Metadata quality evaluation of a repository based on a sample technique.

Marc Goovaerts & Dirk Leinders

Hasselt University Library, Belgium
{marc.goovaerts, dirk.leinders}@uhasselt.be

Abstract. In this paper, we evaluate the quality of the metadata of an OAI-compliant repository based on the completeness metric proposed by X. Ochoa and E. Duval. This study focuses on the completeness of the metadata records as defined by M.A. Sicilia et al, where machine-understandability is a mandatory requirement for completeness. The goal is to use the completeness metric as a tool for harvesters and repository managers to evaluate easily the quality of the metadata of a repository. We focus on the metadata used by the communities of agriculture, aquaculture and environment from the VOA3R project. The OceanDocs repository serves as a use case. The completeness metric is used on a sample of records from the repository. The paper concludes that in the opinion of the authors quality evaluation is not a global process, but depends on the context. The completeness metric have to be used on the specific elements, relevant for the specific community.

Keywords: Metadata quality, Institutional repository, VOA3R, OceanDocs

1 Introduction

The quality of metadata is crucial for service providers who want to develop enhanced services. VOA3R [1] is a 3-year European project launched in June 2010 and funded by the European Commission under the seventh framework ICT Policy Support Program. The VOA3R platform is a service provider who integrates existing open access repositories as well as digital libraries, sharing scientific and open access research related to Agriculture, Food, Aquaculture and Environment. VOA3R is dedicated to providing a community-oriented platform based on social networking, micro-blogging and social bookmarking. To support the quality of the harvested metadata a specific application profile was created: VOA3R AP [2]. But an application profile does not guarantee the quality of the content.

Metadata quality has not been given adequate attention in the repository community. The definition of Dublin Core as a standard for OAI-PMH [3] brought the granularity of the metadata for institutional repositories to a basic level. The Guidelines for Repository Implementers for OAI-PMH suggests that specific communities can use other standards, but in practice most of the repositories follow the standard implemen-

tation in packages like DSpace, which supports a Dublin Core qualified at maximum [4].

An important reason for this choice is the fact that most submitters are not information specialists. Authors do not like the administrative work of creating metadata. Therefore a minimal format seems a nice solution in an environment where authors get the responsibility to submit their papers.

The success of general search engines with simple text search made it feel superfluous to create rich metadata. In the last fifteen years, the internet has changed completely the way researchers are looking for information. Yahoo, Google and other search engines limited the search technique to a simple word search supported by a powerful ranking system. Why should people bother about rich metadata?

But rich services need better metadata. Service providers like VOA3R want to create relations between pieces of metadata automatically. Therefore you need more refined and precise metadata. Ontologies are again becoming relevant, surely if terms and concepts can be defined uniquely by an identifier or resource URI. For example, the systematic use of AGROVOC keywords makes it possible to relate research topics in AGRIS. The use of resource URIs for every AGROVOC keyword supports multilinguality.

Metadata formats guarantee the level of granularity. A full MODS, not one translated from Dublin Core qualified as in DSpace, is much more refined than Dublin Core. Specific application profiles have been developed like Agris AP [5] and VOA3R AP, as more granular formats than Dublin Core.

2 Definition of metadata quality

How can we define quality for metadata content? T.R. Bruce and D. Hillman [6] proposed seven parameters for metadata quality: completeness, accuracy, conformance to expectation, logical consistence, accessibility, timeliness, and provenance.

In this study we focus on completeness, as the most important parameter for the service provider, with the following definition: ‘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’. [7] But completeness is related to granularity and precision.

Objects are described by metadata elements. Granularity defines the refinement of these elements. For example, Dublin Core has only one element for the source element (bibliographicCitation), while MODS has the possibility to split up the source description in multiple elements. The example below shows part of a MODS description, where journal title, volume, start and end page are available separately.

```
<relatedItem type="host">
  <titleInfo>
    <title>Bulletin Scientifique de l'IMROP</title>
  </titleInfo>
  <part>
    <detail type="volume">
          <number>28</number>
    </detail>
    <extent  unit="page">
          <start>1</start>
          <end>31</end>
    </extent>
  </part>
</relatedItem>
```

Fig. 1. Example of a source description in MODS, specifically the reference to a journal.

The use of authority control, ontologies and unique identifiers defines content unequivocally. Because of its unambiguity a DOI or a handle is sometimes more relevant than a whole abstract. The use of resource URIs for author names, journal titles or thesauri terms makes these values uniquely defined. Institutional repositories are mainly based on text. Harvesters like OAIster collect structured text, but to create rich services a machine-readable approach is essential in a world with Linked Open Data. M.A. Sicilia et al. 'consider machine-understandability as a mandatory requirement for completeness of metadata records' [8].

