community of practice can extend it with specific parameters. For the same reason, the model does not prescribe nor describe any quality measurement that is needed; rather, it leaves this decision to the community of practice that will instantiate the model in its application domain. Approximately half of the quality parameters are potentially suited for assessing metadata quality either because are "content quality" parameters (11 parameters including *authenticity*, *integrity*, *freshness*) or because are "generic" ones (9 parameters including *reputation*, *compliance with standard*, *sustainability*).

The 5SQual (Moreira et al., 2009) is a tool supporting the evaluation of core elements in Digital Libraries, i.e., digital objects, metadata, and services. Such a tool actually implements a theoretical quality model for Digital Libraries (Gonçalves et al., 2007) which has been defined by having the 5S framework (Gonçalves et al., 2004) as the underlying model for characterizing Digital Libraries. In such a model a number of quality dimensions have been proposed and associated with Digital Library concepts. In particular, 3 quality parameters

are associated with "metadata specification", i.e., *accuracy*, *completeness*, and *conformance*, and 7 quality parameters are associated with "digital objects", i.e., *accessibility*, *pertinence*, *preservability*, *relevance*, *similarity*, *significance*, *timeliness*. For each of them, the authors propose potential approaches for their calculation and the 5SQual offers an implementation of a subset of them.

3.6. Remarks on Frameworks

There are two main aspects characterising every data quality assessment framework: the definition of the qualities to be assessed through the framework, i.e., *parameters* or *dimensions*, and the methods to be used to assess the identified qualities, i.e., *metrics*. Although among the discussed frameworks there is no one that is expected to supplant the others, it can be observed that: (*i*) diverse frameworks tend to share a number of quality parameters, e.g., accuracy, completeness; (*ii*) the dimensions captured by a framework tend to grow in number when the goal is to accommodate the needs of diverse communities of practice; (*iii*) it is not expected to reach a global agreement on which dimensions are to be defined and their exact meaning, simply because this is a community specific requirement, (*iv*) there is the need to develop frameworks having multiple metrics for assessing a given quality parameter as to be able to accommodate diverse settings; (*v*) there is the strong need to develop a comprehensive set of mappings supporting the transformation of quality parameters assessed according to a given framework into quality parameters assessed via another framework.

Batini et al. (2009) provide a comprehensive and systematic description of methodologies for selecting and applying data quality assessment and improvement techniques. In essence, they draw similar conclusions, e.g., there is a set of basic quality dimensions that occurs across the frameworks like accuracy, completeness, consistency and timeliness, yet no shared meaning. This confirms that in essence, data and metadata quality issues are two very close worlds. Moreover, they observe that the whole data quality research field is evolving, it cannot be considered mature, and it is moving towards considering a wider number of data types and information systems.

4. Approaches to Metadata Quality Issues

The set of frameworks discussed so far are mainly conceived to characterize the metadata quality concept and provide methods for measuring to what extent a given resource has to be considered a quality one according to the given characterization. In this section, we discuss concrete approaches aiming at dealing with metadata quality issues.

Yasser (2011) analysed and compared problems reported in the literature and identified five categories of problems. These categories are: (*i*) *incorrect values*, i.e., metadata records contain values that do not represent a given resource correctly even though elements are applied correctly; (*ii*) *incorrect elements*, i.e., the values assigned are appropriate to describe the resource but have been

assigned to the wrong element; (*iii*) *missing information*, i.e., the metadata record is not complete; (*iv*) *information loss*, i.e., some details characterising the information are lost (this generally occurs during the process of converting metadata from one scheme to another); (*v*) *inconsistent values*, i.e., different values associated with an element may equally represent a characteristic of the resource, but they may be different enough in recorded form to undermine system functionality.

To resolve these problems a number of approaches are described in the literature. Depending on the solution proposed, four categories can be identified: approaches aiming at achieving a common understanding of metadata (cf. Sec. 4.1); approaches aiming at highlighting the problems affecting metadata (cf. Sec. 4.2); approaches aiming at supporting the generation of metadata (cf. Sec. 4.3); and, approaches aiming at repairing and homogenising metadata (cf. Sec. 4.4). Remarks on these categories are given in Section 4.5.

4.1. Metadata Guidelines, Standard and Application Profiles

Metadata guidelines are agreed policies potentially governing every aspect of metadata including the values and elements to be used to characterise the resources the system manages. Cataloguing guidelines have a long and well recognized tradition, and their role in the creation of quality metadata has necessarily been recalled in the context of digital libraries and repositories, where the practice of metadata creation by authors is much diffused. In fact, guidelines are among the most used approaches (Park and Tosaka, 2010b) and they can be a very effective tool to convey rules and principles thus to establish a common knowledge.

Metadata standards and applications profiles are two approaches for realising metadata guidelines, e.g., Guy et al. (2004) and Hillmann and Phipps (2007) discuss on how to establish guidelines by using application profiles.

