A Metadata Quality Assurance Framework

Péter Király

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
Göttingen, Germany
peter.kiraly@gwdg.de

September, 2015

Contents

I. Introduction and hypothesis

In lots of large digital collections the only way a user could access a digital object is via its metadata. If metadata is not precise, contains inappropriate or less information users miss the object, and data creators lost the energy they put in the creation and maintenance of it. In past years there were some efforts in different digital collections to run research projects regarding the metadata quality: define what it is, how to measure it, and suggesting methods to improve it. Europeana[1] – Europe's digital library, museum and archives – just published a report [1] about the current state of art of this topic within the institution. This report states: „There was not enough scope for this Task Force to investigate elements such as metrics for metadata quality or how EDM schema validation could affect metadata quality."[2]. This current research

[1] `http://europeana.eu`
[2] [1] p. 52. EDM stands for Europeana Data Model, Europeana's metadata schema, see [2]

would like to start at exactly this point; we would like to set up a framework, which tests incoming and existing records, checks them against quality requirements, gives the Europeana community a dashboard about metrics, showing the historical changes of those metrics, and provides tips and suggestions to data creators, aggregator institutions and different Europeana teams.

The starting hypothesis is that it is possible to measure some factors of the data quality with computational tools. Here we list just a few of those data quality features:

- the „completeness" of the records: the ratio of the fields filled and unfilled
- there are mandatory fields, and if they are empty, the quality goes down
- whether the value of an individual field matches the rules of the metadata scheme
  - there are fields which should match a known standard, for example ISO language codes - you can apply rules to decide whether the value fits or not
  - the „data provider" field is a free text - no formal rule - but no individual record could contain unique value, and when you import several thousands of new record, they should not contain more than a couple new values
  - there are fields which should contain URLs or emails or dates, we can check whether they fit for formal rules, and their content are in a reasonable range (we should not have record created in the future for example)
- the probability that a given field value could be unique (the less other records have the same value, the higher information value of a field instance)
- the probability that a record is not duplication of another record

The measurements gives information on three different areas: the data quality of a record, a metadata scheme field, and a collection of records (all the records of a data provider, or all the records in an ingestion session). The most of the measurements happen on record level. Here we examine a number of metrics, and create an overall view of the record. By comparing metrics of different records we can filter out "low quality" and exceptionally good records. Moreover we can find records with similar features. The field level means, that with aggregating the information collected on records level, plus analysing the completeness of individual instances we examine the real usage patterns of a field (to see how it fits with the intended usage), and the developers of the metadata scheme and applications can draw conclusions regarding to the rules they applied on the field (for example the can provide good examples and anti-patterns for data creators, or they can improve the user interfaces). On collection level there is no original measurements, here we simple aggregate the data from the record and field level and provide explanatory analyses of the local cataloguing/data transformation habits.

If we could extract the data quality features we can draw conclusions and use them in several tasks. Some of these benefits:

- Working together with data creators on improving individual records or changing the local metadata creation habits
- Working together with aggregators on refining the metadata transformation workflows, to call attention to exceptions and local habits
- Improving ingestion workflow, filtering out records in an early stage
- Refining metadata scheme and documentation
- Refining end-user services (web interfaces and API) to build more on reliable fields, prepare for typical exceptions