Granularity and precision influences our view on the completeness of the metadata. These aspects will be used in the study further on.

Traditionally, metadata quality is evaluated manually based on a questionnaire. Basically, there are many subjective aspects in this approach. It is also a work intensive job. In their article 'Automatic evaluation of metadata quality in digital repositories' [9] X. Ochoa and E. Duval describe a very complete method of automatic evaluation of the seven metadata quality parameters of by T.R. Bruce and D. Hillman. The method does not only evaluate the metadata but also the relation to the content and the user expectations. In this article we focus only on the completeness of metadata. We used the second completeness metric of X. Ochoa and E. Duval.

3 Evaluation of metadata quality: OceanDocs case

We propose a simple statistical approach using a random sample of records to evaluate the metadata quality. The completeness metrics of X. Ochoa and E. Duval are devised to analyze digital libraries and repositories with a full record set. In many cases it can be more practical to work with a limited sample. For example, when a service provider requests a sample for evaluation before harvesting the targeted repository.

While X. Ochoa and E. Duval measure the metadata quality using complete records, we focused on key elements which are machine-readable. Some elements of

the metadata are difficult to evaluate because they are not always available or because they are not mandatory. For example some publications do not have an author, some journals do not have an ISSN. Every community has specific key elements, depending on their focus. The aquatic community uses for example ASFA keywords, while in agriculture AGROVOC is used. Evaluation criteria have to be adapted to the needs of the community.

The OceanDocs [10] repository, our study case, is used in the aquatic community and also harvested by VOA3R. Therefore, the ASFA and AGROVOC thesauri are relevant metadata elements for respectively the aquatic and the agricultural community.

Ex. Agrovoc term:

- Marine fisheries:
  - with term code: c_4611
  - with resource URI: http://aims.fao.org/aos/agrovoc/c_4611

ASFA term:

- Fisheries biology
  - with term code: c_5839
  - resource URI: http://aims.fao.org/aos/asfa/c_5839

Fig. 2. Example of AGROVOC and ASFA terms with their unique identifiers.

The key elements analyzed were the keywords and the source description. We evaluated the precision (availability of controlled vocabulary and unique identifiers for keywords) and the granularity (source description).

The analysis went through three steps:

1. Two random samples of OceanDocs records were taken with data about keywords and source description.
2. The confidence interval was measured for each of the keyword elements and for each of the elements of the source description to check whether the sample was representative [11].
3. For both group of elements, the results were computed with Ochoa and Duval's completeness metric [12].

3.1 Creation of samples of metadata records from OceanDocs

Samples were taken from the OceanDocs repository[1]. We generated the samples by applying a simple random sampling technique. Each record was included in the sample with equal probability, which was determined by the desired sample size. The first sample of 100 records gave a large confidence interval. The second sample of 300

[1] The data is available in the OAI-MODS format. (ex. http://www.oceandocs.org/odin-oai/request?verb=GetRecord&metadataPrefix=mods&identifier=oai:www.oceandocs.org:1834/1500).

records gave an acceptable confidence interval. Note that the size of the sample is not related to the size of the database, but to the standard deviation of the sample. Therefore, even for larger databases the technique allows to work with relatively small samples.

From each record we collected the following elements:

- Keywords: Free keywords, ASFA term, ASFA term code, AGROVOC term, AGROVOC URI
- Source of journal contribution: journal title, volume, issue, start page, end page. Only for sample 2, we calculated the availability.

If an element is available in a record the value is 1, if unavailable 0. Table 1 shows the results.

	Free keyword	ASFA keyword	ASFA termcode	AGROVOC keyword	AGROVOC URI
Sample 1 – 100 records					
Records with	51	71	64	51	51
Average ($\bar{x}$)	0,51	0,71	0,64	0,51	0,51
Sample 2 – 300 records					
Records with	168	211	196	161	161
Average ($\bar{x}$)	0,56	0,703	0,653	0,537	0,537

Table 1. Availability of elements for the different samples

3.2 Defining the confidence interval of the samples

For every keyword element of the samples, we defined the confidence interval using the formula proposed by L. Egghe and R. Rousseau. We calculated for the average $\bar{x}$ of sample size N a confidence interval with 95% certitude.

$$\left[\bar{x}-1.96\sqrt{\frac{\bar{x}\left(1-\bar{x}\right)}{N-1}},\bar{x}+1.96\sqrt{\frac{\bar{x}\left(1-\bar{x}\right)}{N-1}}\right] \quad \textit{(95\% confidence interval)} \qquad (1)$$

From L. Egghe & R. Rousseau (2001). Elementary Statistics for Effective Library and Information Service Management. London, Aslib. p. 86

The results are listed in table 2 and 3.

