# Exploring Metadata Providers Reliability and Update Behavior

Sarantos Kapidakis(✉)

Laboratory on Digital Libraries and Electronic Publishing,
Department of Archive, Library and Museum Sciences, Ionian University,
72, Ioannou Theotoki Str., 49100 Corfu, Greece
sarantos@ionio.gr

**Abstract.** Metadata harvesting is used very often, to incorporate the resources of small providers to big collections. But how solid is this procedure? Are the metadata providers reliable? How often are the published metadata updated? Are the updates mostly for maintenance (corrections) or for improving the metadata? Such questions can be used to better predict the quality of the harvesting. The huge amount of harvested information and the many sources and metadata specialists involved makes prompt for answers by examining the actual metadata, rather than asking about opinions and practices. We examine such questions by processing appropriately collected information directly from the metadata providers. We harvested records from 2138 sources in 17 rounds over a 3-year period, and study them to explore the behaviour of the providers. We found that some providers are often not available. The number of metadata providers failing to respond is constantly increasing by the time. Additionally, the record length is slightly decreasing, indicating that the records are updated mostly for maintenance/corrections.

**Keywords:** OAI · Metadata · Harvesting · Reliability · Services · Record enrichment

## 1 Introduction

In this work we examine the harvesting of metadata and how it evolves over time. Metadata harvesting is used very often, to incorporate the resources of small or big providers to large collections. The metadata *harvesters,* like National Science Digital Library (NSDL) and Europeana, accumulate metadata from many *collections* (or *sources*), belonging to *metadata providers* mostly memory institutions, by automatically contacting their *servers* and storing the retrieved metadata locally. Their goal is to enable searching on the huge quantity of heterogeneous content, using only their locally store content. Metadata harvesting is very common nowadays and is based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). As examples, we mention the Directory of Open Access Repositories (OpenDOAR) that provides an


N. Fuhr et al. (Eds.): TPDL 2016, LNCS 9819, pp. 417–425, 2016.
DOI: 10.1007/978-3-319-43997-6_36

authoritative directory of academic open access repositories[1], and the OAIster[2] database of OCLC with millions of digital resources from thousands of providers.

In [6] Lagoze et al. discuss the NSDL development and explains why OAI-PMH based systems are not relatively easy to automate and administer with low people cost, as one would expect from the simplicity of the technology. It is interesting to investigate the deciciencies of the procedure.

National or large established institutions consistently try to offer their metadata and data reliably and current and to keep the quality of their services as high as possible, but local and smaller institutions often do not have the necessary resources for consistent quality services – sometimes not even for creating metadata, or for digitizing their objects. In small institutions, the reliability and quality issues are more prominent, and decisions often should also take the quality of the services under consideration.

The evaluation and quality of metadata is examined as one dimension of the digital library evaluation frameworks and systems in the related literature, like [2, 7, 10]. Fuhr et al. in [2] propose a quality framework for digital libraries that deal with quality parameters. The service reliability falls under their System quality component, and the metadata update under their Content quality component.

In [9] Ward describes how the Dublin Core is used by 100 Data Providers registered with the Open Archives Initiative and shows that is not used to its fullest extent. In [4] Kapidakis presents quality metrics and quality measurement tool, and applied them to compare the quality in Europeana and other collections, that are using the OAI-PMH protocol to aggregate metadata. In [5] Kapidakis further studies the responsiveness of the same OAI PMH servers, and the evolution of the metadata quality over 3 harvesting rounds between 2011 and 2013.

From the different aspects of the quality of digital library services, the quality of the metadata is the one that has been mostly studied. Some approaches are applied on OAI-PMH aggregated metadata: Yen Bui and Jung-Ran Park in [1] provide quality assessments for the National Science Digital Library metadata repository, studying the uneven distribution of the one million records and the number of occurrences of each Dublin Core element in these. Another approach to metadata quality evaluation is applied to the open language archives community (OLAC) in [3] by Hughes that is using many OLAC controlled vocabularies. Ochoa and Duval in [8] perform automatic evaluation of metadata quality in digital repositories for the ARIADNE project, using humans to review the quality metric for the metadata that was based on textual information content metric values.