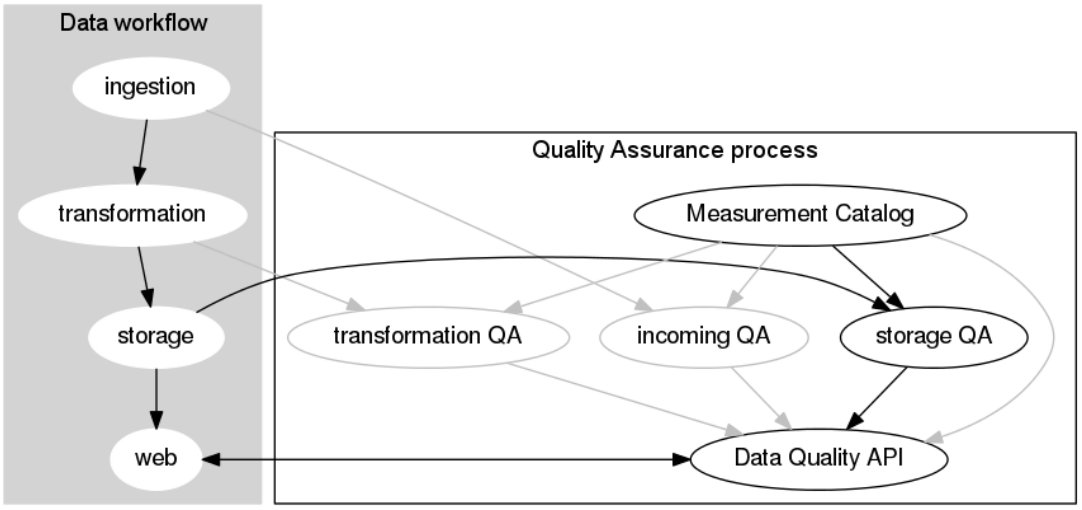

Figure 1: *The steps of data quality assurance.*

I would like to call to the reader's attention to a feature of any statistical tools, that they will not give us one hundred percent reliable results, but an overall tendency. Fortunately we can estimate the validity of our predictions with a level of accuracy and other metrics (in ideal case with manual comparison we even can measure the recall and precision of our predictions). Shortly: the results of the measurements usually are not intended to make decisions, but to help in the decision making process.

II. Workflow and architectural plan

There are three main points in current Europeana data management workflow where we can check data quality. The first stage is the entry point of incoming records. In this point the records are in some "raw" format, they might or might not be EDM schema record, so if we want to run checks, we have to prepare for all formats Europeana ingests. When Europeana ingest a record, it transforms it to EDM, and runs a couple of enhancement processes. That is the second main point where a quality assurance has take place. And finally the third check could happen after everything is stored in its final format. The first two stages require access points from the ingestion process, and I have no sufficient information the internal details of that to verify whether that is a realistic plan to implement those checks in the initial phase of the research, so in Figure 1 I draw them with grey denoting as optional steps. Another uncertainty is whether we can give back the measurements and metrics we get through the analyses, so we suggest to create a Data Quality API with its own storage system. This API can be called from the ingestion workflow, from the existing web interfaces, or can be used as standalone service.

The general metadata metrics along with other measurable metadata issues are stored in a formal way in Measurement Catalog. The entries of it contain operational codes (which the QA tools can use as plugins), description of the problem or measurable feature, the checking stage, the level of measurement (field, record, collection). When a QA process is running, it reads the runnable tasks from the Measurement Catalog. The process should provide back some information about the actual overhead of the task, so in planning the next run, the data quality operator (who is responsible for operate the tool) could plan the time and resource requirements of each tasks.

III. Quality metrics

The conceptual framework for the following metrics were set up by Bruce and Hillmann [9]. Ochoa and Duval [20] worked out the computations for the Ariadne project.[3] These metrics are the research will begin with, but there will be more metrics in the future.

I. Completeness

A metadata instance should describe the resource as fully as possible. Also, the metadata fields should be filled in for the majority of the resource population in order to make them useful for any kind of service. While this definition is most certainly based in static library instance view of metadata, it can be used to measure how much information is available about the resource.

$$Q_{comp} = \frac{\sum_{k=1}^{N} P(i)}{N} \tag{1}$$

Where

- P(i) is 1 if the i^{th} field has a no-null value, 0 otherwise.
- N is the number of fields defined in the metadata standard.

This approach doesn't make differences between the importance of the fields, however in practice some fields are more important than others. In the Europeana Data Model (EDM) [2] (and almost every metadata scheme) there are mandatory and optional elements. The web site and API use aggregated fields and facets. The Task Force report also talks about the differences between fields. So we have enough information to set weights for each fields. To calculate weighted value we have to apply the following equation:

$$Q_{wcomp} = \frac{\sum_{k=1}^{N} a_i \times P(i)}{\sum_{k=1}^{N} a_i} \tag{2}$$

Where

- a_i is the relative importance of the i^{th} field

II. Accuracy

The information provided about the resource in the metadata instance should be as correct as possible. Typographical errors, as well as factual errors, affect this quality dimension. However, estimating the correctness of a value is in not always a right/wrong choice. There are metadata fields that should receive a more subjective judgement. For example, while it is easy to determine whether the file size or format are correct or not, the correctness of the title, description or difficulty of an object has much more levels that are highly dependent of the perception of the reviewer. In Europeana's case this metrics is problematic, since we don't have the resource itself, only its metadata, so we are not able to extract term frequency of both.

$$Q_{accu} = \frac{\sum_{i=1}^{N} tf(resource_i) \times tf(metadata_i)}{\sqrt{\sum_{i=1}^{N} tf(resource_i)^2 \times \sum_{i=1}^{N} tf(metadata_i)^2}} \tag{3}$$

[3] The equations and bulk of the text of this section are copy from these two articles.