Sample 1 (100 records)	Average ($\bar{x}$)	Confidence interval
Free Keywords	0,51	[0,412 ; 0,608]
ASFA keyword	0,71	[0,621 ; 0,799]
ASFA code term	0,64	[0,545 ; 0,735]
AGROVOC	0,51	[0,412 ; 0,608]
AGROVOC URI	0,51	[0,412 ; 0,608]

Table 2. Confidence interval of sample 1

Sample 2 (300 records)	Average ($\bar{x}$)	Confidence interval
Free Keywords	0,56	[0,504 ; 0,616]
ASFA keyword	0,703	[0,651 ; 0,755]
ASFA code term	0,653	[0,599 ; 0,707]
AGROVOC	0,537	[0,480 ; 0,594]
AGROVOC URI	0,537	[0,480 ; 0,594]

Table 3. Confidence interval of sample 2

Sample 1 (with sample size N=100) gave a confidence interval of about 20%. Therefore we took a second sample with 300 records which had an acceptable confidence interval for our further analysis.

3.3 Evaluation of the metadata quality by using a completeness metric

X. Ochoa and E. Duval have defined two completeness metrics. The basic completeness metric counts the number of fields in each metadata instance that contain a no-null value. In the case of multi-valued fields, the field is considered complete if at least one instance exists. They also proposes a metric with a weighting factor for the different metadata fields of the record. A higher degree of relevance of a field will be translated in a higher weighting factor. We used this weighted completeness metric.

$$Qwcomp = \frac{\sum_{i=1}^{N} \alpha_i * P(i)}{\sum_{i=1}^{N} \alpha_i} \qquad (2)$$

From X. Ochoa & E. Duval (2009), Automatic evaluation of metadata quality in digital repositories. In Int. J. Digit. Libr., 10, (2-3), p. 71.

Note that the maximum value for Qwcomp is 1 (all fields with importance different from 0 are non-empty) and the minimum value is 0 (all fields with importance different from 0 are empty).

As discussed above, while X. Ochoa and E. Duval measure the metadata quality using complete records, we focused on key elements of the metadata.

We evaluated the metadata quality of OceanDocs, specifically the use of keywords. We only looked at the averages of sample 2 because of their smaller confidence inter-

val. For the aquatic community the use of the ASFA thesaurus is relevant. The agriculture community uses the AGROVOC thesaurus. Therefore we gave different weighting factors to each keyword element. Free keywords received the lowest and unique identifiers the highest weighting factor.

The quality of keywords elements for the aquatic community was measured by the use of free keywords, ASFA keywords and ASFA term codes. We gave them the following weighting factors for their relevance.

- Free keywords = 1 – ASFA keywords = 2 – ASFA term code = 3
  The completeness value from aquatic perspective:
  Qwcomp= (1*0,56 + 2*0,703+3*0,653)/(1+2+3) = 0,654

The quality of keywords elements for the agriculture community was measured by the use of free keywords, AGROVOC keywords and AGROVOC URIs with the following weighting factors.

- Free keywords = 1 - AGROVOC keywords = 2 - AGROVOC URI = 3
  The completeness value from agriculture perspective:
  Qwcomp= (1*0,56 + 2*0,537+3*0,537)/(1+2+3) = 0,540

In both cases we put a heavy weighting on the unique IDs. We believe that accuracy can be achieved mostly by using authority control and resource URIs are the most relevant exponent of it.

Based on the second sample, we also evaluated the completeness of the source description, specifically of journal contributions. From the 300 records in sample 2, 162 were journal contributions. We evaluated the source description on the existence of journal title, volume + issue, start page and end page. Volume and issue were combined - if one of both was available then it got a value - because some journals use only one of both.

The results are shown below in table 4.

Sample 2	Journal title	Volume-issue	Start page	End page
Records with	146	143	142	142
Average ($\bar{x}$)	0,901	0,883	0,877	0,877

Table 4. Availability of elements for publications in journals from sample 2

The following weighting factors were used, for:

- Journal title = 3 - volume + issue = 2 - start page = 2 - end page=1
  The completeness value for source (journal contribution):
  Qwcomp=(3*0,901 + 2*0,883 + 2*0,877 + 1*0,877)/(3+2+2+1) = 0,888.

