Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems (CRIS2016)

Community curation in open dataset repositories: insights from Zenodo

Miguel-Angel Sicilia[a], Elena García-Barriocanal[a], Salvador Sánchez-Alonso[a]

[a]*University of Alcalá, Pza. San Diego s/n, 28871 Alcalá de Henares (Madrid), Spain*

Abstract

The increasing concern for the availability of scientific data has resulted in a number of initiatives promoting the archival and curation of datasets as a legitimate research outcome. Among them, dataset repositories fill the gap of providing long-term preservation of diverse kinds of data along with its meta-descriptions, and support citation. Unsurprisingly, the concern for quality arises as in the publication of papers. However, repositories support a larger variety of use cases, and many of them implement minimal control on the data uploaded by users. An approach to tackle with quality control in repositories is that of letting communities of users to filter the relevant resources for them, at the same time providing some form of trust to users of the data. However, there is a lack of knowledge of the extent to which this social approach that relies on communities self-organizing actually contributes to the effective organization inside repositories. This paper reports the results of a study on the Zenodo repository, describing its main contents and how communities have emerged naturally around the deposited contents.

Selection and peer-review under responsibility of Elhadi M. Shakshuki.

Keywords: dataset repositories, Zenodo, communities

1. Introduction

The preservation and availability of research data is a major concern as it affects core principles of scientific practice, including repeating and contrasting experiment and sharing findings. This concern has resulted in a number of initiatives and services that offer long-term preservation of research data in a broad sense, many of them exposing data and its description in open access form. The key features of those initiatives is offering mechanisms for the persistent identification and archiving of datasets together with services for sharing them and making them citeable. This later feature allows for using datasets as a complementary source for research output evaluation as suggested elsewhere [1].

Unsurprisingly, the concern for quality arises as in the publication of papers. There exist data journals or journals that require archival of data as associated to articles. However, repositories support a larger variety of use cases,

* Corresponding author. Tel.: +0-000-000-0000.
E-mail address: msicilia@uah.es

Selection and peer-review under responsibility of Elhadi M. Shakshuki.

and many of them implement minimal control on the data uploaded by users. This is in cases due to the fact that repositories are intended for a wide or even global user community, and it is not economically feasible to implement some kind of centralized quality control. One of the approaches to non-centralized quality control is that of relying on user communities that organize around collection of resources picked from the repository using in many cases topical or discipline-specific criteria. These communities explicitly or implicitly carry out some form of quality control, thus becoming *de facto* the delegates of the quality control of the overall repository. The approach has the interesting attribute of being scalable, as it grows with the community of users of the repository, and externalizes the work of applying selection criteria. Yoon[7] found that these user communities are one of the factors influencing user *trust* in digital repositories.

The approach of removing controls on update, and then using a social or community approach as a quality control mechanism has been used in other online repositories in the past. An example is *Connexions*, a learning material repository that implemented a similar approach to the so-called *lenses*. Kelty, Burrus & Baraniuk[3] describe that approach as a post-publication process. In addition to pointing out to scalability as a property of the approach, they identify additional added values, as "[...] reveal relationships, new contexts of use, and possibilities for reuse that would not be possible if the objects in the repository had a single evaluation by a single reputable source".

Zenodo (https://zenodo.org/) is an online repository hosted at CERN which allows sharing publications and supporting data. The repository was launched May 2013 and it was designed specifically to help 'the long tail' of researchers based at smaller institutions to share results in a wide variety of formats across all fields of science. Some communities are already using Zenodo[2] in their archival workflows, taking benefits also from their integration with the Github platform (https://github.com/).

As Zenodo is intended to support individual researchers, it features no mechanisms to control for the data uploaded. In words of Lars Holm Nielsen, a software engineer based at CERN, who has been working to create the repository, "Researchers can upload files to Zenodo and there's minimal validation of what goes in there, but these community collections essentially allow everyone to create and curate the content and this solves the issue of us otherwise having to validate everything that's uploaded". In its few years of existence, a number of communities have appeared in Zenodo. As Zenodo does not restrict the creation of communities by registered users, their creation and functioning respond only to the will of individuals and communities engaged with the repository. This makes the repository an interesting exemplar of a data curation repository in which researcher behavior manifests both in the growth and actual use of the repository and also in the selection made by communities.

In this paper, we report on a empirical analysis and exploration of the collection of user resources of Zenodo, with an emphasis in looking at the structure of communities.