Where

- $tf(resource_i)$ and $tf(metadata_i)$, are the relative frequency of the i^{th} word in the text content of both the resource and the metadata.
- N is the total number of different words in both texts.

III. Conformance to Expectations

The degree to which metadata fulfils the requirements of a given community of users for a given task could be considered as a major dimension of the quality of a metadata instance. If the information stored in the metadata helps a community of practice to find, identify, select and obtain resources without a major shift in their workflow it could be considered to conform to the expectations of the community. According to the definition of quality ("fitness for purpose") used in this paper, this is one of the most important quality characteristics. We should measure this metrics differently for categorical values (where a field could take a value from a limited set), and free text values, which does not have this restriction.

Calculation for categorical fields:

$$infoContent(cat\_field) = -log(f(value)) \tag{4}$$

Where

- f(value) is the relative frequency of value in the categorical field for all the current instances in the repository. This relative frequency is equivalent to the probability of value.

normalized form:

$$infoContent(cat\_field) = 1 - \frac{log(times(value))}{log(n)} \tag{5}$$

Where

- times(value) is the number of times that the value is present in that categorical field in the whole repository.
- n is the total number of instances in the repository.
- When times(value) is 0 (the value is not present in the repository), the infoContent is 1. On the other hand, if times(value) is equal to n (all the instances have the same value), the infoContent is 0.

$$Q_{cinfo} = \frac{\sum_{i=1}^{N} infoContent(field_i)}{N} \tag{6}$$

Where

- N is the number of categorical fields.

Calculation for free text fields:

$$infoContent(freetext\_field) = \sum_{i=1}^{N} tf(word_i) \times log(\frac{1}{df(word_i)}) \tag{7}$$

Where

- $tf(word_i)$ is the term frequency of the i^{th} word

- $df(word_i)$ is the document frequency of the i^{th} word.
- N is the number of words in the field.

$$Q_{tinfo} = log(\sum_{i=1}^{N} infoContent(field_i)) \tag{8}$$

IV. Logical Consistency and Coherence

Metadata should be consistent with standard definitions and concepts used in the domain. The information contained in the metadata should also have internal coherence, that means that all the fields describe the same resource.

Consistency:

$$brokeRule_i = \{\, 0\ ; if\ instance\ complies\ with\ ith\ rule 1; otherwise \tag{9}$$

$$Q_{cons} = 1 - \frac{\sum_{i=1}^{N} brokeRule_i}{N} \tag{10}$$

Coherence:

$$distance(f1, f2) = \frac{\sum_{i=1}^{N} tfidf_{i,f1}\ timestfidf_{i,f2}}{\sqrt{\sum_{i=1}^{N} tfidf_{i,f1}^2 \times \sum_{i=1}^{N} tfidf_{i,f2}^2}} \tag{11}$$

Where

- $tfidf_{i,field}$ is the Term Frequency Inverse Document Frequency of the i^{th} word in the textual field f.
- N is the total number of different words in the field 1 and 2.

$$Q_{coh} = \frac{\sum_{i}^{N}\sum_{j}^{N} \{\, d\ istance(field_i, field_j); if\ i < j 0; otherwise}{\frac{N \times (N-1)}{2}} \tag{12}$$

V. Accessibility

Metadata that cannot be read or understood have no value. If the metadata are meant for automated processing, for example GPS location, the main problem is physical accessibility (incompatible formats or broken links). If the metadata are meant for human consumption, for example Description, the main problem is cognitive accessibility (metadata is too difficult to understand). These two different dimensions should be combined to estimate how easy is to access and understand the information present in the metadata.

