Consistency and Interoperability on Dublin Core Element Values in Collections Harvested using the Open Archive Initiative Protocol for Metadata Harvesting

Sarantos Kapidakis[a]

Department of Archival, Library and Information Studies, University of West Attica,

Keywords: Dublin Core, Metadata, OAI-PMH, Harvesting, Language, Controlled Vocabularies, Controlled Terms, Repeated Values, Linked Open Data, Dendrogram.

Abstract: When resource descriptions use the exact same value for an entity, this value is easier parsed, identified and utilized by automatic procedures. The use of controlled values, even when it is common and very useful, it is usually not enforced during the data entry. In this paper we study the use of the controlled values in many harvested collections and we study all Dublin Core elements and also their similarity. We mainly focus in the element *language*, as there is a lot of standardization on how to denote language values, followed by other elements that normally use controlled values. We discovered values that are repeated many times and in many collections and many more values that are used only once! The lack of coordination among collections during their creation results to many variations for each value, even when the value is used consistently and many times inside a collection. The study uses dendrogram to reveal the current usage of the Dublin Core elements inside and among active collections by clustering the collections with similar values and helps adopting better guidelines, designing better tools and improving the effectiveness of the collections.

1 INTRODUCTION AND RELATED WORK

In many cases, metadata elements of many resources may share the same value. E.g. some resources may have the same creator, who should always be denoted with the exact same value, so that all these resources will be retrieved on each appropriate search request.

Many elements can take controlled values and often most of them do. The *type* element takes controlled values most often, and has many value repetitions, and less variety of values, been followed by the element *language*. But the values of element *language* are more inter-operable, as there are more standards and good practices defining them.

Elements expressing language and type of the resource often take controlled values, forming groups of resources with the same values. Many metadata elements can use controlled values and the library guidelines define which metadata elements should only use controlled values, and how these values should be selected.

When controlled values are not used, metadata creation may be easier and natural language processing tools can be used later to identify the denoted entities. Although this approach provides a huge improvement in the retrieval procedure, the errors that can be introduced are not acceptable in many cases by the library tradition, and libraries continue to use controlled values.

While libraries concentrate on improving their own collections, there are issues that arise when searching across different services, and Maltese in (Maltese, 2018) described the approach that the University of Trento followed. Libraries develop and use various controlled vocabularies for years, and in (Harper and Tillett, 2007) Harper at al. explains how such tools that were developed by the Library of Congress can be used in the development of robust Web services and Semantic Web technologies.

The Linked Open Data[1] (*LOD*) defines a modern trend to universally adopt controlled terms, where each entity of interest (e.g. person, subject, geographic location) has a unique URL that should be

[a] https://orcid.org/0000-0002-8723-0276

[1] http://linkeddata.org

Kapidakis, S.
Consistency and Interoperability on Dublin Core Element Values in Collections Harvested using the Open Archive Initiative Protocol for Metadata Harvesting.
DOI: 10.5220/0010112001810188
In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 2: KEOD*, pages 181-188
ISBN: 978-989-758-474-9

adopted and used to represented it, eliminating any ambiguities and permitting easy correlation and navigation to resources through their common entities. Similar issues apply to the whole Cultural Heritage Information, and in (Baca, 2003) Baca deals with practical issues in applying metadata schemes and controlled vocabularies. For scientific data, the FAIR Guiding Principles, as described by ”Wilkinson et al. in (Wilkinson, 2016), also put specific emphasis on enhancing the ability of machines to automatically find and use the data, which includes unique identifiers, such as controlled values.

An approach to metadata quality evaluation is applied to the open language archives community (OLAC) in (Hughes, 2005) by Hughes that is using many OLAC controlled vocabularies.

The evaluation and quality of metadata is examined as one dimension of the digital library evaluation frameworks and systems in the related literature, like (Moreira et al., 2009; Zhang, 2010). Additionally, Fuhr et al. in (Fuhr et al., 2007) and Vullo et al. in (Vullo et al., 2010) propose a quality framework for digital libraries that deal with quality parameters. Király at al, in (Király et al., 2019), examine the data quality in Europeana, for the purpose of multilinguality, and they consider the consistently of the Dublin Core *language* element in these records.