In this paper we want to examine how solid is the metadata harvesting procedure. Are the metadata providers really reliable, responding when they are accessed, so that we have current information? How often are the published metadata updated? Are the updates mostly for maintaining (correcting) or for improving the metadata? The rest of the paper is organized as follows: In Sect. 2 we explore the reliability of the servers over many harvesting rounds. In Sect. 3 we study the frequency and the nature of the updates of the harvested metatata, and we conclude on Sect. 4.

[1] http://www.opendoar.org.

[2] http://www.oclc.org/oaister.en.html.

## 2 Reliability of the Servers Over Time

The reliability of the servers is important for ensuring current information. If the metadata harvesting service is not responding, the corresponding metadata records will not be updated at that time. This is not a very serious deficiency, as the updates will eventually be performed on the next successful metadata exchange, and will be normally used afterwards. Nevertheless, the unreliability - downtime of the metadata harvesting server usually indicate a proportional unreliability or downtime of the resource providing service, which always resides on the local sites, where both the local and the harvested metadata link to. When the resources are not available, the corresponding user requests are not satisfied, affecting the quality of the service.

In order to see if the metadata provider services are reliable and work over the years, we organized 17 *harvesting rounds*, over three years, from 2014 to 2016 (27 months). Each round took a few days to complete and was held apart of the previous round between one and two months. We tried to harvest the 2138 OAI sources listed in the official OAI Registered Data Providers[3] on January of 2014. We expected the providers listed there to be the most used and from the most seriously involved ones, and seriously considering their content and services.

In the first round, in January 2014, 1221 servers responded on all our rounds, only 1338 of the 2138 servers (63 %) responded at least one time. All servers were harvested on all rounds, but the remaining 800 servers never responded. Any OAI valid response was considered a satisfying server response, even in the rare cases that the communication with the server failed later on, because subsequent communication could eventually force the server to provide all its records, in most cases. It seems that the initial list, although official, was not up-to-date. Most failures were attributed to the OAI server not been found, to the OAI server not responding, or to protocol errors/incompatibilities. We performed the harvesting rounds by developing an application based on the pyoai[4] python library version 2.4.5. In order to have comparable results, we did not change the harvested servers in any of our rounds. On Fig. 1 we can

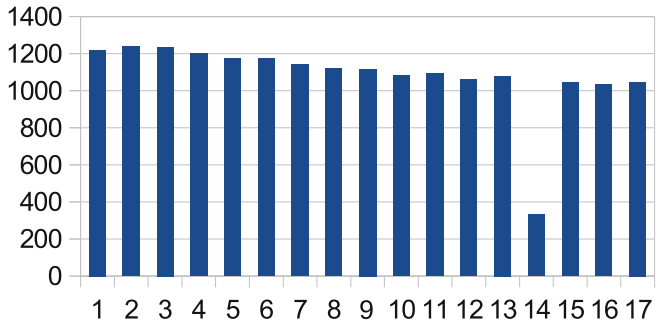


**Fig. 1.** The number of responses (y-axis) on each of the 17 harvesting rounds (x-axis).

[3] https://www.openarchives.org/Register/BrowseSites.

[4] https://pypi.python.org/pypi/pyoai.

see the number of servers responding on each round. We observe that temporary issues may affect the harvesting:

In one round (December 2015) there were much fewer responsive servers (330) than all other rounds. There must have been network connection problems closer to our harvesting client, which may not affect the harvesting from other clients. Therefore, we ignore this round in the rest of our study. The numbers of responses on all other rounds seem natural: In the first round (January 2014) 1221 servers responded, and on the last one (March 2016) 1047 servers. The most servers (1238) responded on the second round (February 2014) and the fewest (1033) on the one prior to the last (February 2016).

