

Towards Cleaning-up Open Data Portals: A Metadata Reconciliation Approach

Alan Tygel+, Sören Auer*, Jeremy Debattista*, Fabrizio Orlandi*, Maria Luiza Machado Campos+

+ Graduate Program on Informatics – PPGI – UFRJ, Brazil
{alantysel, mluiza}@ppgi.ufrj.br

* University of Bonn and Fraunhofer IAIS, Germany
{auer, debattis, orlandi}@cs.uni-bonn.de

Abstract—This paper presents an approach for metadata reconciliation, curation and linking for Open Governmental Data Portals (ODPs). ODPs have been lately the standard solution for governments willing to put their public data available for the society. Portal managers use several types of metadata to organize the datasets, one of the most important ones being the tags. However, the tagging process is subject to many problems, such as synonyms, ambiguity or incoherence, among others. As our empiric analysis of ODPs shows, these issues are currently prevalent in most ODPs and effectively hinders the reuse of Open Data. In order to address these problems, we develop and implement an approach for tag reconciliation in Open Data Portals, encompassing local actions related to individual portals, and global actions for adding a semantic metadata layer above individual portals. The local part aims to enhance the quality of tags in a single portal, and the global part is meant to interlink ODPs by establishing relations between tags.

I. INTRODUCTION

Analysing large amounts of data plays an increasingly important role in today’s society. However, new discoveries and insights can only be attained by integrating information from dispersed sources.

One approach for addressing the problem of data dispersion are data catalogues, which enable organizations to upload and describe datasets using comprehensive metadata schemes. Similar to digital libraries, networks of such catalogues can support the description, archiving and discovery of datasets on the Web. Recently, we have seen a rapid growth of data catalogues being made available to the public. The data catalogue registry datacatalogs.org, for example, already lists 285 data catalogues worldwide.

Data catalogues where data is supposed to be open, at least in the licensing sense, are usually called Open Data Portals (ODPs). Implementations that show the increasing popularity of ODPs can be seen, for example, in open government data portals, data portals of international organizations and NGOs, as well as scientific data portals.

These ODPs comprise large amounts of structured data, mostly in the form of tabular data such as CSV files or Excel sheets. They aim to be a one-stop-shop for citizens and companies interested in using public data produced by governments or civil society organisations. Examples are the US’ data portal, the UK’s data portal, the European Commission’s

portal as well as numerous other local, regional and national data portal initiatives.

In the research domain ODPs also play an important role. An example of a popular scientific open data portals is the Global Biodiversity Information Facility Data Portal. Also many international and non-governmental organizations operate ODPs such as the World Bank Data Portal or the data portal of the World Health Organization.

Despite its recent popularity, Open Data and Open Data Portals still face significant impediments, as richly described in [25]. Zuiderwijk et al. collected 118 socio-technical impediments for use of open data from interviews, workshops and literature. Some cited impediments were “absence of commonly agreed metadata”, “insufficiency of metadata”, “the lack of interoperability” and “difficulty in searching and browsing data”, showing that a great challenge for ODPs is the organization of data.

The open data organization challenge can be subdivided into two aspects: 1) structuring and organizing the datasets themselves and 2) providing well-structured and organized metadata for the datasets. The first aspect was, for example, tackled by approaches for semantic lifting of data by [5] and [4], who tried to build general strategies for putting large open government datasets in the Link Data cloud. For the standardized structuring metadata, the Data Catalog Vocabulary (DCAT)¹ [3] was developed. However, the cross-portal metadata alignment and reconciliation can not be addressed by DCAT.

The metadata used to organize datasets in an ODP comprises categories or groups and most importantly labelling with free-text words or sets of words – the tags. The concept of tagging became popular within Web 2.0 services and aggregation tools like del.icio.us. The main advantages of tagging are the ease of classifying, and the crowd effect – resulting in the so called folksonomies – because all users were allowed to tag and share their contents. Tagging datasets in an ODP cannot be considered as folksonomies, because the process is mainly driven by portal managers and data publishers, and not by the actual users. As a result of this, the structuring effect of crowd-tagging and folksonomies is

¹Available at <http://www.w3.org/TR/vocab-dcat/>

missing in ODPs.

A quick look over some ODPs reveals that most of them suffer from a very confusing organization of datasets. The first level of categorization uses the concept of groups. In general, they are stable and meaningful, but normally contain a large number of datasets. A more detailed classification should be done via tags, whose use in ODPs has the following issues:

- *Synonyms*: In most ODPs, there exists large number of synonymous tags, e.g., `crops` and `seeds`;
- *Different writings of the same word*: Several tags are incorrectly written, or have differences in capitalization or accents, e.g., `baden-wuerttemberg` and `Baden-Württemberg`;
- *Lack of relationships*: There is no explicit relationships between the tags, e.g., `Community Centres` is clearly a specialization of `Community`, but this is not explicit;
- *Ambiguity*: As tags are written as pure text, ambiguity is prevalent in ODPs, e.g., the tag `apple`, which could refer to the fruit or to the company; and
- *Incoherence*: Tags do not allow any connection between different portals that use the same or equivalent tags, e.g., two datasets tagged with `budget` in different portals are not connected.