Some guidelines are associated with metadata standards, e.g., Dublin Core metadata guidelines (Apps, 2005). Others are tailored to promote a certain use of a metadata standard in a given context, e.g., the CDP Dublin Core Metadata Best Practices (CDP Metadata Working Group, 2006) provide guidelines for digitized cultural heritage resources by using the Dublin Core element set. Others are oriented to enhance the quality of metadata offered via web-based services, e.g., the CrossRef guidelines define nine easy steps aiming at enhancing the produced metadata (crossref.org, 2012). Others are oriented to support the realisation of services resulting from aggregating metadata from different "providers", e.g., Vanderfeesten et al. (2008) defined guidelines characterizing a number of aspects including metadata standards, OAI-PMH implementation, best practices and semantics such as the use of "inverted name" syntax for "creator" which is also a mandatory element.

However, these approaches are not problem free. Park and Tosaka (2010a,b) evaluate metadata creation practices in digital repositories and collections. Such studies – conducted on cataloging and metadata professionals dealing with a great variety of digital objects – highlighted that although the analysed sample uses a wide range of metadata standards, only a few are widely used, namely

MARC and DublinCore. DublinCore, although widely used, was considered difficult to apply because of semantic overlaps and ambiguities.

4.2. Metadata Evaluation Approaches

Metadata evaluation approaches consists in assessing quality dimensions with computer assisted facilities (Bruce and Hillmann, 2004). However, Nichols et al. (2009) observe that because of the community specific nature of the metadata quality dimensions, evaluation approaches aiming to identify general quality problems may not meet the requirements of a specific application environment. Thus such tools can be "only" a valid help for identifying metadata problems, but informed interpretation is necessary to understand the actual problems and take the correct decisions in the specific context.

These approaches fall in two main categories: *analytic-oriented approaches* and *crowdsourcing-oriented approaches*.

Analytic-oriented approaches are aimed at extracting quality dimensions. For instance, Hillmann and Phipps (2007) underline the potential of *Application Profiles* to support metadata quality automated validation. They observe that when metadata has to conform to a specific Application Profile an automatic validator should be able to validate metadata characteristics such as the presence of appropriate encoding schemes, as well as the correctness of the vocabulary terms themselves. For example, the validator should be capable to determine whether a metadata element is qualified by the correct encoding scheme, or whether a value term is correctly expressed according the related syntax encoding scheme. However, the authors have to recognize that not all characteristics can be validated automatically, e.g., an automatic validator may not be able to determine the correctness of a date expressed in free text format.

Dushay and Hillmann (2003) proposed SpotFire, a software instrument that produces visual representations of data thus allowing humans to recognize visual patterns and derive appropriate conclusions. In particular, such a tool offers a number of data visualization and analysis facilities including operations such as checking for conformance to a particular controlled vocabulary or string pattern, and looking for anomalies in data such as typographical errors and bad values. A similar tool is also proposed by Nichols et al. (2008). The main functionalities offered are a summary description of metadata elements, a sorted presentation of metadata element lists and a completeness oriented visualization.

Crowdsourcing-oriented approaches are based on user feedbacks. Feedbacks can be obtained from the activities of system' final users, as well as from the work of digital library and repository administrators. For instance, Manghi et al. (2010a) proposed a facility which allows final users to submit data curation feedbacks in the form of "delete", "add" and "update" annotations in order to help improve the quality of the aggregated content. Savino and Schulze (2011) propose two tools for collecting feedback: (*i*) a "content checker" aiming at helping archivists to discover and signal errors in their metadata elements once represented in the aggregated metadata format as the result of existing mapping rules; (*ii*) a "vocabulary checker" aiming at helping managers to discover elements that do not match the agreed controlled vocabularies.

4.3. Semi-automatic Metadata Generation Approaches

Semi-automatic metadata generation approaches promote the creation of metadata by combining software facilities with human practices (Greenberg et al., 2006; Park and Lu, 2009). These facilities range from metadata editors to tools aiming at automatically generating metadata.

Park (2009) discusses an approach for using metadata guidelines by embedding them within a tool for semi-automatic metadata generation, so to enable catalogers or document authors to create metadata in compliance with the guidelines. Greenberg et al. (2001) studied how a simple Web form with textual guidance and selective use of features such as drop-down menus and pop-up windows could assist document authors in the generation of quality metadata.