The rest of this paper is structured as follows. Section 2 describes the materials and methods for acquiring and processing the required data. Then, the analysis of the data is provided in Section 3. Finally, conclusions and outlook are in Section 4.

2. Materials and methods

This section describes the methods used for getting the data and the tools used for their processing.

2.1. Metadata harvesting

The complete collection of Zenodo metadata records was obtained by using the OAI-PMH endpoint provided by the repository. The records were obtained from the `user-zenodo` collection that includes the entire repository, and the recommended format, `oai_datacite3`.

The `Sickle` Python library[1] was used to write a simple OAI-PMH client. As the current version of `Sickle` only supports Dublin Core record extraction, a custom class `DataCite3Record` was coded to extract the data from DataCite format to a flattened structure for further processing.

Metadata was extracted according to the specification of the DataCite metadata schema version 3.1. The DataCite Metadata Schema is a list of core metadata properties chosen for consistent identification of a resource for citation

[1] https://pypi.python.org/pypi/Sickle

and retrieval purposes, along with recommended use instructions. The resource that is being identified can be of any kind, but it is typically a dataset in its broadest sense, meaning any form of data, not necessarily numerical.

DataCite metadata elements are categorized in three groups:

- Mandatory (M) properties
- Recommended (R) properties are optional, but strongly recommended.
- Optional (O) properties are optional and provide richer description.

All the relevant metadata elements that are regularly filled in Zenodo were extracted. Optional metadata elements *size*, *format* and *version* were discarded as they are nor relevant to the current analysis. Recommended elements *related identifier*, *description* and *geolocation* were also discarded for similar reasons.

2.2. Gathering community information

The records harvested contain a URL to their corresponding Zenodo Web page in the metadata field `alternateIdentifier`. This was used to implement a batch scraping process extracting Community information for each resource.

Finally, the data was processed and analyzed using the Anaconda environment for data science (https://www.continuum.io/why anaconda), that builds on the SciPy framework (http://www.scipy.org/) of Python scientific libraries.

3. Results

The `zenodo-user` collection included 3,828 records at the time of the data extraction. This is a subset of the resources searchable in the query interfaces of Zenodo, which may correspond to the fact that Zenodo includes previous collections also.

3.1. Descriptive analysis

Resources harvested are identified via DOIs in around 91% of the cases. Language in metadata is only present in 15% of the resources, of which 97% corresponds to English.

Data repositories are mainly motivated by easing the reproduction of experiments and giving credit for them. This usually entails that the resources are shared under some form of open access license. The rights field in the metadata gathered has "open access" in 89% of the cases, with an additional 2.5% with "Embargoed Access" (which may eventually become open after the embargo period). Closed and restricted access account for the rest. This is consistent with one of the initial motivations of Zenodo, that was giving support for sharing openly research results funded totally or partially by the European Comission.

Dates in DataCite are specified in two elements. *Publication year* is intended in the following way: "the year when the data was or will be made publicly available". The multi-valued and labelled *date* element uses a controlled vocabulary. We only found a significant difference for the case of the subtype "Available". This is in principle surprising, but may be related to this usage note in DataCite: "to indicate the end of an embargo period, use Available". It might be reflecting a shift from the creation of the record to the end of that embargo period, but this would require further study.

A first interesting finding is in the type of resources. The *resourceType-general* variable shows the distribution in the left figure of Fig 2. Resource type-general is a controlled vocabulary recommended by DataCite 3.

Most of the resources are still text-based resources (54%), but the second most abundant resource type (34%) is of *Software* type (described in DataCite as "A computer program in source code (text) or compiled form."), which is worth exploring further. This might be reflecting the increasing concern for reproducible research, in which sharing the software is crucial for research to be transparently and effectively reproduced (as the computer code used to analyze a dataset is the only record that permits others to fully understand what a researcher has done). *Datasets* are in a third position (described as "Data encoded in a defined structure.") with a 12% of the total resources.

The right part of Figure 2 is a decomposition in subtypes of the general ones. However, it has missing values for elements of general type *Software* and *Dataset*, with respectively . As it can be appreciated, it is basically a breakdown of text resource types, and articles are majority.