As a result, the navigation, exploration and search within individual, but in particular also across ODPs is significantly hampered.

In this paper, we present an approach and its implementation for improving the tag curation within and across ODPs. Our main contributions are:

- A comprehensive analysis of tag usage in 90 ODPs, which justifies the need and benefits of better tools for managing tags;
- An approach for cleaning and reconciliation of tags in ODPs; and
- An approach for collaboratively connecting ODPs through meaningful shared tags.

The remainder of this paper is organized as follows. After deriving a basic characterization, we present an analysis of the use tags in ODPs in Section III. Section IV shows our approach, both at the local (individual ODPs) and global levels (Tag Server). Section V describes the implementation of the open source tools, while Section VI describes some of the achieved results. The final sections outline related work and draw the conclusions of the paper.

II. CHARACTERIZATION OF AN OPEN DATA PORTAL

According to [2], an Open Data Portal is “a collection of systems set up to make Open Data used and useful”. A formal definition of an ODP can be found in [22]. However, in that case, the focus is general metadata analysis, which turns their definition unsuitable to be used here.

Figure 1 shows the relevant entities and relations that are used in the remainder of this paper. An *Open Data Portal*, in this context, is a collection of datasets, which hold open data resources online. Each *Dataset* belonging to an ODP can

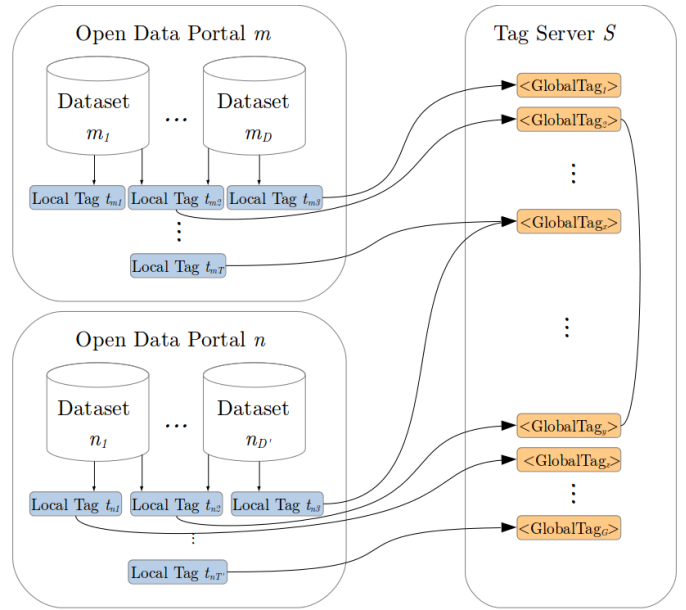


Fig. 1. Relevant elements of the Semantic Tags for Open Data Portals system.

be tagged with local tags. Each *Local Tag* also belongs to an ODP, and can be used to tag one or more datasets. In this architecture, local tags are connected to *Global Tags*, stored in a collaborative *Tag Server*. Several local tags from different ODPs can be associated to a single global tag, which can also have semantic relationships with other global tags.

III. ANALYSIS OF THE USE OF TAGS IN OPEN DATA PORTALS

In this section, we profile the use of tags in Open Data Portals. The analysis is restricted to systems running CKAN², the standard open-source software for ODPs. The CKAN community publishes a census³, where 139 portals are listed. Through the API offered by the software, we tried to obtain data from all portals, but only 90 responded adequately when the assessment was performed (Sep. 2015). Reasons for the lack of availability were mainly that the portal was completely offline, the API was disabled or not responding at the same URL of the website or the portal was using an outdated version of CKAN.

Most of the ODPs are related to governments and public administrations at local, regional, national or continental levels. Some of them are also focusing on specific themes, such as energy or geothermal data. Although most ODPs are authoritative and run by governments and public administrations, some of the portals were built as civil society initiatives. A complete list of the analysed ODPs is available online⁴.

The analysed ODPs are quite heterogeneous. The number of datasets in each portals varies from 3 to 147,485, and the number of tags, from 3 to 49,189. Regarding the quality of

²Available at <http://ckan.org>

³Available at <http://ckan.org/instances>

⁴<http://bit.ly/1NGygtk>

TABLE I
SUMMARY OF DATA USED IN THE EXPERIMENT.

Portals	90
Datasets	389,913
Overall tags	220,567
Unique tags	148,657
Average tags per portal	2451
Average tags per dataset	3.88
Association with semantic resources	36%
Groups	1500
Average groups per portal ⁶	21.43
Average datasets per group ⁷	67.45

the portals, although there is no general benchmark, *Open Data Monitor* attests a high heterogeneity within European ODPs. An informal quality assessment using the Five Stars of ODPs [2] also shows that portals vary from simple data registries (one star) to a common data hub (five stars).

A summary of the experiment data is shown in Table I. The code used to collect and analyse the data is available as an open-source project⁵.

The analysis is divided in two groups: local metrics, to analyse the quality of tags in a particular ODP, and global metrics, looking at the interrelations between portals, and with the Linked Open Data (LOD) cloud.