In Fig. 1 we observe that the number of servers responding on each round (excluding the ignored round) decreases almost linearly (and decreased by 174 between the 16 rounds), so we assume that many servers stopped working or responding permanently. In order to verify that, we depict in Fig. 2, for each round (except the last), the number of servers that responded for the last time during that round, and never responded afterwards.

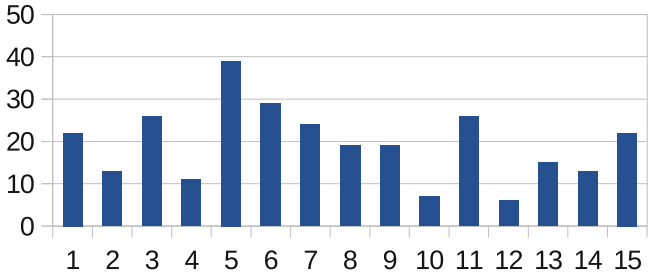


**Fig. 2.** The number of servers (y-axis) that responded for the last time in each round (but the last) (x-axis)

In the 16th round, 1047 servers responded. The servers that did not respond on the rounds close to the last one have higher probability of responding later on, and may not be dead. Nevertheless, on each round there were some (from 6 to 39, with an average of 20) servers that did respond for the last time during our 16 rounds. We counted the number of times each server responded and we clustered the servers by their count of responses, which we present on Fig. 3. The 800 servers that never responded are also included.

In Fig. 3 we can see that 721 servers, the vast majority of the responding servers, responded in all 16 rounds, 223 servers responded in 15 rounds and did not respond only once while 69 servers responded in 14 rounds and did not respond twice. The servers that responded only once were 27, and the number of servers responding from 1 to 13 times are similarly distributed, with 25 servers on average and a minimum of 14 servers (those responding 10 times) and a maximum of 39 servers (those responding 5 times).

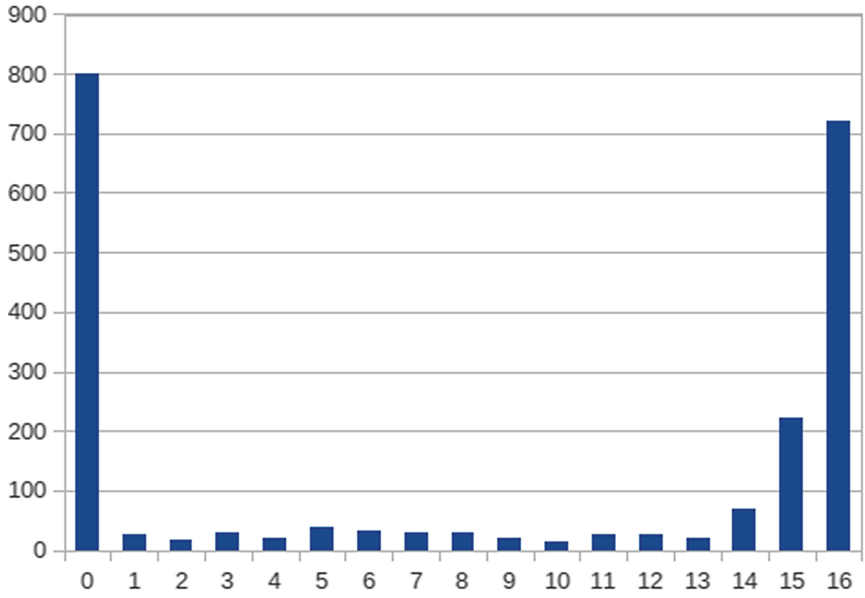


**Fig. 3.** The number of servers that responded exactly from 0 up to 16 rounds.

It is not expected that some servers fail accidentally many times, and this fact usually indicates a more permanent reason. In Fig. 3 we observe the accumulation of two quantities, the servers that responded few times and ceased working afterwards (and these should normally be about equal for each round and form a horizontal line on Fig. 3) and those that only failed occasionally a few times – and these are accumulated on the last few columns, which are sharply increasing, possibly forming an exponential line.