In (Kapidakis, 2018) Kapidakis studies the presence and the repetitions of the values of the metadata elements from many harvested metadata, examining the number of their statements and the text they contain. He also studied how they are evolving over time. In (Harper, 2016) Harper processes the items of the Digital Public Library of America (DPLA) and demonstrates the “metadata fingerprints” (D3 Star Plots) to visualize the metadata characteristics. He used them to summarize the number of metadata statements per item from different providers, across multiple fields, and also to compare the signatures of the items versus those with at least 1 hit, using google analytics. He also used these “fingerprints” to visualize the word counts, to comparatively study the different Dublin Core elements. He did not try to process, examine the consistency or find patterns in repeated or controlled values. In (Kapidakis, 2019) Kapidakis studies the presence and the repetition of the values of the 15 Dublin Core elements, and comments on individual unexpected values, using a 1000 record sample from each collection. He found that most Dublin Core elements are using controlled values in some collections, and are using free values in many others and that the values in a collection need better conforming to common rules.

In this paper we study the values of each Dublin Core element in many collections, where we mainly focus in the elements that their values have many repetitions, such as the *language* element followed by *type*, *format*, *rights*, *coverage* and *publisher*. We show how the effort to conform to known practices or standards form groups of values with similar patterns. We examine records from publicly available official collections that will normally be used in combination with other collections.

In order to study the Dublin Core metadata we first have to collect them from the *services* that provide them. We run a *harvesting task* for each service, asking for all its records. In many cases such tasks fail, due to permanent or temporary errors. The records that were returned constitute a *collection*, even an incomplete one, in case of an error in the communication resulting to less records.

The rest of the paper is organized as follows: In section 2 we describe our methodology and how we selected our services and used the software we made to create our data-set, and we examine general characteristics of the harvested metadata. In section 3 we study the individual values for the Dublin Core element *language* in all metadata records that the harvesting tasks returned for all collections. We also study how the controlled values are unique or common over many collections, when all metadata are used as one collection. In section 4 we examine the sets of values for the Dublin Core element *language* used in each collection and we present a dendrogram showing their similarity, to reveal the differences among different services. Similar analysis is performed for the other Dublin Core elements, and in section 5 we present dendrograms for some other elements, to demonstrate the differences in the patterns of their values. Finally on section 6 we conclude and present issues for further research.

2 HARVESTING METHODOLOGY AND CHARACTERISTICS OF THE DATA

To harvest and study the provided metadata, we created an OAI-PMH client using the `oaipy` library and used it to ask each service from a list of OAI-PMH services to provide all its metadata records, and to process them. Harvesting tasks are common for the OAI-PMH services, which periodically satisfy harvesting requests for the new or updated records, and involve the exchange of many OAI-PMH requests and responses, starting from the negotiation phase for the

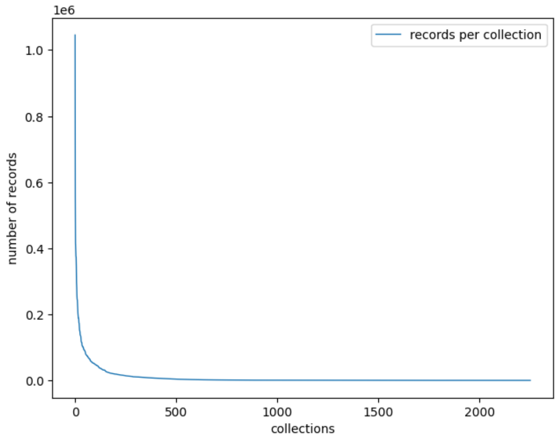

Figure 1: The number of records retrieved per collection for the 2251 collections.

supported OAI-PMH features of the two sides.

The sources listed on March of 2020 in the official OAI-PMH Registered Data Providers[2] site were used as the list of services to harvest and contained 4627 entries. Sometimes the tasks time out resulting to abnormal termination of the task. Instead of using a flat timeout deadline, which would be inappropriate to harvest a large number of records, we set partial timeout deadlines of 15 minutes for each task, which would be extended many times (up to 1 day - we never reached this hard deadline) as long as the task did return any new records. We interrupted any incomplete task when it stopped delivering records and moved on to the next task.