Regarding the other main tool for organizing ODPs – groups – Table I also shows the number of groups per portal, and the number of datasets inside each one. While the tags are attributed to an average 3.88 datasets, groups contain a mean value of 67.45 datasets. This makes groups less selective than tags, which justifies our decision to focus on tags in this work. Moreover, while all 90 portals use tags, 20 do not use groups to organize data.

A. Local Metrics

1) *Tag Reuse*: The objective of this metric is to assess whether a single tag is being used to characterize several datasets, just a few or even only one. Creating new tags for each dataset can be considered a bad tagging practice. If tags are reused for several datasets, tag-based information retrieval will be more effective. Figure 2 shows the distribution of the percentage of tags used only once for each portal. The graphic shows a peak around 70% of the tags used only once. From the 90 portals, 78 use more than 50% of the tags only once. As a conclusion, tag reuse can be considered very low, thus effectively preventing the tags to be a suitable means to improve navigation, exploration and retrieval of datasets from ODPs.

2) *Tags per dataset*: This metric assesses the number of tags used per dataset. The goal is to verify, as in [22], if the tag metadata is being actively used in the portals. However, the results of this metric cannot lead to further conclusions, since there is no optimal value for the number of tags per

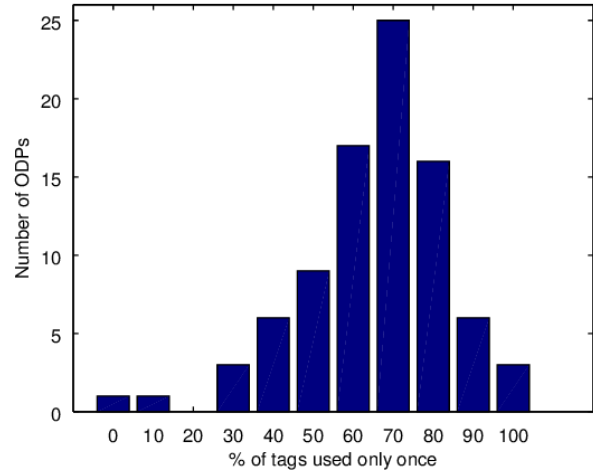


Fig. 2. Re-use of tags inside a portal. The graphic shows the distribution of the percentage of tags used only once.

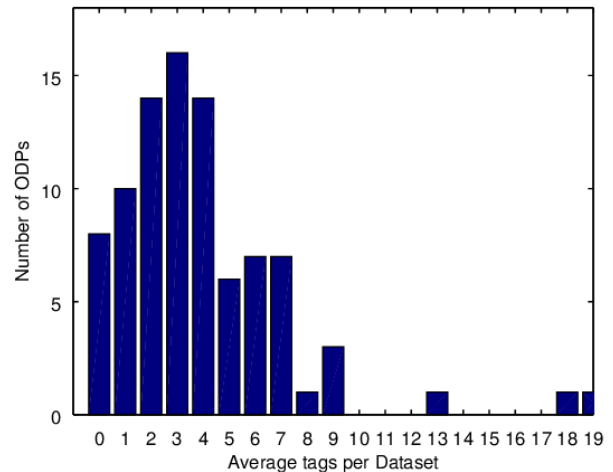


Fig. 3. Distribution of the average number of tags used per dataset in Open Data Portals.

dataset. Using few and consistently used tags may support the organization of datasets better than many incoherently used ones. On the other hand, few tags may not label the content adequately. Figure 3 shows the distribution of the average tags per dataset for each portal. We can see that most ODPs apply between 1 and 7 tags to each dataset, with a peak around the value of 3. In general, we can affirm that describing datasets with tags is a common procedure in ODPs.

3) *Tag similarity*: By looking at the ODP tags, one can readily recognize that many tags differ only on capitalization, accents or singular and plural forms. Thus, this metric assesses whether several tags are being used with the same meaning. While recognizing these cases is easy for humans who understand the language of the tags, an automatic discovery of tags with the same meaning is not always straightforward. A simple approach is to convert the tags to lowercase and

⁵<https://github.com/alantysel/StodAp>

⁶Excluding ODPs which do not use groups.

⁷Excluding void groups.

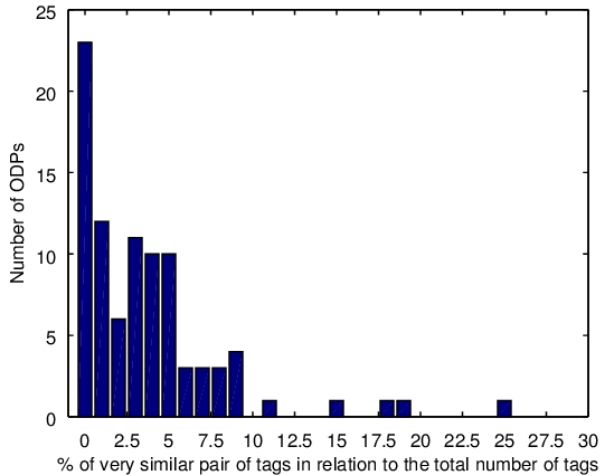


Fig. 4. Proportion of similar tags in ODPs, where the difference lies only capitalization or special characters.

unaccented strings for comparison. Despite its simplicity, this method catches a significant number of cases such as `birth` and `Birth`.

