# Meta-Line: Lineage Information for Improved Metadata Quality


Sascha Tönnies
L3S Research Center
Appelstrasse 9a
30167 Hannover
Germany
toennies@l3s.de

Benjamin Köhncke
L3S Research Center
Appelstrasse 9a
30167 Hannover
Germany
koehncke@l3s.de

Wolf-Tilo Balke
IFIS TU Braunschweig
Mühlenpfordtstrasse 23
38106 Braunschweig
Germany
balke@ifis.cs.tu-bs.de



## ABSTRACT
Controlled content quality also in terms of indexing is one of the major advantages of using digital libraries in contrast to general Web sources or Web search engines. However, considering today's information flood the mostly manual effort in acquiring new sources and creating suitable (semantic) metadata for content indexing and retrieval is already prohibitive. A recent solution is given by automatic generation of metadata, where various methods currently become more widespread. But in this case neglecting quality assurance is even more problematic, because heuristic generation often fails and the resulting low-quality metadata will directly diminish the quality of service that a digital library provides. To address this problem, we propose a metadata quality model to determine the overall quality of a metadata set and validate individual requirements imposed on that metadata set. Furthermore, lineage information is provided to trace the quality evolution of a metadata set.



## Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Storage and Retrieval – *Information Search and Retrieval.*

## General Terms
Design, Performance, Experimentation.

## Keywords
Metadata Quality, Quality Lineage, Model


## 1. INTRODUCTION
With the growth of the Internet as a publication platform and especially with the growing number of open access journals and 'grey literature'/preprint servers, more and more high quality content is made available constantly. Beside the content that is provided by domain specific publishers also this Web content is increasingly important for digital library users. Hence many digital libraries have extended their collections by harvesting content from the Web. However, unlike the controlled content that arrives from traditional publishers, assessing the quality of harvested content poses severe challenges. Whereas the general quality of each item or Web information source can usually be assessed quite well by the community of users (e.g. using feedback in Web 2.0 interfaces), the problem of correctly indexing the new content for retrieval with both bibliographic data and content-based index terms remains with the digital library provider. Whereas gathering information from the Web has often been compared to 'trying to drink from a fire hydrant', indexing all this information with controlled quality seems like 'trying to drink from a fire hydrant while assessing the water quality of each sip'. This leads to a trade-off for digital libraries between offering broad and up-to-date document collections and providing high quality metadata for retrieval.

We propose that information about the generation process of (not only semantic) metadata is essential and that in practice this kind of information is not available. To solve this problem, we have to trace the evolution of metadata records during their life cycle. Furthermore, this information must be globally available and machine readable to use it during automatic metadata enrichment and quality checks. This entails that we have to define a unique standard with the power to describe this kind of information. It seems obvious to reuse an established concept from the domain of data warehousing the so called data lineage [1] or data provenance [2]. Information stored in data warehouses is collected from many different sources and integrated as materialize views in local databases. One can clearly see the similarities to our scenario. Also digital libraries collect data from different sources and integrate them into their repository. With the usage of lineage information, it would be possible to trace the evolution of metadata records. But how could such lineage information look like? All the more if we consider the different requirements of the different domains, e.g. chemistry, computer science, architecture or mechanical engineering, do we have to confess that finding the "one size fits all" solution will not work?

Considering related work already done in the field of data provenance, we can see that this kind of information is always highly domain dependent. Thus, we can find literature about lineage models in specific domains, e.g. [3], or based on well-defined workflow engines, e.g. [4]. Emphasizing the main objective, i.e. the automatic quality assessment of metadata records, we may reduce this overwhelming problem of defining a global and domain spanning metadata lineage schema into a smaller one. We need a generic quality model which makes it possible to determine the quality of a specific metadata record in a specific version at any time.



*JCDL'12*, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

## 2. QUALITY MODEL

Data quality can be measured by several metrics. These metrics are objective or subjective and highly tasked dependent. This implies that a quality model has to integrate the concept of a metric as well as the concept of individual requirements. Our model (see Figure 1) is divided into three areas: *local*, *local*/global, and *global*. In this case, global means that the respective model parts are universally valid and therefore have to be globally available. The local part includes information that is highly content dependent and should therefore be stored within the respective organization, i.e. digital library. For the parts assigned to *global/local* such a strict allocation is not possible. Here, we have to decide case-by-case.