Greenberg (2004) discusses two methods for automatic metadata generation: *metadata extraction*, i.e., metadata are produced by relying on the resource content, and *metadata harvesting*, i.e., metadata are collected from human-created tags embedded in the header source code of the resources. Greenberg et al. (2006) provide a very brief overview of automatic metadata generation approaches and revise both experimental research and application developments with the goal to identify functionalities for tools supporting automatic metadata generation. Park and Lu (2009) analysed the extent and types of research initiatives and systems, and discuss their practical application. All these studies conclude that (*a*) automatic processes will never replace human evaluation or production, rather they have to aid humans while creating metadata; (*b*) two fundamental functionalities are (*i*) helpers supporting the acquisition of metadata that a human can evaluate and edit; and (*ii*) helpers supporting the integration of content "standards" (e.g., subject thesauri, name authority files) into the metadata generation applications.

4.4. Metadata Cleaning, Enhancement, and Augmentation Approaches

This family of approaches complements the others since they focus on "repairing" existing metadata rather than identifying potential issues like evaluation approaches or producing quality-level metadata like guidelines and semi-automatic generation approaches. Because of this characteristic, its coverage is quite large and multifaceted.

Among the issues receiving a lot of attention due to their difficulty there are those related with name disambiguation. Lee et al. (2005) observe that name ambiguity in bibliographic citations can be divided into two specific subproblems: *mixed citation* (a.k.a. homonym problem) and *split citation* (a.k.a. synonym problem). Mixed citation occurs when the same name refers to more than one person, family or organization; this may happen due to abbreviations or because the different entities have exactly the same name spelling. Split citation, on the other hand, occurs when a person has different name variations which are treated as if they belonged to different persons; this may be due to pseudonyms, differences in language or script, transcription errors, abbreviations, as well as change in the order of the name components or change of name for many reasons such as marriage or divorce. The authors present solutions

based on one of the state-of-the-art sampling-based approximate join techniques declaring them as scalable yet highly effective. Laender et al. (2008) propose a solution for name disambiguation consisting in a heuristic-based hierarchical clustering (HHC) method, stemming from the following considerations: (*i*) it is very rare that two authors with very similar names and sharing a common co-author are two different persons in the real world and (*ii*) authors tend to publish in the same subjects and publication venues for a considerable portion of their careers. The authors claim that HHC performs competitively when compared with existing supervised machine-learning methods, without requiring any training phase. Beall (2010) analyzes strengths and weaknesses of manual and automatic approaches and concludes that a hybrid approach may become the most successful and widely used, especially for resources in the open world of Internet. Beall also discusses some features of name metadata records (such as birth and death dates, family, life events, institutional affiliation) and how these might help. Moreover, Beall highlights the role that services like the Library of Congress Authorities, the Virtual International Authority File (VIAF), or Wikipedia, might have. Kan and Tan (2008) propose to use uninformed string matching (e.g., the cosine distance or the edit distance) and informed record matching (i.e., record similarity is calculated by combining string similarity in a weighted formula thus to consider the different type of elements). Smalheiser and Torvik (2009) surveyed the literature related to name disambiguation and proposed a probabilistic model based on a multi-dimensional vector space for features representation. Recently, Manghi and Mikulicic (2011) presented an open source tool for authority control which aims to (*i*) offer administrative user interfaces for customizing the structure of authority files, (*ii*) tune-up probabilistic disambiguation of authority files through a set of similarity functions for detecting record candidates for duplication and overload, (*iii*) curate such authority files by applying record merging and splitting actions, and (*iv*) expose authority files to third-party consumers in several ways.

Another problem that has been extensively discussed is the homogenisation and enhancement of metadata when they are integrated into an aggregative system aiming at offering a unifying access to the resources and a number of added value services (Manghi et al., 2010b). Hillmann et al. (2004) propose to apply "safe transforms" to metadata records, i.e., an automated technique which is designed for addressing some of the common quality problems (e.g., missing data, incorrect data, confusing data, insufficient data), and which can be applied to enhance the information of the original metadata with no risk of degradation, e.g., (*i*) by *removing "noise"* like metadata with no information, (*ii*) by *detecting and identifying controlled vocabularies*, and (*iii*) by *normalizing metadata presentation*. Moreover, they propose an approach for *metadata recombination and augmentation* which is based on the idea that a metadata record can be built by aggregating metadata "statements" included in different "records" coming from diverse providers. A similar idea comes from Phipps et al. (2005), that suggest to create an "orchestra" of automated services for aggregating source statements into enhanced descriptions and exposing them to users. The "orchestra" include services for metadata augmentation, safe