A second possibility is to use the well known Levenshtein edit distance, which can also be suitable for detecting gender and plural differences, in some languages. However, this method fails with tags containing numbers. For example, the Levenshtein edit distance between `budget-2010` and `budget-2011` is the same as between `Access` and `access`. Semantic-oriented methods, as detailed in [9], could also be used to detect synonymous tags.

Figure 4 shows a distribution of the percentage of similar tags inside each ODP. Similarity was checked using the simple approach. The occurrence of a significant rate of similarity reveals that there are few portals adopting a systematic tagging procedure. Despite the low percentage for some portals, in many of them similar tags still occur. Only 20 portals, out of overall 90, revealed no similar tags at all. It should be noticed that these portals use far less tags (148 per portal) than the average of all portals (2451 per portal), which may also be a sign of careful tagging.

B. Global Metrics

1) *Coincident tags between portals*: Different ODPs, especially governmental ones, can publish related data, which may also be tagged similarly. Using the same tag comparison approach as described in the local tag similarity metric, we found that 73,316 tags appeared in more than one ODP, which represents 33% of the total tags. If we are interested in datasets from different ODPs tagged similarly, an overestimation bias may come from the fact that some portals act only as datasets harvesters, replicating the same datasets (and related tags). On the other hand, because portals are available in several languages, different tags could have the same meaning in different languages, what in turn tends to be an underestimation bias. In any case, the figure clearly indicates that there exists

TABLE II
EXPRESSIVENESS OF TAGS

	Absolute Occurrence	Weighted by Usage
Associated to a meaning	23.46%	23.71%
Not associated to a meaning	68.38%	64.20%
Not considered	8.16%	12.09%

great potential for linking tags between open data portals. In fact, with this metric, our aim is to justify and motivate the development of a semantic tag curation approach for open data portals, which will be described in IV-B.

2) *Tag expressiveness*: A way of taking the tagging process one step further is to associate tags with resources or terms described in knowledge bases. In [19], while building the MOAT ontology⁸, the authors designed the association of each tag with a meaning, represented by one or more URIs in the LOD cloud. With the expressiveness metric, our aim is to check if a tag is suitable to be connected to LOD cloud, i.e., if there are possible resources to represent its meaning.

In order to search for candidate resources for the tags, we used Lexvo.org [17], a service that offers connections to different semantic knowledge bases, in several languages. By providing a term (in our case, the tag) and its language, Lexvoc.org returns the corresponding resources, either as `rdfs:seeAlso` or `lexvo:means`.

Table II shows the results. The majority of tags (68.38%) did not correspond to any semantic resource according to this method. 8.15% of the tags were not evaluated either because they contain numbers, or because their length was equal or smaller than three. In those cases, results are mostly wrong. For 23.46% of the tags, at least one meaning or equivalent term was found, and their use represent a similar magnitude of 23.71%.

It is not possible to guarantee that all associations were meaningful, and even worse, that the meaning intended by the tagger was correctly captured. The tag language was estimated by the ODP locale, which can also be a source of errors if not correctly set. Further evaluations are needed in order to estimate the potential that ODP tags have to be connected to the LOD cloud. However, we see that at least one fifth of the tags correspond directly to a semantic resource. Providing context and a stemming pre-processing would probably enhance this result. Thus, we can say that some semantic potential is present on the tags.

After this analysis, we can affirm that: (i) tags in ODPs are widely used, but in a non-systematic way, which hinders their capacity of supporting information retrieval, and (ii) there is a potential for using these tags as connecting elements between ODPs, and for raising semantics from them. Next, we describe our proposal based on these statements.

⁸<http://muto.socialtagging.org/mirror/moat.rdf>

IV. SEMANTIC TAGS FOR OPEN DATA PORTALS - THE STODAP APPROACH

In this section, we describe a tag reconciliation approach for cleaning up ODPs, supported by software tools both at the local and global contexts. The objective is to tackle the main problems identified by the metrics described in the previous section, and thus to facilitate data organization and linking through metadata descriptions of ODPs.

Figure 5⁹ shows an overview of the proposed approach. Data publishers in charge of ODPs are offered tools for local tag curation. These tags are then connected to global tags hosted in a central Tag Server, which can be collaboratively edited both by data consumers and publishers. They can add new semantic descriptions to the global tags, establish relations between them, and also create new links between global and local tags. Data consumers have the option to retrieve data directly from ODPs, or through references gathered from the central server. The description of these actions is shown in the sequel.

A. Local Approach: cleaning up tags inside an ODP

Subsection III-A showed that ODPs suffer from low reuse of tags, and that there is a significant tags duplication due to slight spelling differences. In fact, both problems – low reuse and duplication – are connected, since merging similar tags improves tag reuse. However, low tag reuse can be also attributed to the lack absence of a standard tagging procedure, which would guide users in this task.