## 3 Frequency and Nature of the Updates

We want to get estimation on how often the published metadata are updated, which should affect the harvesting round, and also can indicate how important it is to reliably harvest successfully. For that we tried to harvest 1000 metadata records on each of the previously mentioned rounds. We examined the records of each collection and we recorded how many times they change during successive harvesting rounds, counting even the smallest change in any of their harvested records. This represents the frequency of update of the sample records, and is an indication of the update frequency of the whole collection. We depict that on Fig. 4, where the x-axis represents the number of rounds with value changes of the examined records and the y-axis represents the number of sources with that property.

We see that the largest group is the one with the records that never change. We counted 228 such sources that their content is therefore always the same. The remaining sources, which are the majority of them, occasionally update their content – some of them quite frequently. The records of the rest of the sources changed from 1 to 15 times,

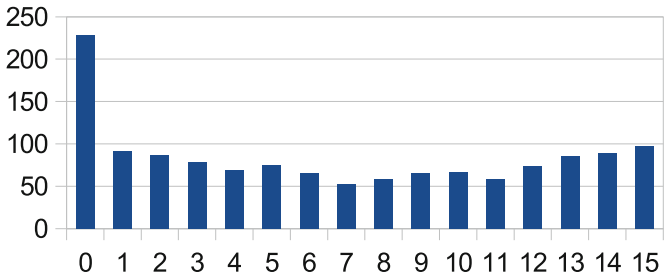


**Fig. 4.** The number of sources (y-axis) that their records changed from 0 up to 15 times (x-axis).

with less sources in the middle area (around 7 times, which forms a minimum of 52 sources) and more sources closer to the two edges (1 and 15 times which forms a maximum of 98 sources). The uniform distribution of the sources according to the number of the content updates (on average we have 74 sources for each number of source content changes, with standard deviation is 13.74) shows that the sources do not change too fast, faster than our sampling period, except maybe those with the highest number of changes counted (98 for 15 changes, 89 for 14 changes, etc.).

We conclude that, although there are sources that are never updated, the rate or quantity of the updates is ranging almost uniformly to any value and there are not any other obvious update patterns.

The harvested metadata records can always be mapped to Dublin Core. Thus, in addition to examining the whole record, we can examine individual Dublin Core elements. In Fig. 5 we can see in separate columns the number of sources that their Dublin Core elements title, creator, subject, description, type and rights changed, sorted by the number of times they change.

The columns in Figs. 4 and 5 are related, but not derived one from the other, as they count the records or the individual elements when they change. Nevertheless, they have a similar shape. In Fig. 5 we can see that most metadata elements do not change many times.

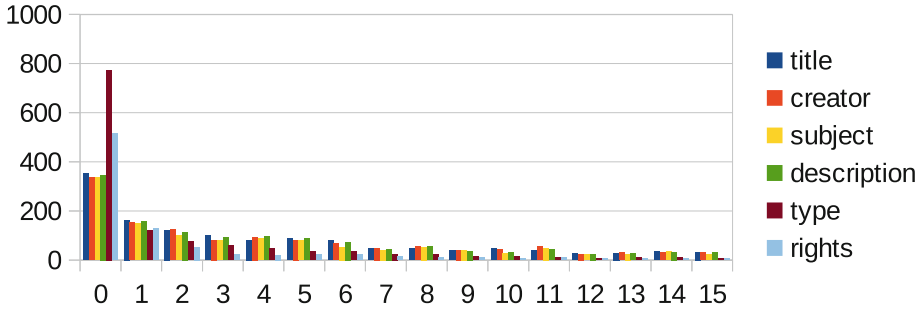


**Fig. 5.** The number of sources (y-axis) that some Dublin Core elements changed from 0 up to 15 times (x-axis) (Color figure online)