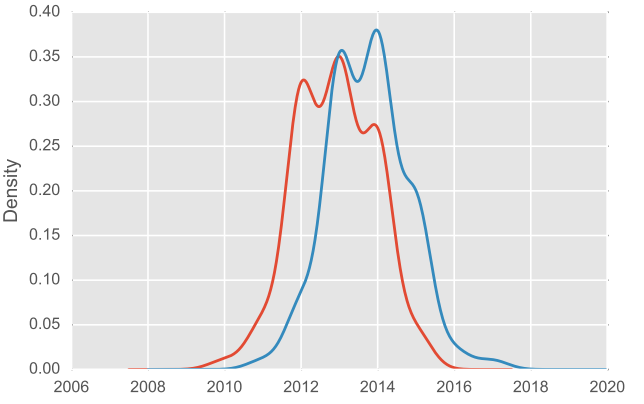

Fig. 1. Kernel density estimation of available day (blue) and publication date (red)

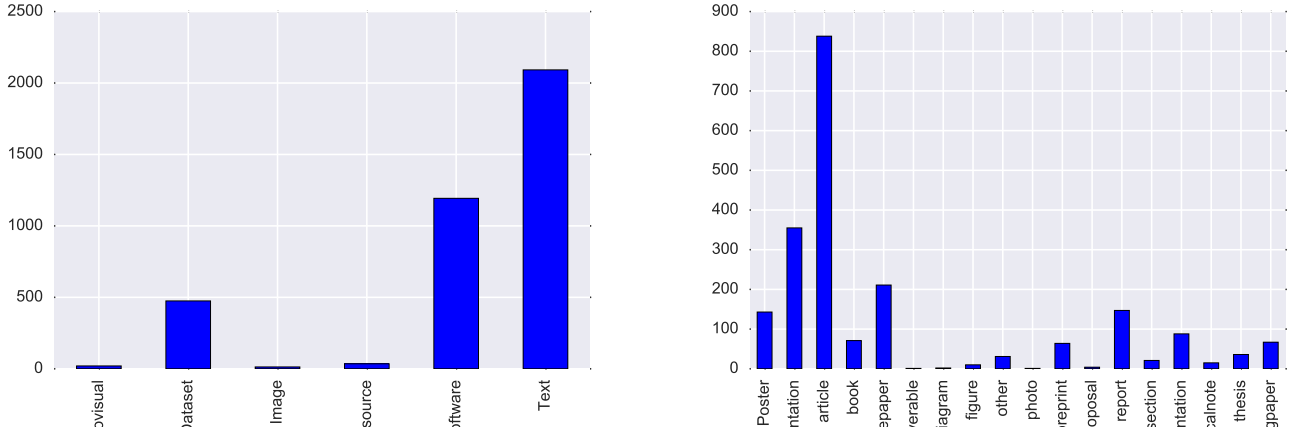

Fig. 2. Distribution of resource types-general (left) and resource-types

3.2. Analysis of community structure

A total of 149 communities were found to be associated with resources in the harvested dataset, and they included around a 46% of the resources.

The distribution of resources per category shows a typical "long tail" with many communities including only one or a few resources and a small number of communities containing much larger numbers. Figure 3 shows a logscale histogram and probability density estimation for the dataset.

The communities above the third quartile are provided in Table 1. As it can be seen there, the list is dominated by the OpenAIRE collection of EC funded research, curated by the Zenodo repository itself. In consequence, the majority of resources are still under the central control of the repository as of today. The second largest community is that of ENI, a middleware platform partially funded by the EC. The third has again an European scope: the branch of the International BASIN Program.

An interesting analysis is that of the scope of communities. For example, the list includes open access publishers as F1000, and contains datasets associated with F1000Research article. The *F1000Research* editorial team curates this collection. This is a case in which a repository complements a publisher model. Another example in this direction is the *British Journal of Medicine and Medical Research*, which follows a open peer review model. It also includes experimental journals as the *Journal of Brief Ideas*.