We have evaluated the metadata of the OceanDocs repository on the quality of the keywords and the source description, through a sample of 300 records. We obtained the following completeness values.

ASFA	0,654
AGROVOC	0,540
Source (Journal contribution)	0,888

Table 5. Completeness values of ASFA, AGROVOC and Journal contributions

The level of metadata completeness for AGROVOC was low in OceanDocs. It is an oceanographic repository, therefore we expected a higher completeness level for ASFA. In our opinion, the completeness level for ASFA is still low. What level can a service provider expect to create services with these elements ? With a result of 0,654, about 35% of the records was not accessible through the keyword elements. On the other hand the completeness level of the source description was high. It demonstrates the granularity of the OceanDocs metadata.

Other aspects of the metadata could be studied like the description of relations (DOI, URLs, versioning, …). But the two parameters, keywords and source - ASFA and AGROVOC are similar parameters from different communities - are basic indicators of the completeness and the quality of metadata.

4 Conclusions

This contribution presents a quick and easy evaluation method of the metadata quality of institutional repositories. It evaluates the completeness and granularity of the content using a sample of records. From these records, machine-readable elements were selected to be evaluated in their context. With the completeness metric of X. Ochoa and E. Duval the quality was measured. The OceanDocs repository was used as a case study.

If harvesters want to create extra services on top of the basic search functionalities, they have to control the quality and specifically the completeness of (specific parts of) the metadata. From our test case, we see that different communities, in our case agriculture and oceanography, will have different focuses: e.g. AGROVOC against ASFA. The quality and its evaluation will depend on the standards of the community. In our opinion quality evaluation is not a global process, but depends on the context. The completeness metric will then be used on the specific fields, relevant for the specific community.

It is difficult to define the threshold values for metadata completeness based on one case study. Further studies will be necessary, but already it is clear that a high level of completeness is necessary to create rich services on the harvester level.

Metadata quality is relevant for the services that are required and can be delivered to a community by a harvester like VOA3R. Rich metadata is for us complete, granular and precise metadata. Central in this approach is the use of authority control systems with controlled vocabularies, ontologies and ultimately the use of resource URIs as unique identifiers which guarantees the accuracy and the reusability of the metadata.

References:

1. VOA3R (Virtual Open Access Agriculture & Aquaculture Repository) - http://voa3r.eu
2. N. Diamantopoulos, C. Sgouropoulou, K. Kastrantas and N. Manouselis (2011). Developing a Metadata Application Profile for Sharing Agricultural Scientific and Scholarly Research Resources. In Metadata and Semantic Research. (Communications in Computer and Information Science.,240), pp 453-466. (http://voa3r.confolio.org/scam/5/resource/15)
3. Guidelines for Repository Implementers: 2.1 Dublin Core and Other Metadata Formats - http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#MinimalImplementation-DC
4. DSpace, version 1.8.x. Functional Overview. Metadata - https://wiki.duraspace.org/display/DSDOC18/Functional+Overview#FunctionalOverview-Metadata
5. The AGRIS Application Profile for the International Information System on Agricultural Sciences and Technology Guidelines on Best Practices for Information Object Description - http://www.fao.org/docrep/008/ae909e/ae909e00.htm
6. T.R. Bruce & D. Hillmann (2004). The Continuüm of Metadata Quality: Defining, Expressing, Exploiting. In Metadata in practice. Chicago, ALA Editions. pp. 238-256.
7. X. Ochoa & E. Duval (2009). Automatic evaluation of metadata quality in digital repositories. In Int. J. Digit. Libr., 10, (2-3), p. 69
8. M.A. Sicilia, E. Garcia, C. Pages, J.J. Martinez, and J.M. Gutierrez. 2005. Complete metadata records in learning object repositories: some evidence and requirements. Int. J. Learn. Technol. 1, 4 (May 2005), 411-424.
9. X. Ochoa & E. Duval (2009). Automatic evaluation of metadata quality in digital repositories. In Int. J. Digit. Libr., 10, (2-3), p. 67-91, DOI: 10.1007/s00799-009-0054-4
10. OceanDocs - http://www.oceandocs.org
11. L. Egghe & R. Rousseau (2001). Elementary Statistics for Effective Library and Information Service Management. London, Aslib. p. 68-73.
12. Ochoa & E. Duval (2009), Automatic evaluation of metadata quality in digital repositories. In Int. J. Digit. Libr., 10, (2-3), p. 71