The resulting curve is decreasing, which means that the number of sources is decreasing by the number of times they change during the harvesting rounds. The sources that do not change at all are even more. The situation is similar for all examined Dublin Core elements, but more intense for the rights and type elements than with the other elements, the ones that are mostly used: title, creator, subject, description. This can be explained as the rights and type elements contain more standard information that does not get enriched and more rarely need improvement or correction than the more descriptive elements In general, the changes on the contents of the elements vary by the elements. Furthermore, the changes to the individual descriptive elements seem to have similar pattern and therefore to take place at the same round, so most changes are done record-wise rather than element-wise.

Finally, we examine if the size of the record, or the number or the content of its elements increases in an obvious (statistically significant) way, then we can assume that the metadata updates are actually enrichments, adding new information to the corresponding records. On the other hand, small, insignificant, changes in the size are most probably just corrections, that do not add any additional information. Therefore we examine the size of the records on their first and last harvesting round, to see if the updates are mostly for maintenance (corrections) or for improving the metadata.

Figure 6 shows the difference of the size of the record content, measured in words, between the last and first harvesting round for each source. In most cases, the changes/additions are zero or very small, and many times they have a negative sign: more words are removed than added. Very few sources have higher differences, and in most cases they are negative! The average increase of words per source is −14.6, but the standard deviation is 101.7, that indicates that we cannot derive a specific pattern.

We conclude that most changes are not record enrichments but rather maintenance changes, that happen to contain more removal of words. In most other cases the quality improvement was small, if any, indicating that it is hard to make extensive or consistent improvements on large metadata collections.

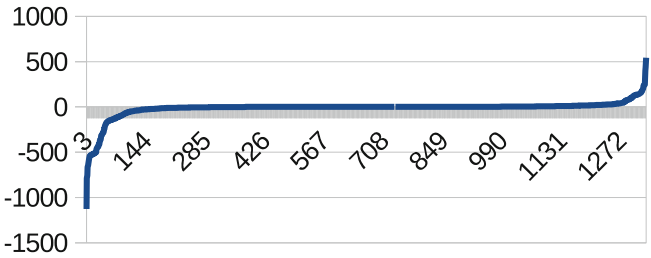


**Fig. 6.** The average increase of words (y-axis) contained in the records, for each source (x-axis)

This conclusion is also supported from Table 1, where we can see the minimum, maximum and average increase of words per collection. We can also see the standard deviation, which is always much higher than the corresponding average, and the number of sources involved in each calculation. We present in different rows the data for the whole record and those for the individual Dublin Core elements. The average

**Table 1.** Statistics on the increase of words in the source records and elements.

| | Min | Max | Mean | SDev. | Number |
|---|---|---|---|---|---|
| Record | −1128.3 | 542.1 | −14.6 | 101.7 | 1330 |
| Title | −21.1 | 15.4 | 0.0 | 1.5 | 1327 |
| Creator | −46.1 | 21.0 | −0.4 | 3.4 | 1296 |
| Subject | −31.9 | 44.4 | 0.1 | 3.4 | 1198 |
| Type | −6.2 | 4.2 | 0.0 | 0.6 | 1264 |
| Date | −6.5 | 5.2 | 0.0 | 0.4 | 1303 |
| Language | −0.8 | 2.9 | 0.0 | 0.1 | 1080 |
| Identifier | −27.4 | 32.0 | 0.1 | 2.1 | 1327 |
| Description | −718.9 | 540.1 | 0.1 | 41.6 | 1289 |
| Contributor | −39.6 | 24.5 | 0.1 | 2.7 | 760 |
| Format | −7.9 | 6.4 | 0.0 | 0.5 | 1044 |
| Publisher | −8.0 | 8.0 | 0.1 | 0.9 | 1209 |
| Relation | −178.6 | 167.5 | 0.0 | 8.8 | 840 |
| Coverage | −11.0 | 6.8 | 0.0 | 1.2 | 249 |
| Source | −268.9 | 23.3 | 0.2 | 10.1 | 763 |
| Rights | −1118.0 | 205.0 | −23.7 | 108.8 | 856 |

increase is almost always minimal, close to 0 (and mostly negative), except in the element “rights” and in the whole records, where there is a clean decrease in size – but still with a much higher standard deviation.