A significant part of the conducted Open Archive Initiative harvesting services did not respond. As a result, only 2251 services returned records. We did not try to recover from harvesting errors or to restart any failing harvesting tasks, but we processed the records we received. Overall, our sequential execution of the record harvesting tasks from their services took more fifteen days to complete.

In Fig. 1 we can see the distribution of the number of returned records for each service. Most collection are small, but there are much bigger ones among them. In Fig. 2 we depict the number of statements per collection, sorted in decreasing order, for each dc element. We can see similar and even overlapping curves for the elements, where most collections have very few statements and few collections have most of the statements.

To better observe the distribution, in Fig. 3 we only show the number of *language* statements, sorted in decreasing order. Only 2062 collections contained *language* statements and 2010 of them contained at least 10 *language* statements.

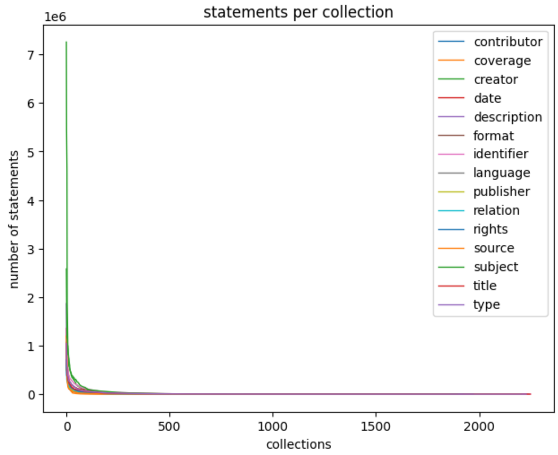

Figure 2: The number of statements for the 15 dc elements, in the 2251 collections.

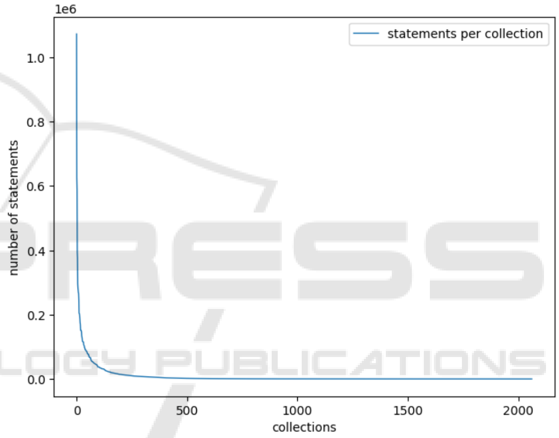

Figure 3: The number of *language* statements, for the 2062 collections.

A big challenge was to harvest the metadata needed for such a study. Enough data have to be collected despite of temporary or permanent errors. The data we used are all publicly available, but they change slightly over time. Getting as many records as possible is essential to get safe conclusions. Overall, the amount of information harvested from these services was huge, and as it was difficult to store, process and manage, we kept part of it, only the one that was needed for our study. In fact, we stored and processed (in memory) about 28GB.

Our harvested metadata consists of 19713013 records with 356105991 statements. In Table 1 we show, for each Dublin Core element (column *element*), the number of collections that include such statements (column *collections*), and the number of statements in all these collections (column *state-*

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

Table 1: Summary data for the occurrences of each DC element in the 2251 collections. The elements with most repeated values are first.

element	collections	statements	values	across
language	2062	16840201	15328	1029
type	2229	25921548	13089	1982
format	2064	13494470	604885	28353
rights	1972	15024492	364138	71249
subject	1896	54073813	7493705	1383596
relation	1919	13139458	7365138	366647
coverage	403	5850572	474233	17530
publisher	2204	14638984	690294	85598
creator	2241	62918155	11107239	2053622
contributor	1167	9415885	2063538	117905
title	2249	21656612	17879231	1149977
identifier	2250	40712655	31833633	1295127
description	2232	21995791	13787533	684706
source	1858	11120103	2626893	126302
date	2232	29303252	3510478	339012

ments). Column *values* counts the number of distinct (after normalization) values in these statements that reveals the degree of repetition of the values. Column *values* contains the number of these values that appear on more than one collection and reveals if the values have a broader (than inside a collection) scope. As an example, 16840201 are *language* statements and contain 15328 distinct values (after normalizing 16059 values), of which 1029 appear on more than one collection.