$$Q_{link} = \frac{links(instance_k)}{max_{i=1}^{N}(links(instance_i))} \tag{13}$$

$$Q_{read} = \frac{\sum_{i}^{N} Flesch(fieldtext_i)}{100 \times N} \tag{14}$$

VI. Timeliness

Metadata should change whenever the described object changes (currency). Also, a complete metadata instance should be available by the time the object is inserted in the repository (lag). The lag description made by Bruce and Hillman, however, is focused in a static view of metadata. In a digital library approach, the metadata about a resource is always increasing which each new use of the resource. The lag, under this viewpoint, can be considered as the time that it takes for the metadata to describe the object well enough to find it using the search engine provided in the repository.

$$Q_{curr} = Q_{avg} = \frac{\sum_{i=1}^{N} \frac{Q_i - minQ_i}{maxQ_i - minQ_i}}{N} \tag{15}$$

Where

- Q_i is the value of the i^{th} quality metric (for example Qcomp, Qtinfo or Qread),
- $minQ_i$ and $maxQ_i$ are the minimum and maximum value of the i^{th} metric for all the instances in the repository.
- N is the total number of metrics considered in the calculation.
- Qavg is then the average of the different quality metrics for a given instance.

$$Q_{time} = \frac{Qcurr_{t2} - Qcurr_{t1}}{Qcurr_{t1} \times (t2 - t1)} \tag{16}$$

Where

- t_1 is the time when the original currency ($Qcurr_{t1}$) was measured
- t_2 is the current time with is corresponding value of instantaneous currency ($Qcurr_{t2}$).

prediction for t_3:

$$Qcurr_{t3} = (1 + Qtime_{t2-t1})^{(t3-t2)} \times Qcurr_{t2} \tag{17}$$

Where

- $Qtime_{t2-t1}$ is the calculation of the Qtime metric during the interval between t_1 and t_2.
- t_3 is the time to which the Qcurr estimation is desired.

VII. Provenance

The source of the metadata can be another factor to determine its quality. Knowledge about who created the instance, the level of expertise of the indexer, what methodologies were followed at indexing time and what transformations the metadata has passed through, could provide insight into the quality of the instance.

$$Qprov = Reputation(S) = \frac{\sum_{i=1}^{N} Qavg_i}{N} \tag{18}$$

Where

- $Qavg_i$ is the Average Quality of the i^{th} instance contributed by the source S.
- N is the total number of instances produced by S.
- The Qprov of an instance is equal to the reputation of its source.

IV. Tools

The following existing tools are at our disposal during the research:

- Europeana Search API – This API gives us possibilities to search for terms, field content from a document's perspective, and via the facets we can examine the nature of fields.
- Europeana Apache Solr index – The Search API is based on Apache Solr[4], but naturally it doesn't expose all the possibilities Solr provides. We might install extra Solr plugins which provide extra statistical metrics. The Solr has however a drawback, what it returns is heavily depends on the Solr schema settings. Unfortunately in Europeana the full database reindexing happens very rarely, however from time to time the schema has been changed – it has a consequence, that the current database is not exactly shows what a record contains, so some indexes reflect a historical state of a schema setting.
- Europeana Record API – The Search API provides a limited view of the the full record however some metrics require accessing the full record.
- Europeana MongoDB database – The Europeana records are stored in a MongoDB database[5]. The Record API doesn't provide us any aggregation functions, and since Solr index might not contain every aspect of a record, we might use the underlying Mongo database for its aggregation functionality.
- A statistical software tool – We can extract a number of statistical features of the record set from the above mentioned tools, they are not statistical tools, and at the time of writing we expect, that we will need a proper statistical software tool (for example R[6] or some fork of it, such as Julia[7] or renjin [8]).
- Graph database – The *accessibility* metrics should be calculated by calculation of implicit and explicit connections between records, which might involve an application of a graph database such as Neo4j[9]. The introduction of any new application to the tool stack involve a number of difficulties, so before decision making we should consider the equilibrium of the benefits and disadvantages.

The Framework should able to manage big data sets, and thus should be able to scale up, and run the analyses both in a single machine, and parallely in multiple nodes in a distributed hardware environment. My preliminary researches [10] were based on MapReduce programming model and the tools provided by Apache Hadoop software stack: Hadoop Distributed File System and Java implementation of MapReduce.