## 4 Conclusions and Future Work

In this work we tried to find answers that will help predicting the quality of the harvesting. The huge amount of harvested information and the many sources and metadata specialists involved in the metadata creation process makes impossible to get answers from the sources themselves. We therefore prompt for answers by examining the actual metadata, rather than asking people about opinions and practices. We examine such questions by processing harvested information directly from the metadata providers, even though we understand that the results derived from a small sample cannot be very accurate. Therefore, we presented our numbers mostly in charts, and not verbatim, to show the derived tendency. Also, temporary issues may affect the harvesting.

A significant part of the OAI servers cease working in an almost constant rate of about 14 servers per month, while many other serves occasionally fail to respond. Over the years, the sources are updated, and the updates are mostly record-wise rather than element-wise, because changes to the individual descriptive elements seem to have similar pattern and therefore to take place at the same round. The rights and type elements change less often than the other, more descriptive, elements. Most changes seem to be rather maintenance changes and not record enrichments, as they do not increase the words in the content.

In the future, we can also examine the non responding servers in more detail, and derive and cluster the failure reasons.

## References


1. Bui, Y., Park, J.: An assessment of metadata quality: a case study of the national science digital library metadata repository. In: Moukdad, H. (ed.) CAIS/ACSI 2006 Information Science Revisited: Approaches to Innovation. Proceedings of 2005 Annual Conference of the Canadian Association for Information Science Held with the Congress of the Social Sciences and Humanities of Canada at York University, Toronto, Ontario (2005)
2. Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., Sølvberg, I.: Evaluation of digital libraries. Int. J. Digit. Libr. **8**(1), 21–38 (2007). Springer
3. Hughes, B.: Metadata quality evaluation: experience from the open language archives community. In: Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E., Lim, E.-p. (eds.) ICADL 2004. LNCS, vol. 3334, pp. 320–329. Springer, Heidelberg (2004). doi:10.1007/b104284
4. Kapidakis, S.: Comparing metadata quality in the Europeana context. In: Proceedings of 5th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA 2012), Heraklion, Greece, 6–8 June 2012. ACM International Conference Proceeding Series, vol. 661 (2012)
5. Kapidakis, S.: Rating quality in metadata harvesting. In: Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA 2015), Corfu, Greece, 1–3 July 2015. ACM International Conference Proceeding Series (2015). ISBN 978-1-4503-3452-5
6. Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., Saylor, J.: Metadata aggregation and "automated digital libraries": a retrospective on the NSDL experience. In: Proceedings of 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2006), pp. 230–239 (2006)
7. Moreira, B.L., Goncalves, M.A., Laender, A.H.F., Fox, E.A.: Automatic evaluation of digital libraries with 5SQual. J. Inform. **3**(2), 102–123 (2009)
8. Ochoa, X., Duval, E.: Automatic evaluation of metadata quality in digital repositories. Int. J. Digit. Libr. **10**(2/3), 67–91 (2009)
9. Ward., J.: A quantitative analysis of unqualified dublin core metadata element set usage within data providers registered with the open archives initiative. In: Proceedings of 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2003), pp. 315–317 (2003). ISBN 0-7695-1939-3
10. Zhang, Y.: Developing a holistic model for digital library evaluation. J. Am. Soc. Inf. Sci. Technol. **61**(1), 88–110 (2010)