There are many different values representing the same quantity, such as a language, and we tried to reduce them in an unattended, automatic, element-independent way, by normalizing the values and using only the normalized values afterwards. Our normalization ignores letter case, punctuation and spacing.

Our data reveal that all elements do have both many and few repetitions on different collections. The number of collections with many and few repetitions is different, and *language*, *type*, *rights* and *format* are mostly repeated elements, followed by *relation*, *subject*, *coverage* and *publisher*, both inside a collection and also across collections.

3 INDIVIDUAL LANGUAGE VALUES

In Fig. 4 we depict the number of collections that share each dc element value, sorted in decreasing order. We can see similar distribution with extreme values for all elements. The values for element *subject* are higher than the other elements. This can be explained partly by the fact that there are more subject statements, and also by adopting more common subject schemes in the different collections.

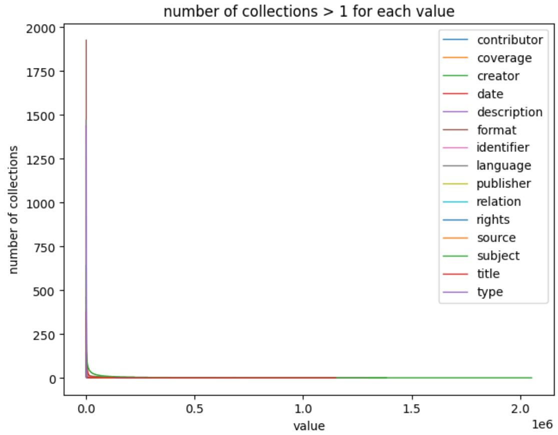

Figure 4: The number of collections that share each dc element value, for the values that are found in more than one collection.

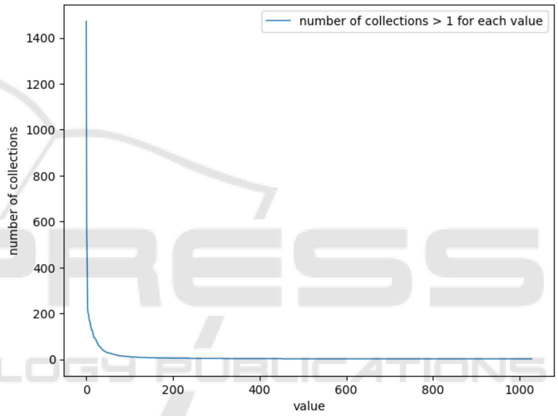

Figure 5: The number of collections that share each *language* value, for the 1029 values that are found in more than one collection (from the 15328 values in total).

To observe the distribution better, Fig. 5 shows the number of collections, sorted in decreasing order, that share each *language* value, for the 1029 values that are present in more than one of the 2062 collections (from the 15328 values in total). The fact that only 1029 language values appear on more than one collection indicate that most other values are most probably ill-formed, as it is very unlikely that so many languages are actually in use. The *language* statements included 15328 distinct values (after normalization).

Compared to other resource properties that are also often repeated (e.g., keywords or resource format), the *language* description depends less on the type or topic of the resources, its semantics is well understood by most people and there exist standards and good practices describing how to denote any language. Additionally, the description of the resource *language* is mostly repeated: many resources contain

eng	1472
en	563
english	127
en_us	89
ingles	10
e	7
enm	5
united states	5
english old ca 450 1100	3
english language	2
Ingilizce	1
egnlish	1

Figure 6: Some values of the element *language* denoting *English* only, with the number of collections they are present.

more than one language, most resources in a collection will be in a few languages, and different collection often have in common some of their languages.

The most common values found are “eng”, “en” and “spa”, while values like “English”, “en_US” and “en-US” are also common. The actual languages of the resources are not so many, but the languages are, unfortunately, mostly specified in many different ways, which makes automatic processing of such values much harder.