V. Control data sets

The promise of the statistical approach is that it independent from the actual metadata scheme, and it could be applied to any kind data collection based on a given metadata scheme. To verify this promise we plan to work concurrently with two different kind of control groups. The first one is a library catalog created in MARC. The natural choice would be the catalog of Niedersächsische Staats- und Universitätsbibliothek, Göttingen (Germany)[11]. The second

[4] http://lucene.apache.org/solr/
[5] https://www.mongodb.org/
[6] https://www.r-project.org/
[7] http://julialang.org/
[8] http://www.renjin.org/
[9] http://neo4j.com/
[10] Codebase on GitHub: https://github.com/pkiraly/europeana-qa, first results: http://pkiraly.github.io/2015/09/23/number-one/.
[11] http://www.sub.uni-goettingen.de/

one would be several different scientific datasets created at the Georg-August-Universität Göttingen[12], and intended to store for long term preservation. For long term preservation of scientific dataset the metadata quality is a crucial factor, since metadata is usually the one and only access points to the records, and bad metadata quality makes big (and expensive) datasets inaccessible/uninterpretable.

VI. Related works

Within W3C the Data on the Web Best Practices Working Group is preparing a Data Quality Vocabulary [23], which describes how to publish data quality measurements in a Linked Data context. The Framework should be able to report in a way which is conformant to this future standard.

Within Digital Public Library of America (DPLA) there is a similar project focusing on the data quality metrics of that project [24].

Within the CARARE project (Connecting Archeology and Architecture for Europeana) Gavrilis and his colleagues created and implemented a metadata quality evaluation model, which proposed a new weighting model to measure metrics from viewpoints of different usage scenarios [28].

VII. Project tasks

The project has three – parallely running – phases, which strongly affect each others.

1. Planning and study phase (3-6 months)
   (a) Studying similar international projects (such as metadata quality and information quality studies/conferences, DPLA project)
   (b) Defining what to measure, how to measure
2. Engineering phase (2 years)
   (a) Create the general Quality Assurance framework concentrating on the core tasks.
   (b) Creating the Measurement Catalog, translating the initial set of metrics into software codes.
   (c) Run the code in a smaller set of records.
   (d) Evaluating the results of the measurements
   (e) See whether we should introduce other kind of metrics.
   (f) Improve the tool to make it scalable and adapt to workflow of the host institutions (Europeana, Niedersächsische Staats- und Universitätsbibliothek, Göttingen eResearch Alliance, GWDG).
3. Dissemination phase (6 months)
   (a) Suggesting changes in the data creation, ingestion, and user interface development tasks at the host institutions.
   (b) Writing articles and maybe a Ph.D. dissertation – given if it has enough material in the project.
   (c) Propagating the method, and planning of its application for other data sets

As a result of email discussions between the Research and Development and the Technology teams of Europeana and myself, they created a wiki and ticketing space [13] within Europeana's project management system for further discussions about the topic.

[12] http://www.uni-goettingen.de/
[13] https://europeanadev.assembla.com/spaces/europeana-r-d/wiki/Task_Group_on_data_quality_

References

[1] Dangerfield, Marie-Claire et. al (2015): Report and Recommendations from the Task Force on Metadata Quality http://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf

[2] Europeana Data Model Documentation (2015) http://pro.europeana.eu/page/edm-documentation

[3] [CODE4LIB] How to measure quality of a record https://www.mail-archive.com/code4lib@listserv.nd.edu/msg27628.html

[4] Phillips, Mark (2015a): Metadata Quality, Completeness, and Minimally Viable Records (2015-01-05) http://vphill.com/journal/post/4075/

[5] Phillips, Mark (2015b): DPLA Metadata Analysis: Part 1-4 http://vphill.com/journal/post/5553/

[6] Twitter #metadataquality hashtag: https://twitter.com/hashtag/metadataquality

[7] Harris, Jim (2011): The Metadata Crisis (2011-10-27) http://www.ocdqblog.com/home/the-metadata-crisis.html

[8] Bruce, Thomas (2013): Metadata Quality in a Linked Data Context (2013-01-24) https://blog.law.cornell.edu/voxpop/2013/01/24/metadata-quality-in-a-linked-data-context/

[9] Bruce, Thomas R. – Hillmann, Diane (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In Metadata in Practice, Hillmann and Westbrooks, eds. http://www.ecommons.cornell.edu/handle/1813/7895