To address this problem locally at a particular ODP, we propose an approach for reconciliation of tags. First, we offer three levels of semi-automatic tag merging strategies:

- 1) With high confidence, we suggest merging tags that differ only by capital letters or special characters. In many ODPs, this strategy will already achieve significant results, as shown in Figure 4.
- 2) After running the first strategy, the Levenshtein distance is computed for all remaining pairs of tags. Tags with distance one or two are suggested for merging, in order to catch plural/gender variations, such as *worker* and *workers*. However, false-positives like *widow* and *window* may appear. Tags containing numbers (to avoid merging tags containing years) or less than 4 characters are not included.
- 3) Finally, we use semantic measures [8] to determine the semantic similarity between two tags. In this case, the tags *autumn* and *fall* have a high similarity, and thus will be suggested for merging.

It must be noted that all these approaches have originally quadratic time complexity, because every pair of tags has to be computed. However, sorting tags alphabetically turns the problem into linear in strategies 1 and 2 (however, with possible losses in 2), and ignoring tags without correspondence in dictionary reduces the dimension in strategy 3.

⁹Icons by SimpleIcon from www.flaticon.com are licensed under CC BY 3.0.

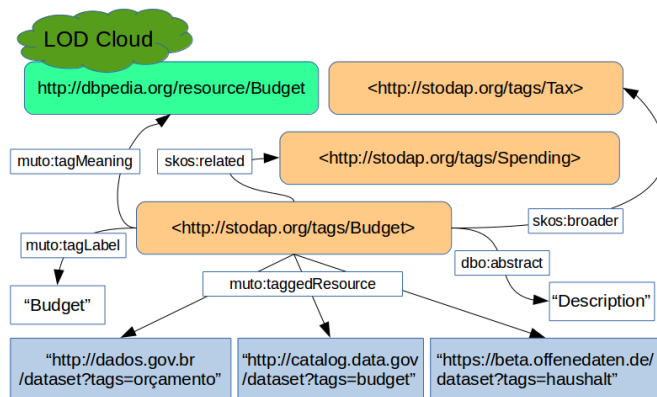


Fig. 6. Example of the StodAp model showing relationships of the global tag <http://stodap.org/tags/Budget>.

After this cleaning procedure, we offer users the opportunity to link each local tag to a global correspondent at the tag server, described in the sequel.

B. Global Approach: a model for linking tags between ODPs

With the aim of building a common and collaborative basis for interlinking ODPs, we developed a Global Tag Server. The conceptual rationale is:

- 1) To assist individual ODPs enhancing the quality of their tags, by assigning a common agreed meaning to them;
- 2) To create a collaborative platform for meaningfully linking ODPs.

The Global Tag Server hosts the description of global tags. Each global tag may be associated to one or more Linked Open Data resources, representing their semantic meanings. Linking to the local tags is accomplished via the URIs which represent a local tag in its context. The global tags can also have several types of relations between each other, such as *skos:broader*, *skos:narrower* or *owl:sameAs*. Figure 6 illustrates the concept with an example.

The example shows the global tag identified by the URI <http://stodap.org/tags/Budget>. With this global tag, a meaning and some URIs of local tags are associated. The global tag is also semantically related to other global tags, using the SKOS vocabulary. The MUTO ontology¹⁰ is used to define some concepts and relations between the tags, like *muto:Tag*, *muto:taggedResource*, *muto:hasTag* and *muto:hasMeaning*.

V. IMPLEMENTATION

In order to support the StodAp approach, we describe in this section some software tools that were implemented. For the local tag curation, we implemented two CKAN plugins: (i) *CKAN Tag Manager*¹¹ and (ii) *CKAN Semantic Tags*¹².

¹⁰<http://muto.socialtagging.org/core/v1.html>

¹¹<https://github.com/alantygel/ckanext-tagmanager>

¹²<https://github.com/alantygel/ckanext-semanticstags>

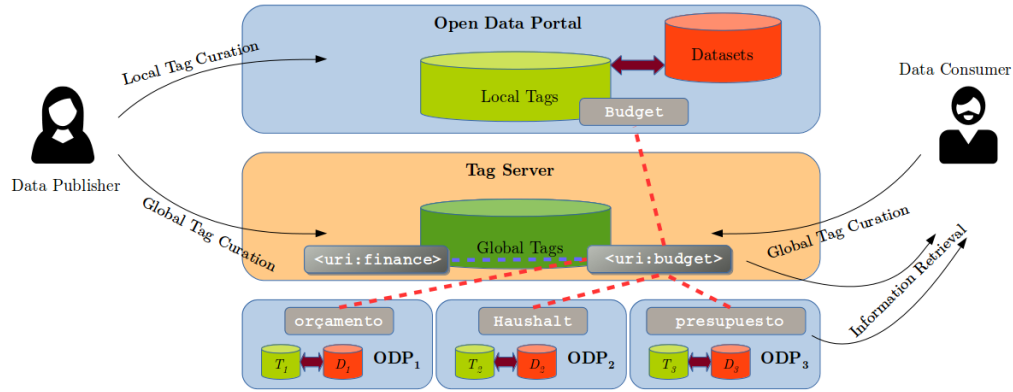


Fig. 5. Overview of the StodAp approach. Local tags are connected to a corresponding global tag within a central tag server. Data managers responsible for an ODPs may use tools for local tag curation, as well for maintaining the tag server. This task is also expected to be performed by data consumers.