Some of the issues in the language specifications are:

- No standard or good practice is followed. E.g., for specifying *English* alone, all of the values in Fig. 6 have been used (only those that are present on more than one collections are shown). The most common value, `eng`, is used 3185233 times in 1472 collections, but the other values are used considerably less.
- Instead of adopting a bare control value, the value includes a verbose description. E.g. `title in Arabic at head of title, preface and afterword in German`, `text in Yiddish script`, `predominately English some Hebrew and Chinese`. These additional details should normally be included in a description element.
- More than one value is stored inside a single statement. E.g. `text in Yiddish; prayers in Hebrew with Yiddish translation and elaboration`, `primarily in Russian with some Hebrew`, `mostly German; some English and French`. Normally, multiple statement should be used for multiple values.
- There are spelling errors in the value. E.g. `Englsih`, `Englishx`. The value content on the original system should provide a choice of the legal values, to avoid such errors
- The value is completely irrelevant, probably inserted by mistake. E.g. `born digital`, `39 x 53 cm`, `1946-1951 in english`, `<--please select language-->`.
- The value uses local language or a non standard character set encoding. E.g. `Anglais`. There is no established provision on specifying a character encoding on specific metadata elements, and most good practices define values in ASCII (which are identical in UTF-8). Other encodings should be avoided for global use.

Similar remarks apply for the values of the other Dublin Core elements.

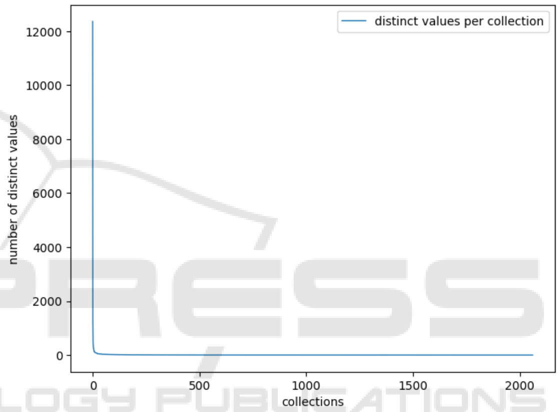

Figure 7: The number of *language* distinct values in each of the 2062 collections.

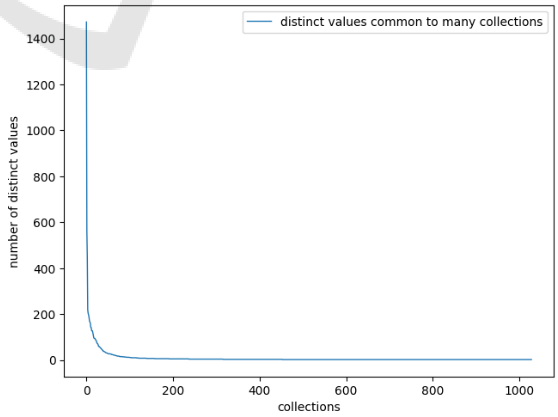

Figure 8: The number of *language* distinct values in each of the 2062 collections, that are found in two or more collections.

The collections we used were created independently and probably have different description guidelines, but are registered to the OAI-PMH registry, to allow metadata aggregation and use in an aggregated context, resembling one big collection. There are no agreed rules among those collections, but we expect them to mostly adopt common practices.

Fig. 7 demonstrates how the same values are found inside a collection and Fig. 8 in different from our 2062 collections, not taking into account how many times a value appears in each collection.

4 CLUSTERING LANGUAGE VALUES

One would expect to see as values for the element *language* either the three-letter language codes (e.g. `eng`, `spa`, `jpn`, ...) or the two-letter language codes (e.g. `en`, `fr`, `de`, ...). Values like these really are the most common ones.

But other values are also reasonable, such as values containing a dialect specification (e.g. `en_US`, `en_ZA`, `pt_BR`, ...) or using language names in English (e.g. `English`, `Korean`, `German`, ...), as they could be automatically converted to more appropriate values.

To cluster the sets of values used, and reveal their similarity, we used as a clustering metric the euclidean distance of the value vectors with their frequencies. A collection joins a cluster if its similarity (inverse distance) to any of the collections in the cluster is above a given threshold.