[10] DCMI Metadata Provenance Task Group http://dublincore.org/groups/provenance/

[11] Dodds, Leigh (2010): Quality Indicators for Linked Data Datasets http://answers.semanticweb.com/questions/1072/quality-indicators-for-linked-data-datasets

[12] Flemming, Annika (2010): Quality Criteria for Linked Data Sources http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_Criteria_for_Linked_Data_sources&action=history

[13] Fürber, Christian – Hepp, Martin (2011): Towards a Vocabulary for Data Quality Management in Semantic Web Architectures. In Presentation at the First International Workshop on Linked Web Data Management, Uppsala, Sweden. http://www.slideshare.net/cfuerber/towards-a-vocabulary-for-data-quality-management-in-semantic-web-architectures

[14] W3C, Provenance Vocabulary Mappings http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings

[15] Stvilia, B. – Gasser, L. – Twidale, M. B. – Smith, L. C. (2007): A framework for information quality assessment. Journal of the American Society for Information Science and Technology, 58: 1720–1733. doi:10.1002/asi.20652 http://myweb.fsu.edu/bstvilia/papers/stvilia_IQFramework_p.pdf

[16] Lee, Sang Hyun – Haider, Abrar (2011): A Framework for Information Quality Assessment Using Six Sigma Approach. In Communications of the IBIMA Vol. 2011 (2011), Article ID 927907, DOI: 10.5171/2011.927907 http://www.ibimapublishing.com/journals/CIBIMA/2011/927907/927907.pdf

[17] Hill, Gregory (2009): A Framework for Valuing the Quality of Customer Information (PhD dissertation at Department of Information Systems, Faculty of Science, The University of Melbourne) http://ghill.customer.netspace.net.au/docs/summary.html

[18] Stvilia, Besiki (2006): Measuring information quality (PhD dissertation at University of Illinois at Urbana-Champaign) http://search.proquest.com/docview/305328745

[19] Besiki Stvilia's homepage http://myweb.fsu.edu/bstvilia/

[20] Ochoa, Xavier – Duval, Erik (2009): Automatic Evaluation of Metadata Quality in Digital Repositories. International Journal on Digital Libraries 10 (2-3), 67-91 https://lirias.kuleuven.be/bitstream/123456789/255807/2/xavuxavier-pre.pdf

[21] Erik Duval's homepage https://erikduval.wordpress.com/about/

[22] New Zealand national library local aggregation service metadata quality reports such as http://metadata.digitalnz.org/nzresearch/127

[23] Data on the Web Best Practices: Data Quality Vocabulary. (W3C Editor's Draft 29 September 2015) http://w3c.github.io/dwbp/vocab-dqg.html

[24] Harper, Corey (2015a): Can Metadata Be Quantified? Presentation at DPLAFest (2015-04-18) http://schd.ws/hosted_files/dplafest2015/c1/CanMetadataBeQuantifiedSlides.pdf

[25] Harper, Corey (2015b): dpla-analytics on GitHub https://github.com/chrpr/dpla-analytics

[26] Metadata Quality Research Brainstorming. (A shared document created by DPLA and Europeana staff members to discuss issues of cooperations in metadata quality research) https://docs.google.com/document/d/15pmA276_fxShkCEagoloJwCXH89PhrF3qWBgB8xSrag/edit#heading=h.mq4e51njyj6

[27] Debattista, Jeremy – Lange, Christoph – Auer, Sören (2014): Representing Dataset Quality Metadata using Multi-Dimensional Views. In Proceedings of the 10th International Conference on Semantic Systems (SEM '14). Leipzig, 2014, pp. 92-99 http://dx.doi.org/10.1145/2660517.2660525, http://eis-bonn.github.io/Luzzu/papers/semantics2014.pdf

[28] Gavrilis, Dimitris – Makri, Dimitra-Nefeli – Papachristopoulos, Leonidas – Angelis, Stavros – Kravvaritis, Konstantinos – Papatheodorou, Christos – Constantopoulos, Panos (2015): Measuring Quality in Metadata Repositories. In Research and Advanced Technology for Digital Libraries. Proceedings of 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015 Poznań, Poland, September 14–18, 2015. (Volume 9316 of the series Lecture Notes in Computer Science). pp 56-67.