The CKAN Tag Manager one offers an environment for tag curation directly inside the CKAN platform. It comprises basic functions such as deletion and editing of tags, and advanced function aimed to enhance the quality of tags. In this sense, the plugin checks:

- Very similar tags, differing by capitals or special characters;
- Similar tags, with a Levenshtein distance ≤ 2 (after lowercasing and unaccenting)
- Possible synonyms, using Natural Language Toolkit [1].

In all those cases, the user is offered the option of merging the respective pair of tags. Figure 7 shows a screenshot of the CKAN Tag Manager.

The CKAN Semantic Tags plugin implements the connection between a CKAN instance and the Global Tag Server. Each local tag can be associated to a global tag from the server. After the association, datasets linked with a local tag also point to the global server, as shown in Figure 8.

The tag server is implemented using the collaborative *MediaWiki*. Specially, the *Semantic MediaWiki* extension [13] is used in order to include properties and integrate the global tags in the LOD Cloud, through the export of RDF files. The page of a global tag is shown in Figure 9. Each global tag page is build using semantic templates and forms, in order to facilitate consistency and coherency and to be more user-friendly.

VI. RESULTS

We describe in this section some results achieved with the STODaP approach. At the global level, it was possible to implement the global tags server and to test the performance.

A. STODaP Server

In order to test the system, an open-source implementation of STODaP was created and deployed at <http://stodap.org>. The following approach was used create 663 global tags at the server:

- From the 220,567 tags harvested, we selected the 663 that were used in more portals, representing all tags used in 10 or more portals;

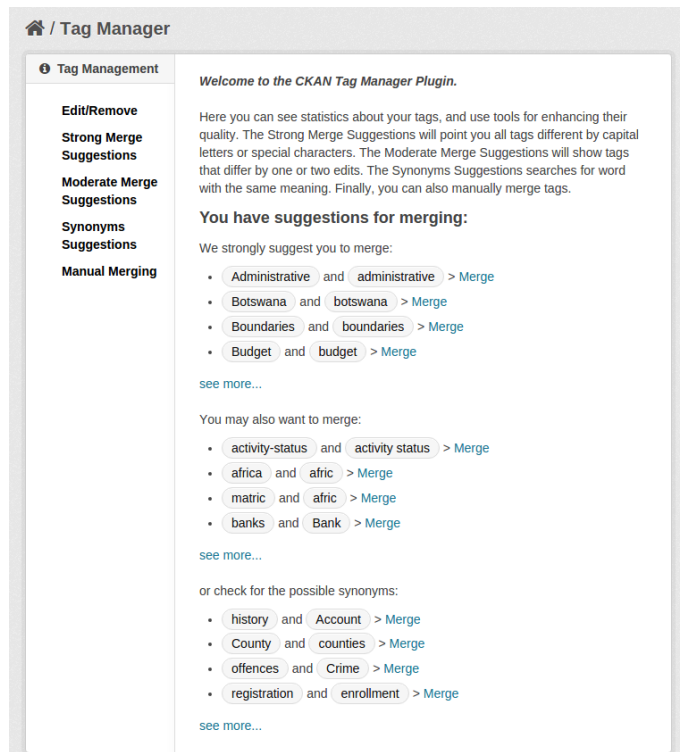


Fig. 7. Local tag curation in a CKAN instance. The plugin offers possibilities of manual and semi-automatic tag merging. The first block contains only valid suggestions, while the second block shows 2 false-positives. The synonym module also detected plurals. Tags in this example were extracted from the africapendata.org portal.

- Using the Lexvo.org service, we found URI candidates to represent the tag meaning via the `lexvo:means` property;
- Using the Lexvo.org service, we found translations and synonyms for the tags via the `rdf:seeAlso` and `lexvo:translate` properties;
- We searched for the translations and synonyms in the harvested tags and included the results as



Fig. 8. Detail of dataset in an ODP. The dataset is tagged with two tags, and one of them (alimentos) is connected to the global tag <http://stodap.org/tags/Food> through the `muto:hasTag` property.

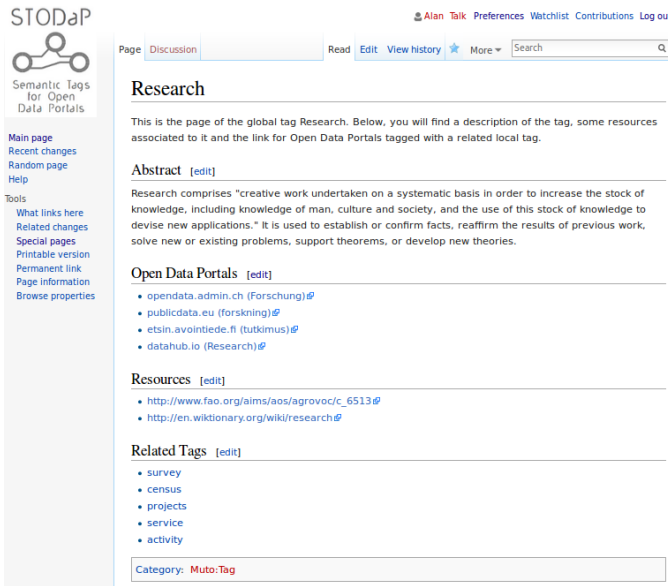


Fig. 9. Semantic Tag Server for Open Data Portals. Simplified example of the global tag `Research`, which is linked to 48 Open Data Portals and 16 semantic resources. In the screenshot, we illustrate a few of them.