Thus, if a collections specifies only language `eng` and another one specifies only language `fre`, they have no similarity. When another collections includes both languages `eng` and `fre`, its similarity to both the previous collections is established, although its exact value depends on the frequency of each language value, and the clustering algorithm uses this collection to bridge the distance between the collections and to (eventually) form a cluster with the collections with common language names.

On the other hand, if a collection specifies languages `eng` and `fre` and another collection `en` and `fr`, these collections have no similarity, even though they are specifying the same real languages, because they are using different values for them. If all other collections use either the three letter or the two letter specification (but no both) - or another (uncommon) value schema, the above collections will not become part of the same cluster.

A dendrogram is a diagram representing a tree performing hierarchical clustering on data and showing the resulting tree. The top of the U-link indicates a cluster merge. The two legs of the U-link indicate which clusters were merged. The length of the two legs of the U-link represents the distance between the child clusters. We use dendrograms to show the possible division of the collections to sub-clusters and clusters (depending on where we should set the similarity threshold) and to explore the similarity among them.

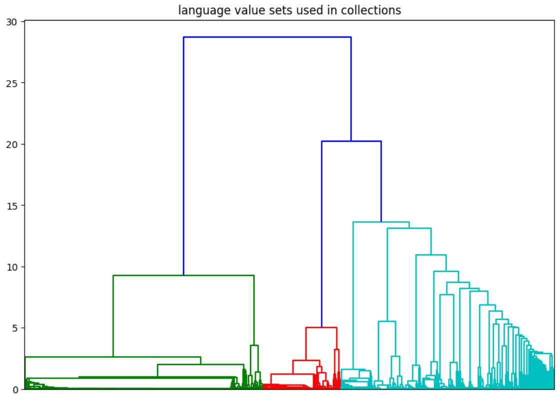

Figure 9: Dendrogram for *language* from 2010 collections.

In the dendrogram of Fig. 9, we can see the cluster distances from the 2010 of our 2062 collections, excluding the 52 tiny ones, with less than 10 *language* statements. We observe that we have 3 big clusters, and we can see the smaller differences inside each cluster.

When two clusters contain mostly different values and any common values are incidental, their distance is high. The high distance of the child clusters also shows that even though the collections use some common values (to justify some similarity), they also contain many different ones, possibly some of them by mistake or from failing to strictly follow rules or practices.

5 OTHER REPEATED ELEMENTS

We also examined the value patterns for other Dublin Core elements. The clusters for the elements *type*, *format*, *rights*, *coverage* and *publisher*, which are the elements with the smallest number of distinct values in more than one collection are also demonstrated through dendrograms.

In the dendrogram of Fig. 10, we can see the cluster distances from the 2211 of our 2229 collections, excluding the 18 tiny ones, with less than 10 *type* statements. We observe that we have 4 big clusters that are quite distant apart (heterogeneous), which are subdivided into further heterogeneous clusters.

In the dendrogram of Fig. 11, we can see the clus-

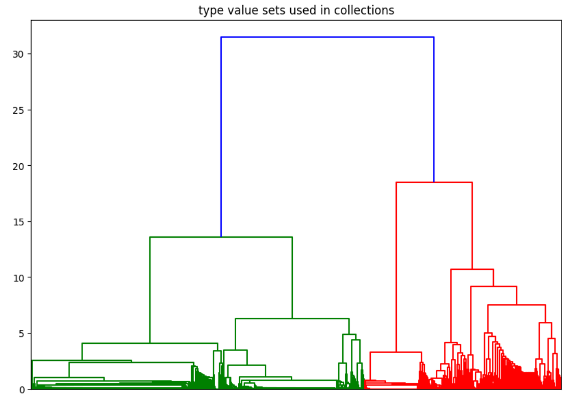

Figure 10: Dendrogram for *type* from 2211 collections.