transformations, equivalence services, crosswalking, archiving and persistence checking, annotation services, and metadata improvement and rating. Hillmann (2008) analyses the difference between *transformative* processes (e.g., modifying metadata based on the structure or values already available in statements) and *augmentative* ones (e.g., adding information based on information gleaned from the resource itself). This distinction is relevant for determining the sequence of processes. Transformations do things like: detection of controlled vocabulary values and attribution of those values to a particular vocabulary; detect and fix common typographical errors; deprecation of "promiscuous defaults", e.g., values that provide no information value, added to metadata only to fill a slot or provide functionality only. Instead, augmentation includes: machine-based processes that add values, for example, topics or formats; human-based augmentation, such as the addition of topics, relationships to educational standards. In some cases transformation can be orchestrated by humans. For instance, Savino and Schulze (2011) described a "metadata editor" through which an authorized user can interact with the entire set of aggregated metadata and modify existing metadata records as well as create new records. While editing an existing record, a user can correct the metadata values, as well as enrich the metadata elements, and then store the modified record back into the aggregating system. The editor automatically performs a check on incorrect values and on missing mandatory elements by relying on established policies and guidelines. de Groat (2009) gives a description of desired services aiming to remediate the metadata in order to achieve certain expectations with respect to the quality of service offered by the aggregator service. Such a report surveys desired and existing tools for the following metadata element typologies: topical subjects, genre, names, geographical information, dates, title information, type of resource, addressable raw object, rights, and identifiers.

The enhancement or augmentation of metadata records can be performed on specific elements of the records. For instance, in case where it is requested to have access to the real resource described by the metadata, it is possible to complete deficient records. Laender et al. (2008) propose a strategy consisting in an extensible service called PaperMetaSearch. This service searches for a document full text on the Web, by submitting parameterized queries to existing search engines (e.g., Google, Yahoo), with the employment of metadata information already available to the user as potential query arguments, e.g., the title and the list of the authors of the document.

4.5. Remarks on Approaches for Resolving Metadata Quality Problems

Each of the approaches discussed above has been introduced to solve specific quality problems. None of them can be considered as the ultimate and optimal solution to all quality issues, especially in complex and heterogeneous contexts as those addressed by the new evolutionary types of digital library systems, like the data infrastructure ones. A summary of the pros and cons of the presented approaches is given in Table 2. Methods forcing metadata with shared meaning (cf. Sec. 4.1) are potentially effective since explicitly declare the agreed aspects,

Metadata Guidelines, Standard and Application Profiles
Pros: potentially effective; if shared among organisations, they promote cross organisation interoperability;
Cons: challenging to agree between different organisations; often end-up being complex combinations of features reflecting the interests of many disparate parties; they infringe autonomy of the entities adopting them;
Metadata Evaluation Approaches
Pros: helpful to identify specific problems;
Cons: based on community specific criteria;
Semi-automatic Metadata Generation Approaches
Pros: helpful to deal with the data deluge;
Cons: human assessment;
Metadata Cleaning, Enhancement, Augmentation Approaches
Pros: fundamental to enable cross-community exploitation of metadata;
Cons: information loss; information inconsistency;

Table 2: Summary of Metadata Quality Approaches

yet it is impossible to identify generic and expressive enough agreements capable of accommodating the needs of every community of practice. Approaches aiming at highlighting the problems affecting metadata (cf. Sec. 4.2) are very useful for calling attention to potential problems, yet the list and semantics of the potential problems is usually community specific. Solutions focused on supporting the semi-automatic generation of metadata (cf. Sec. 4.3) are fundamental in contexts characterized by a large amount of resources, yet the need of human assessment of what has been generated it is an important limitation to the scalability of the approach. Approaches aiming at repairing and homogenising metadata (cf. Sec. 4.4) enable the usage of metadata in contexts different from their initial ones, yet may bring information loss and inconsistency.

Crowdsourcing-based approaches, which offer a problem-solving strategy that well apply to the data infrastructure settings, offer new opportunities as well as potentially introduce new challenges such as how to assess users and their contributions (Doan et al., 2011; Oomen and Aroyo, 2011). For instance, the so diffused social tagging (Huang et al., 2012) might have a very important impact on quality and effectiveness of resources metadata, especially if combined with semantic technologies. However, tags are often community specific and thus difficult to exploit in multidisciplinary contexts.

5. Conclusions

Data and metadata quality is a very important yet challenging issue affecting the effective usage of such a kind of resources playing a key role in our information society. Although no definitional agreement has been achieved yet, it is commonly recognized that metadata quality is a subjective, multi-dimensional

and context-dependent concept. All the issues characterizing it in a given community of practice are further amplified when dealing with "big data" scenarios where data and metadata (a) come from multiple and heterogeneous sources, (b) are collected with different approaches, and (c) are expected to be used in contexts different from their initial ones.

In this work, we have surveyed efforts done so far in modelling and assuring metadata quality in the Digital Library domain with the aim of providing a comprehensive and update picture of the progresses achieved so far in this area. In particular, we have discussed a number of attempts aiming at proposing frameworks characterizing the metadata quality issue and promoting effective methods for assessing the quality of given metadata. Moreover, a number of metadata quality issues arising in the creation and/or the aggregation phases have been discussed and the approaches aiming at mitigating them – e.g., guidelines, (semi-)automatic generation, validation, cleaning, improvement – have been presented. In spite of the fact that automatic approaches fit better with the given setting, best results in all stages of the metadata life cycle are obtained when automatic means are integrated with human intervention.