`muto:taggedResources`, together with the portals tagged with the original term;

- Using the Natural Language Toolkit Library, we searched for semantic similar global tags, which were added as `skos:related`.

The occurrence of the original tags among the portals, and the results after including the translations and synonyms can be seen in Figure 10. The graphic shows the 30 most used tags, and the achieved increment in the number of relations. The occurrence of tags denoting years can also be noticed. Obviously these tags have no synonyms nor translations, and thus no increment is shown. It is also worth mentioning that the tag `test` is the fourth most used one. This fact is probably related to the early stage of development of some portals.

B. Local Level

At the local level, the main potential achievements are at the tag curation process. As shown in Figure 4, a considerable

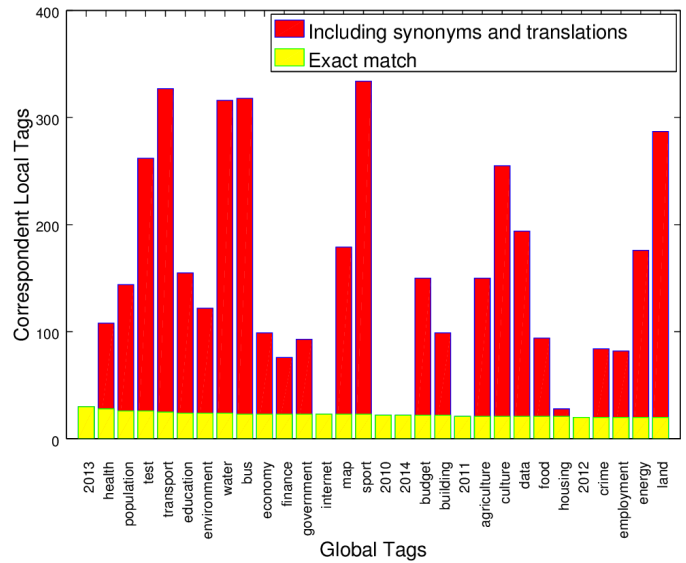


Fig. 10. Correspondence between local and global tags. The yellow bar shows the number of exact occurrences of the tag in ODPs. The red bar shows the improvement when considered translations and synonyms, which can also occur in a same portal. This explains the numbers over 90.

number of pairs of tags differ only by capital or accented characters. Using the naive approach to merge similar tags in every portal would result in reducing the number of 14,168 local tags, which represents 6.4% of the total number of tags. Lowercase and unaccented tags differing by a Levenshtein-distance from 0 to 2 represent a total of 35,066 pairs, or 15.8% from the whole tag universe. However, as discussed above, this approach can lead to false-positives and thus requires manual checking.

VII. RELATED WORK

The vast majority of scientific works about tagging and semantics focus on a different kind of context in relation to ours. Grubbers seminal paper [6], and others such as [7], [16], [12], [18], [10], [11] give interesting perspectives about the tagging activity and its relation to semantics, but always in the folksonomy (or collaborative tagging) context. In this case, tags are attributed to resources by the crowd, passing through a crowd-selection mechanism, which can enhance the tagging quality, but inserts some inherent noise. This is applicable to platforms such as `delicio.us` or `flicker`, where several users can tag the same resource. However, in the open data portals context, tags are only attributed by system managers. Although less noisy, this procedure is biased by few taggers. The tag server approach described in this paper adds collaborative reconciliation layer over the ODPs.

In relation to the metrics for tagging environments, some related ideas could be found in the literature. For example, [22] presents a framework to evaluate the quality of ODPs. Among the applied quality metrics, three of them – *Usage*, *Completeness* and *Accuracy* – are related to metadata keys, which tags are part of. *Usage* establishes which metadata

keys are actually used in a portal; *Completeness* evaluates the presence of non empty values; and *Accuracy* checks if metadata adequately describes the data. However, this metric is not applied for tags.

Laniado and Mika did a similar analysis over hashtags on Twitter [14]. Their work is focused in answering if Twitter hashtags constitute *strong identifiers* for the semantic web. To achieve this, four metrics are used: frequency of hashtags; specificity, which is the deviation from the use of them without being a hashtag; consistency; and stability over time.

The problem of semantic lifting in ODPs was tackled by [5], [4]. In [23], a strategy for lifting datasets in ODPs to the Linked Data cloud is presented. In all these works, however, the semantic lifting refers to the datasets, and not to metadata.

There also has been some work done with regard to metadata reconciliation [15], [24]. However, to the best of our knowledge none of them has been specifically applied to open data portals or leverages tag curation as proposed by STODaP.

VIII. CONCLUSIONS

In this paper, we presented an approach for metadata reconciliation among Open Data Portals. The use of tags was analysed, and several problems were found, such as a low tag reuse rate and the overall existence of different tags for the same meaning. On the analysis we also found that several portals share the same tags, showing that tags have a good potential to be linking elements among datasets. Converting tags into semantic identifiers was also shown as a viable option, even though more sophisticated methods have to be investigated. Based on these findings, we derived the STODaP approach, which comprises two parts: a local one, aimed at cleaning up and enhancing the quality of ODPs tags, and a global one, for connecting ODPs through semantic tags. The implementation of both shows that significant enhancements can be achieved both at the individual ODPs and the global levels.