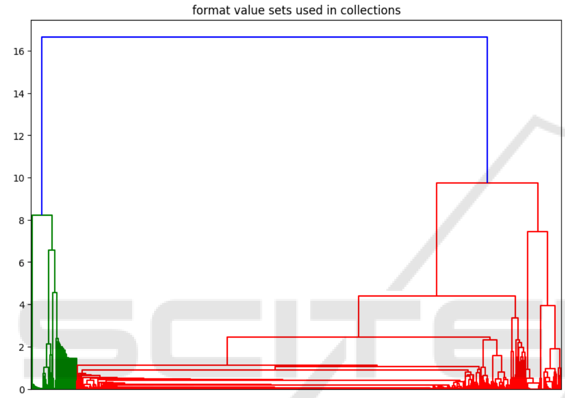

Figure 11: Dendrogram for *format* from 2013 collections.

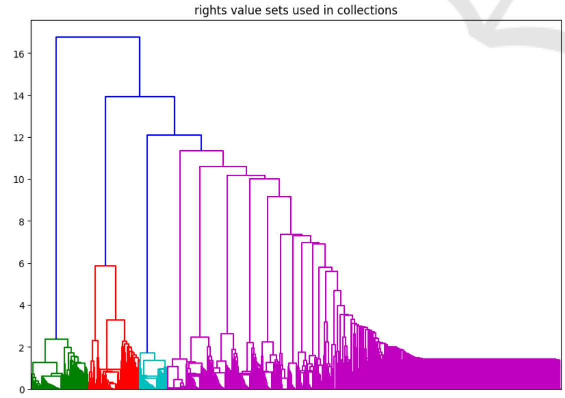

Figure 12: Dendrogram for *rights* from 1873 collections.

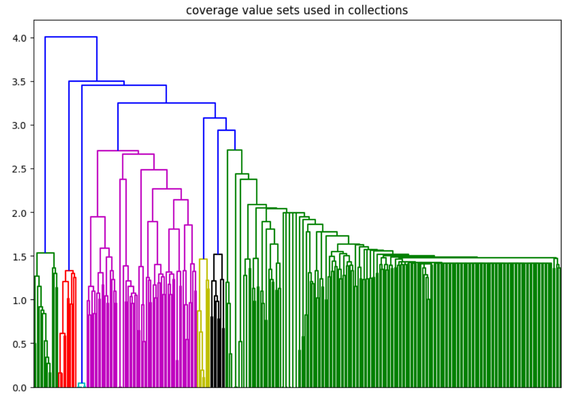

Figure 13: Dendrogram for *coverage* from 319 collections.

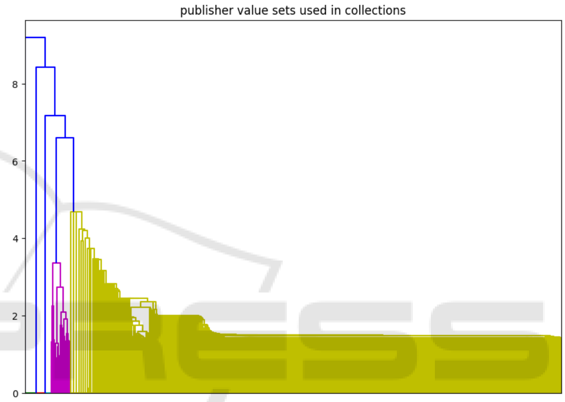

Figure 14: Dendrogram for *publisher* from 2154 collections.

ter distances from the 2013 of our 2064 collections, excluding the 51 tiny ones, with less than 10 *format* statements. We observe that we have 2 very different big clusters, and the cluster to the right has many more collections splitted in groups - with similar values.

In the dendrogram of Fig. 12, we can see the cluster distances from the 1873 of our 1972 collections, excluding the 99 tiny ones, with less than 10 *rights* statements. We observe that we have 4 very different clusters, and the cluster to the right has 70% of the collections and half of them are very similar, while the other half are form many sub-clusters.

The element that is present in the fewest collections is *coverage*. In the dendrogram of Fig. 13, we can see the cluster distances from the 319 of our 403 collections, excluding the 84 tiny ones, with less than 10 from the few collections including *coverage* statements. We observe that we have 7 very different big clusters, and two of them include most of the collections.

The element *publisher* is commonly found in many collections, but its values are in many cases highly repeated but also vendor specific: the collections declare their own publisher values. In the dendrogram of Fig. 14, we can see the cluster distances from the 2154 of our 2204 collections, excluding the 50 tiny ones, with less than 10 *publisher* statements. We observe that we have 1 big cluster, with most collections, another cluster with few collections and few more clusters with one collection each.