In spite of these results, there are still many open issues in dealing with data and metadata quality. The most relevant among these issues are: (i) the need to develop a comprehensive and open framework enabling diverse research communities to characterize their data quality concepts and tools by means of a lingua franca; (ii) the demand for the development a generic (i.e., non-domain-specific) and machine-processable way to capture data quality aspects that can be effectively used to acquire genuine indicators of quality-oriented aspects; (iii) the need to develop generic tools that, by relying on given characteristics of the data and known strategies, can augment quality-oriented aspects and certify the degree of the resulting data with respect to such aspects.

So far systems addressing more complex contexts, like those enabling data infrastructures, have developed solutions that largely resemble those just discussed for Digital Libraries. For instance, the OpenAIRE initiative (Manghi et al., 2012), which is called to develop a data infrastructure for open access research products, has developed guidelines that complement and reuse the DRIVER guidelines (Vanderfeesten et al., 2008) discussed in Section 4.1. Similarly, in the context of the biodiversity domain, Hardisty and Roberts (2013) envisage a data infrastructure serving this domain in which data and metadata quality issues and potential solutions are borrowed from the digital library domain. The US DataOne initiative (Allard, 2012), dedicated to provide an e-infrastructure for Earth observational data, has recently published tutorials on "How to write Quality Metadata" (Henkel et al., 2012) focusing primarily on "accuracy" and "completeness". Finally, in the framework of EUDAT (Lecarpentier, 2011), a large data infrastructure initiative that plans to act as an European aggregator of datasets metadata, solutions close to the one described above are proposed for the initial phase of infrastructure development, however it is recognized that much more has still to be done to deal with multi-disciplinary and multipurpose contexts (de Witt, 2012).

Given the increasing relevance of data-driven research and decision-making it

is expected that research on data and metadata quality will largely reinvigorate in the future. This expectation has motivated the authors in compiling this survey so to offer a reference point for all those that are not familiar with the large amount of work already done on metadata quality in the Digital Library area.

Acknowledgements

The work reported has been partially supported by the ICT-2007.4.3 Information and Communication Technologies Program as part of the DL.org project (Grant Agreement no. 231551). Special thanks go to Maria Bruna Baldacci for her valuable help and suggestions in finalising this paper.

References

Allard, S., 2012. DataONE: Facilitating eScience through Collaboration. Journal of eScience Librarianship 1.

Apps, A., 2005. Guidelines for encoding bibliographic citation information in dublin core metadata. DCMI Recoimmendation, Dublin Core Metadata Initiative, `http://dublincore.org/documents/dc-citation-guidelines/` Data accessed: March 2013.

Ashley, K., Bizer, C., Candela, L., Fergusson, D., Gionis, A., Heikkurinen, M., Laure, E., Lopez, D., Meghini, C., Pagano, P., Parsons, M., Viglas, S., Vitlacil, D., Weikum, G., 2012. Technological & Organisational Aspects of a Global Research Data Infrastructure - A view from experts. GRDI2020 Booklet, `http://www.grdi2020.eu/Pages/SelectedDocument.aspx?id_documento=9a85ca56-c548-47e4-8b0e-86c3534ad21d` Data accessed: March 2013.

Barton, J., Currier, S., Hey, J. M. N., 2003. Building quality assurance into metadata creation: An analysis based on the learning objects and e-prints communities of practice. In: Proceedings of the International Conference on Dublin Core and Metadata Applications - DC-2003, Seattle, 28 September 2 October 2003. Dublin Core Metadata Initiative, pp. 39–48.

Batini, C., Cappiello, C., Francalanci, C., Maurino, A., 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41 (3), 16:1–16:52.

Batini, C., Scannapieco, M., 2006. Data Quality: Concepts, methodologies and techniques. Springer-Verlag, Berlin-Heidelberg.

Beall, J., 2010. Metadata for name disambiguation and collocation. Future Internet 2 (1), 1–15.

Bethard, S., Wetzler, P. G., Butcher, K. R., Martin, J. H., Sumner, T., 2009. Automatically characterizing resource quality for educational digital libraries. In: Heath, F., Rice-Lively, M. L., Furuta, R. (Eds.), Proceedings of the 2009 Joint International Conference on Digital Libraries, JCDL 2009, Austin, TX, USA, June 15-19, 2009. ACM, pp. 221–230.

Borgman, C., 2010. Research data: Who will share what, with whom, when, and why?, China-North America Library Conference, Beijing.

Borgman, C., 2011. The Conundrum of Sharing Research Data. Journal of the American Society for Information Science and Technology 63 (6), 1059–1078.