Future research and development includes a tag suggestion approach for ODPs which takes into account the related tags at the tag server, using collective knowledge as in [21]. Using the possibly structured data of the ODPs in order to improve tagging suggestions is also a research direction that should be followed. At the global level, an interesting approach is to detect the emergence of schemas from the tags, as described in [20]. We will also call for the attention of the open data community in order further to advance collaborative strategies for enriching the tag server. For STODaP to realize its full potential, ODP administrators and users should be involved and (meta)data literacy needs to be improved.

ACKNOWLEDGEMENTS

This work was supported by a grant of the European Commission within the Horizon2020 framework programme for the project OpenBudgets. A. Tygel is supported by CAPES/PDSE grant 99999.008268/2014-02. M. L. M. Campos is partially supported by CNPq–Brazil.

REFERENCES

- [1] S. Bird, E. Loper, and E. Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.
- [2] P. Colpaert, J. Sarah, M. Peter, E. Mannens, and R. V. de Walle. The 5 stars of open data portals. In *7th MeTTeG*, pages 61–67, 2013.
- [3] R. Cyganiak, F. Maali, and V. Peristeras. Self-service linked government data with dc4t and gridworks. In A. Paschke, N. Henze, and T. Pellegrini, editors, *I-SEMANTICS*. ACM, 2010.
- [4] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangquan, J. Flores, D. L. McGuinness, and J. A. Hendler. {TWC} LOGD: A portal for linked open government data ecosystems. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):325–333, 2011.
- [5] I. Ermilov, S. Auer, and C. Stadler. User-driven semantic mapping of tabular data. In M. Sabou, E. Blomqvist, T. D. Noia, H. Sack, and T. Pellegrini, editors, *I-SEMANTICS 2013*, pages 105–112. ACM, 2013.
- [6] T. Grubber. Ontology of Folksonomy: A Mash Up of Apples and Organges. *Int'l Journal on Semantic Web & Information Systems*, 3(2), 2007.
- [7] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW*, page 211, 2007.
- [8] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, 30(5):740–742, 2014.
- [9] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain. Semantic Similarity from Natural Language and Ontology Analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254, 2015.
- [10] H. L. Kim, S. Scerri, J. G. Breslin, S. Decker, and H. G. Kim. The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies. In *DC2015*, pages 128–137, 2008.
- [11] H. L. Kim, S. Scerri, A. Passant, J. G. Breslin, and H. G. Kim. Integrating tagging into the web of data: Overview and combination of existing tag ontologies. *Journal of Internet Technology*, 12(4):561–572, 2011.
- [12] T. Knerr. Tagging ontology-towards a common ontology for folksonomies. *Hochschule Furtwangen University*, 2006.
- [13] M. Krötzsch, D. Vrandečić, M. Völkel, H. Haller, and R. Studer. Semantic wikipedia. *Journal of Web Semantics*, December 2007.
- [14] D. Laniado and P. Mika. Making sense of Twitter. In *ISWC*, 2010.
- [15] R. Lawler, H. Yang, K. Woods, and J. Kaminker. Open reconcile: A practical open-sourced ontology-driven webservice. In *EDOCW, 2012 IEEE 16th International*, pages 124–131, Sept 2012.
- [16] A. Marchetti and M. Rosella. SemKey : A Semantic Collaborative Tagging System. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, volume 7, pages 8–12, 2007.
- [17] G. D. Melo. Lexvo . org : Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web Journal*, 7:1–5, 2013.
- [18] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics*, 5(1):5–15, 2007.
- [19] A. Passant. Linked Data tagging with LODR. In *Semantic Web Challenge (ISWC)2*, number 1, pages 1–8, 2008.
- [20] V. Robu, H. Halpin, and H. Shepherd. Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web*, 3(4), 2009.
- [21] B. Sigurbjörnsson and R. V. Zwol. Flickr tag recommendation based on collective knowledge. *Proceedings of the 17th international conference on World Wide Web - WWW '08*, 6:327–336, 2008.
- [22] J. Umbrich, S. Neumaier, and A. Polleres. Quality assessment & evolution of Open Data portals. In *The International Conference on Open and Big Data*, 2015.
- [23] S. van der Waal, K. Wecl, I. Ermilov, V. Janev, U. Milosevic, and M. Wainwright. Lifting Open Data Portals to the Data Web. In S. Auer, V. Bryl, and S. Tramp, editors, *Linked Open Data – Creating Knowledge Out of Interlinked Data*, chapter 9. Springer, 2014.
- [24] S. van Hooland, R. Verborgh, M. De Wilde, and R. Van de Walle. Free your metadata: a practical approach towards metadata cleaning and vocabulary reconciliation. In *Proceedings of the Digital Humanities 2012 Conference*, pages 29–30. Hamburg University Press, July 2012.
- [25] A. Zuiderwijk, M. Janssen, S. Choenni, R. Meijer, and R. S. Alibaks. Socio-technical Impediments of Open Data. *Electronic Journal of e-Government*, 10(2):156–172, 2012.