6 CONCLUSIONS AND FUTURE WORK

A good practice on metadata descriptions consists of using specific values to identify discrete entities. This is the same approach that leads to the Linked Open Data, when specifying entities from other, open, collections. Additionally, quite often these entities are much fewer than the collection items (e.g. the number of languages of the items, or their type), leading to many value repetitions. This occurs more on some Dublin Core elements (like *language* and *type*), than on some others (like *identifier* and *date*). But it does not occur as often as we would expect: Quite often, there are many values corresponding to the same entity.

We started by studying the *language* element. Even though the language specification information is handled in standard ways in many computer applications, on many harvested metadata the Dublin Core element *language* lacks any standardization.

We examined the *language* values found and their frequencies. In many cases we found illegal or problematic values and we classified them into categories. We used dendrograms to show the similarity of the *language* values among collections.

Nevertheless, there are more common understanding (and also standards) on what the *language* entities are and how to denote them. And still, many items do not adopt the same values, and provide many "unusual" values.

We repeated the procedure for the *type*, *format*, *relation*, *coverage* and *publisher* elements. The situation on all these elements was similar. We could observe the value clustering differences by using dendrograms.

We conclude that we need more standardization on values, more so on *language* values, so that statements across collections follow some good practices. The situation is similar to the other Dublin Core elements, although not identical.

To collect all possible records, we adapted the harvested procedure to handle both reasonable timeouts and large number or records, using a repeated harvesting procedure for many small timeout intervals.

In the future, we could study unique and low repetition values that are similar to other values, on elements with high usual repetition and also repeated values on elements with low usual repetition in order to derive rules and guidelines for automatically creating value mappings.

REFERENCES

Baca, M. (2003). Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information. *Cataloging & Classification Quarterly*, 36(3–4):47–55.

Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.-P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., and Sølvberg, I. (2007). Evaluation of digital libraries. *International Journal on Digital Libraries*, 8(1):21–38.

Harper, C. A. (2016). Metadata analytics, visualization, and optimization: Experiments in statistical analysis of the digital public library of america (dpla). *Code4Lib*, 33.

Harper, C. A. and Tillett, B. B. (2007). Library of congress controlled vocabularies and their application to the semantic web. *Cataloging & Classification Quarterly*, 43(3–4):47–68.

Hughes, B. (2005). Metadata quality evaluation: experience from the open language archives community. *Lecture Notes in Computer Science*, 3334.

Kapidakis, S. (2018). Metadata Synthesis and Updates on Collections Harvested using the Open Archive Initiative Protocol for Metadata Harvesting. *22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, LNCS 10450, Springer*, pages 16–31.

Kapidakis, S. (2019). Repeated values on Collections Harvested using the Open Archive Initiative Protocol for Metadata Harvesting. *11th International Conference on Management of Digital EcoSystems, MEDES 2019, November 12–14, 2019, Limassol, Cyprus, ACM 2019, ISBN 978-1-4503-6238-2*.

Király, P., Stiller, J., Charles, V., Bailer, W., and Freire, N. (2019). Evaluating data quality in europeana: Metrics for multilinguality. In Garoufallou, E., Sartori, F., Siatri, R., and Zervas, M., editors, *Metadata and Semantic Research*, pages 199–211, Cham. Springer International Publishing.

Maltese, V. (2018). Digital transformation challenges for universities: Ensuring information consistency across digital services. *Cataloging & Classification Quarterly*, 56(7):592–606.

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

Vullo, G., Clavel, G., Ferro, N., Higgins, S., van Horik, R., Horstmann, W., and Kapidakis, S. (2010). : Quality interoperability within digital libraries: the DL. *org perspective. In: 2nd DL. org Workshop in conjunction with ECDL*, 2010:9–10.

Wilkinson, M. D. e. a. (2016). The fair guiding principles for scientific data management and stewardship. *Sci. Data*, 3(160018).

Zhang, Y. (2010). Developing a holistic model for digital library evaluation. *Journal of the American Society for Information Science and Technology*, 61(1):88–110.