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, D. W., Laurie, G., O'Neill, O., Rawlins, M., Thornton, D. J., Vallance, P., Walport, M., 2012. Science as an open enterprise. Final report, The Royal Society, `http://royalsociety.org/policy/projects/science-public-enterprise/report/` Date accessed: March 2013.

Bruce, T. R., Hillmann, D. I., 2004. The continuum of metadata quality: Defining, expressing, exploiting. In: Hillmann, D. I., Westbrooks, E. L. (Eds.), Metadata in Practice. ALA Editions, Ch. 15, pp. 238–256.

Candela, L., Athanasopoulos, G., Castelli, D., El Raheb, K., Innocenti, P., Ioannidis, Y., Katifori, A., Nika, A., Vullo, G., Ross, S., 2011a. The Digital Library Reference Model. Deliverable D3.2b, DL.org, `http://referencemodel.dlorg.eu/` Date accessed: March 2013.

Candela, L., Castelli, D., Pagano, P., 2011b. History, Evolution and Impact of Digital Libraries. In: Iglezakis, I., Synodinou, T.-E., Kapidakis, S. (Eds.), E-Publishing and Digital Libraries: Legal and Organizational Issues. IGI Global, Ch. 1, pp. 1–30.

Candela, L., Castelli, D., Thanos, C., 2010. Making digital library content interoperable. In: Agosti, M., Esposito, F., Thanos, C. (Eds.), Digital Libraries - 6th Italian Research Conference, IRCDL 2010, Padua, Italy, January 28-29, 2010. Revised Selected Papers. Vol. 91 of Communications in Computer and Information Science. Springer, pp. 13–25.

CDP Metadata Working Group, 2006. Dublin core metadata best practices. Tech. rep., Collaborative Digitization Program.

crossref.org, 2012. Metadata guidelines. `http://www.crossref.org/02publishers/metadata_guidelines.html` Data accessed: March 2013.

de Groat, G., 2009. Future directions in metadata remediation for metadata aggregators. Tech. rep., Digital Library Federation, `http://old.diglib.org/aquifer/dlf110.pdf` Data accessed: March 2013.

de Witt, S., October 2012. Metadata and EUDAT. EUDAT Presentation, retrieved March 2013 from `http://eudat.eu/system/files/Metadata%20and%20EUDAT.pdf`.

Doan, A., Ramakrishnan, R., Halevy, A. Y., Apr. 2011. Crowdsourcing systems on the World-Wide Web. Commun. ACM 54 (4), 86–96.

Dushay, N., Hillmann, D. I., 2003. Analyzing metadata for effective use and re-use. In: Proceedings of the International Conference on Dublin Core and Metadata Applications - DC-2003, Seattle, 28 September - 2 October 2003. pp. 161–170.

Eppler, M. J., 2006. Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes. Springer.

Foulonneau, M., Riley, J., 2008. Metadata for Digital Resources: Implementation, Systems Design and Interoperability. Chandos Publishing (Oxford).

Gonçalves, M. A., Fox, E. A., Watson, L. T., Kipp, N. A., 2004. Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. ACM Transactions on Information Systems (TOIS) 22 (2), 270–312.

Gonçalves, M. A., Moreira, B. L., Fox, E. A., Watson, L. T., 2007. "what is a good digital library?" - a quality model for digital libraries. Information Processing and Management 43, 1416–1437.

Graham, P. S., September 1990. Quality in cataloging: Making distinctions. Journal of Academic Librarianship 16 (4), 213–218.

Greenberg, J., 2003. Metadata Generation: Processes, People and Tools. Bulletin of the American Society for Information Science and Technology 29 (2), 16–19.

Greenberg, J., 2004. Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. Journal of Internet Cataloging 6 (4), 59–82.

Greenberg, J., Pattuelli, M. C., Parsia, B., Robertson, W. D., 2001. Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. Journal of Digital Information 2 (2), 1–10.

Greenberg, J., Spurgin, K. M., Crystal, A., 2006. Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions. International Journal of Metadata, Semantics and Ontologies 1 (1), 3–20.

Guy, M., Powell, A., Day, M., 2004. Improving the quality of metadata in eprint archives. Ariadne Issue 38.

Hanson, B., Sugden, A., Alberts, B., February 2011. Making data maximally available. Science 331 (6018), 649.

Hardisty, A., Roberts, D., 2013. A decadal view of biodiversity informatics: challenges and priorities. BMC Ecology -, to appear. http://vbrant.eu/sites/vbrant.eu/files/1648427867838466_accepted_article.pdf Date accessed: March 2013.

Heery, R., Patel, M., 2000. Application profiles: mixing and matching metadata schemas. Ariadne Issue 25.

Henkel, H., Hutchison, V., Strasser, C., Rebich Hespanha, S., Vanderbilt, K., Wayne, L., 2012. How to Write Good Quality Metadata. DataONE Education Module, retrieved March 2013 from http://www.dataone.org/sites/all/documents/L08_WriteQualityMetadata.pptx.