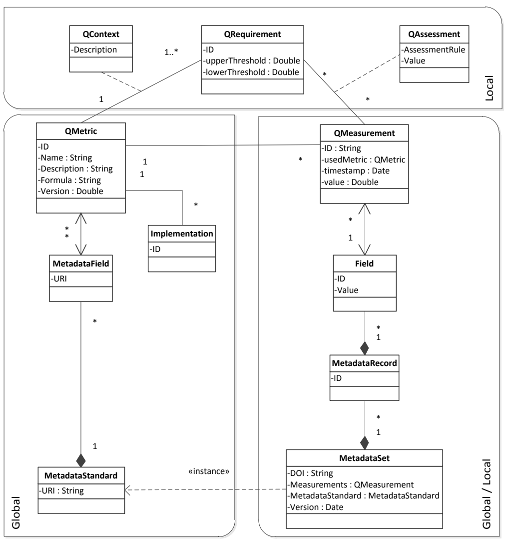


**Figure 1: Metadata Quality Model**

**Global** When talking about digital libraries and document collections, we have to consider different *metadata standards*. Taking also semantic metadata into account, the list of possible standards gets even longer. Every single one is uniquely identifiable by its URI, and defines several metadata fields. Each field has a unique identifier, i.e. its URI.

As already mentioned, the contents vary and so does the quality assessment from field to field. Therefore, it must be possible to assign different metrics to each individual *metadata field* and not only one "global" metric for a whole metadata standard. In our model this is achieved by a quality metric instance associated with a metadata field. Such a quality metric (*QMetric*) is defined as a quantifiable inspection criterion. Therefore, each QMetric has to have a name, a verbal description, the actual formula. Due to compatibility reasons, the outcome of each metric should be normalized to [0,1]. In addition, we can also store references to concrete implementations of the metric.

**Global/Local** In a next step, we need to measure the quality for individual instances of the metadata standard, i.e. metadata set. This set can differ depending on the version. Thus, it is essential for the model, that each set in a specific version can be identified. This can be achieved by assigning a digital object identifier (DOI®) to each set. Now, it is possible to perform a measurement for a specific metric in relation to a metadata set. The reference to the used metric is stored within the *QMeasurement* object. Each QMeasurement has been executed at a specific timestamp and the outcome is a value between [0,1]. Based on these concepts, we are already able to specify the overall quality of a metadata set.

**Local** After we can globally quantify the quality of a specific metadata set, we have to judge this value in the context of a specific digital library. By means of *QRequirement* the quality that a QMetric should have for a metadata field is specified. Typically it defines a set of thresholds which will be used during the quality assessment. It is important that each QRequirement is only valid for a specific context which is verbally described. In this way we will be able to accommodate different users or decision making contexts that have different criteria for assessing the data quality levels that are appropriate for their particular task. After we defined quality requirements for a specific context and we conducted some measurements, we are ready to apply our assessment rules. A *QAssessment* is the result of applying such a rule. Utilizing this model, we are now able to specify and reference one or more metrics for a specific metadata field. In addition, each information provider can use this knowledge to actually perform measurements of owning metadata sets and judge the quality according to his requirements.

Finally, we can adopt the well-known ETL workflow from the area of data warehousing and combine it with a central quality repository. By using a centralized repository, we are able to reuse measurements, thus the quality assessment is faster. This repository can be integrated into a workflow, to evaluate the metadata sets and automatically determine their quality.

## 3. CONCLUSION

We discussed the concept of data lineage process and showed that it is unlikely to come up with a domain spanning lineage model. Still, our main objective, the automatic quality assessment, could be solved by introducing our fine-grained quality model. Our introduced quality model distinguishes between the general metadata standard and concrete instances of this standard. Therefore, we can define several quality metrics for each field of the standard and we can perform concrete measurements of metadata records based on these metrics. In addition, each library can define its own requirements and assess them. For our future work we plan to establish a community based portal for developing quality metrics. The functionality could be similar to the functionality of the *myExperiment* portal. That implies building domain dependent communities, finding, sharing, and creating standards for semantic metadata and their related metrics.

## 4. REFERENCES


[1] Y. Cui and J. Widom, "Practical lineage tracing in data warehouses," in *Proceedings of 16th International Conference on Data Engineering*, 2000, pp. 367-378.

[2] P. Buneman, A. Chapman, and J. Cheney, "Provenance management in curated databases," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data - SIGMOD '06*, New York, New York, USA: ACM Press, 2006, pp. 539-550.

[3] R. Bose and J. Frew, "Composing lineage metadata with XML for custom satellite-derived data products," in *Production*, 2004, vol. 16, pp. 275 - 284.

[4] P. Missier, K. B. Jjame, J. Zhao, and C. Goble, "Data lineage model for Taverna workflows with lightweight annotation requirements," in *IPAW*, 2008, vol. 5272/2008, pp. 17-30.