Hey, T., Tansley, S., Tolle, K. (Eds.), 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, http://research.microsoft.com/en-us/collaboration/fourthparadigm/ Data accessed: March 2013.

Hillmann, D. I., 2008. Metadata quality: From evaluation to augmentation. Cataloging & Classification Quarterly 46 (1).

Hillmann, D. I., Dushay, N., Phipps, J., 2004. Improving metadata quality: Augmentation and recombination. In: Proceedings of the International Conference on Dublin Core and Metadata Applications - DC-2004, Shanghai, China, 11-14 October 2004. Dublin Core Metadata Initiative, pp. 1–8.

Hillmann, D. I., Phipps, J., 2007. Application profiles: Exposing and enforcing metadata quality. In: Proceedings of the International Conference on Dublin Core and Metadata Applications - DC-2007, Singapore, 27-31 August 2007. Dublin Core Metadata Initiative, pp. 53–62.

Huang, S.-L., Lin, S.-C., Chan, Y.-C., 2012. Investigating effectiveness and user acceptance of semantic social tagging for knowledge sharing. Information Processing and Management 48, 599–617.

Hughes, B., 2004. Metadata quality evaluation: Experience from the open language archives community. In: Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E. A., Lim, E.-P. (Eds.), Digital Libraries: International Collaboration and Cross-Fertilization, 7th International Conference on Asian Digital Libraries, ICADL 2004, Shanghai, China, December 13-17, 2004, Proceedings. Vol. 3334 of Lecture Notes in Computer Science. Springer, pp. 320–329.

Ioannidis, Y., August 2005. Digital libraries at a crossroads. International Journal on Digital Libraries 5 (4), 255–265.

Kan, M.-Y., Tan, Y. F., 2008. Record matching in digital library metadata. Communications of the ACM 51 (2), 91–94.

Laender, A. H. F., Gonçalves, M. A., Cota, R. G., Ferreira, A. A., Santos, R. L. T., Silva, A. J. C., 2008. Keeping a digital library clean: new solutions to old problems. In: da Graça Campos Pimentel, M., Bulterman, D. C. A., Soares, L. F. G. (Eds.), Proceedings of the 2008 ACM Symposium on Document Engineering, Sao Paulo, Brazil, September 16-19, 2008. ACM, pp. 257–262.

Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., Saylor, J., 2006. Metadata Aggregation and "Automated Digital Libraries": A Retrospective on the NSDL Experience. In: JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries. ACM Press, New York, NY, USA, pp. 230–239.

Lecarpentier, D., 2011. Towards a pan-European collaborative data infrastructure. iSGTW International Science Grid This Week Issue 9 November 2011.

Lee, D., On, B.-W., Kang, J., Park, S., 2005. Effective and scalable solutions for mixed and split citation problems in digital libraries. In: Berti-Equille, L., Batini, C., Srivastava, D. (Eds.), IQIS 2005, International Workshop on Information Quality in Information Systems, 17 June 2005, Baltimore, Maryland, USA (SIGMOD 2005 Workshop). ACM, pp. 69–76.

Madnick, S. E., Wang, R. Y., Lee, Y. W., Zhu, H., 2009. Overview and Framework for Data and Information Quality Research. Journal of Data and Information Quality 1 (1), 2:1–2:22.

Manghi, P., Bolikowski, L., Manola, N., Schirrwagen, J., Smith, T., 2012. OpenAIREplus: the European Scholarly Communication Data Infrastructure. D-Lib Magazine 18 (9/10).

Manghi, P., Manola, N., Horstmann, W., Peters, D., 2010a. An infrastructure for managing EC funded research outputs – the OpenAIRE project. The Grey Journal (TGJ) 6 (1), 31–40.

Manghi, P., Mikulicic, M., 2011. PACE: A general-purpose tool for authority control. In: García-Barriocanal, E., Cebeci, Z., Okur, M., Öztürk, A. (Eds.), Metadata and Semantic Research: 5th International Conference, MTSR 2011. Vol. 240. pp. 80–92.

Manghi, P., Mikulicic, M., Candela, L., Castelli, D., Pagano, P., March/April 2010b. Realizing and Maintaining Aggregative Digital Library Systems: D-NET Software Toolkit and OAIster System. D-Lib Magazine 16 (3/4).

Moen, W. E., Stewart, E. L., McClure, C. R., 1997. The role of content analysis in evaluating metadata for the u.s. government information locator service (gils): Results from an exploratory study. `http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm` Data accessed: March 2013.

Moen, W. E., Stewart, E. L., McClure, C. R., 1998. Assessing metadata quality: Findings and methodological considerations from an evaluation of the u.s. government information locator service (gils). In: Proceedings of the IEEE Forum on Reasearch and Technology Advances in Digital Libraries, IEEE ADL '98, Santa Barbara, California, USA, April 22-24, 1998. IEEE Computer Society, pp. 246–255.

Moreira, B. L., Gonçalves, M. A., Laender, A. H., Fox, E. A., 2009. Automatic evaluation of digital libraries with 5SQual. Journal of Informetrics 3 (2), 102 – 123.

Nature, 2008. Community cleverness required. Nature 455 (7209), 1.

Nichols, D. M., Chan, C.-H., Bainbridge, D., McKay, D., Twidale, M. B., 2008. A lightweight metadata quality tool. In: Larsen, R. L., Paepcke, A., Borbinha, J. L., Naaman, M. (Eds.), ACM/IEEE Joint Conference on Digital Libraries, JCDL 2008, Pittsburgh, PA, USA, June 16-20, 2008. ACM, pp. 385–388.

Nichols, D. M., Paynter, G. W., Chan, C.-H., Bainbridge, D., McKay, D., Twidale, M. B., Blandford, A., 2009. Experiences in deploying metadata analysis tools for institutional repositories. Cataloging & Classification Quarterly 47 (3/4).

Ochoa, X., Duval, E., 2009. Automatic evaluation of metadata quality in digital repositories. International Journal on Digital Libraries 10 (2-3), 67–91.

Oomen, J., Aroyo, L., 2011. Crowdsourcing in the cultural heritage domain: opportunities and challenges. In: Proceedings of the 5th International Conference on Communities and Technologies. ACM, New York, NY, USA, pp. 138–149.

Park, J., 2009. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. Cataloging & Classification Quarterly 47 (3/4).

Park, J., Lu, C., 2009. Application of semi-automatic metadata generation in libreries: Types, tools, and techniques. Library & Information Science Research 31 (4), 225–231.

Park, J., Tosaka, Y., 2010a. Metadata Creation Practices in Digital Repositories and Collections: Schemata, Selection Criteria, and Interoperability. Information Technology and Libraries 29 (3), 104–116.

Park, J., Tosaka, Y., 2010b. Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. Cataloging & Classification Quarterly 48 (8), 696–715.

Phipps, J., Hillmann, D., Paynter, G., 2005. Orchestrating Metadata Enhancement Services: Introducing Lenny. In: Proceedings of the International Conference on Dublin Core and Metadata Applications - DC-2005, Madrid, 12-15 September 2005. Dublin Core Metadata Initiative, pp. 49–58.

Riley, J., Shepherd, K., 2009. A Brave New World: Archivists and Shareable Descriptive Metadata. American Archivists 72 (1), 91–112.

Savino, P., Schulze, F., 2011. Report on inclusion of archives repositories. Deliverable D2.4, European Film Gateway.

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., Cole, T. W., 2005. Is 'quality' metadata 'shareable' metadata? the implications of local metadata practices for federated collections. In: Thompson, H. (Ed.), Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, April 7-10 2005, Minneapolis, MN. pp. 223–237.

Shreeves, S. L., Riley, J., Milewicz, L., 2006. Moving Towards Shareable Metadata. First Monday 11 (8).

Smalheiser, N. R., Torvik, V. I., 2009. Author name disambiguation. Annual Review of Information Science and Technology (ARIST) 43 (1), 1–43.

Strong, D. M., Lee, Y. W., Wang, R. Y., 1997. Data quality in context. Communications of the ACM 40 (5), 103–110.

Stvilia, B., Gasser, L., Twidale, M. B., Shreeves, S. L., Cole, T. W., 2004. Metadata quality for federated collections. In: Chengalur-Smith, I. N., Raschid, L., Long, J., Seko, C. (Eds.), Ninth International Conference on Information Quality (IQ 2004), November 5-7, 2004. MIT, pp. 111–125.

Stvilia, B., Gasser, L., Twidale, M. B., Smith, L. C., 2007. A framework for information quality assessment. JASIST 58 (12), 1720–1733.

Thanos, C., 2012. Global research data infrastructures: The GRDI2020 vision. GRDI2020 Booklet, `http://www.trust-itservices.com/uploads/GRDI2020%20Roadmap/fc14b1f7-b8a3-41f8-9e1e-fd803d28ba76.pdf` Data accessed: March 2013.

Vanderfeesten, M., Summann, F., Slabbertje, M., 2008. DRIVER guidelines 2.0. Tech. rep., DRIVER, `http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf` Data accessed: March 2013.

Wang, R. Y., Strong, D. M., 1996. Beyond accuracy: what data quality means to data consumers. Journal of Management Infornation Systems 12 (4), 5–33.

Yasser, C., 2011. An Analysis of Problems in Metadata Records. Journal of Library Metadata 11 (2), 51–62.