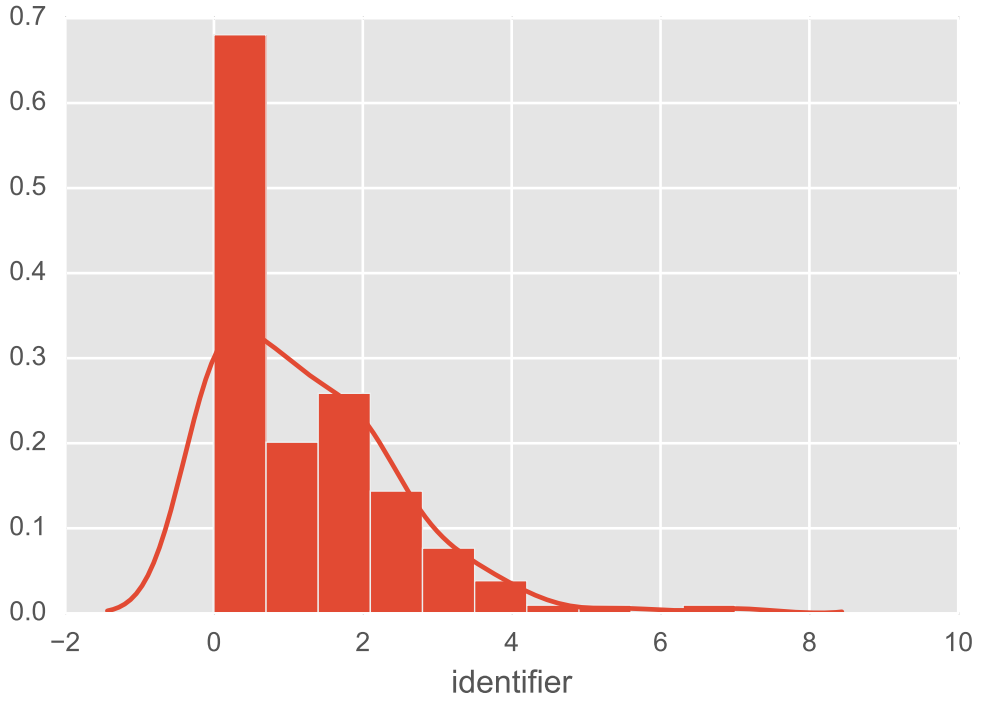

Fig. 3. Distribution of resources by community

It also contains collections from research centers that curate scholarly outputs, as the Harvard-Smithsonian Center for Astrophysics Dissertations & Theses. There are also collections from academic or scientific events.

All the above-mentioned collections are actually traditional sources of curation of research resources, organized around publishing outlets or organizations, rather than actual communities in the sense of non-profit communities or interest organized around digital repositories[4].

In other cases, the collection appears to be topical, and in some cases with a narrow focus, like in the *Drosophila* community. This is reflecting a potential topical organization that may be of interest as an information filtering mechanism. But this is the case of a collection in which the curation policy is based on the object of study.

An special case is that of the Linked Data community that gathers resources related to a particular way of sharing data on the Web. Its curation policy covers however different kinds of resource types: "published peer-reviewed publications; recommendations and standards; software; vocabularies and ontologies; linked data datasets; thesis and student reports". This community is to date containing mostly software resources.

Communities have a small degree of overlap. Only 506 resources appear in more than one community, and as little as 24 in more than two. This points out to independent organizations using the repository for their particular purposes, but shows little reuse. However, other digital repositories that are more focused on reuse have also limited reuse and combination rates[5].

4. Conclusions and outlook

The analysis of the resources of Zenodo has shown that the first non textual resource type more abundant is *software*, followed by *datasets*. This may be a manifestation of the trend towards reproducible research as implemented with open source software packages like R[6].

The mechanism of communities has a potential to become the approach to scaling quality control in the repository. However, to date most part of the resources are still in communities curated by the same repository organization. An

	Community	resources
1	European Commission Funded Research (OpenAIRE)	1102
2	European Middleware Initiative	251
3	EURO-BASIN, North Atlantic Marine Ecosystem Research	74
4	INMiND - Imaging of Neuroinflammation in Neurodegenerative Diseases	63
5	CERN openlab	41
6	F1000Research	40
7	Open Access & Open Science Research	35
8	Small-Scale Air-Sea Interaction	33
9	Nmrlipids	31
10	Research data management (RDM) training materials	28
11	CONFINE project	24
12	British Journal of Medicine and Medical Research	21
13	Harvard-Smithsonian Center for Astrophysics Dissertations & Theses	18
14	BlogForever project	18
15	Biodiversity Literature Repository	17
16	Library and Information Services in Astronomy 7 Posters	16
17	Power Trading Agent Competition	16
18	FP7 DEVOTES project	15
19	British Journal of Environment and Climate Change	14
20	The SAFE project	12
21	10th Biennial HITRAN Database Conference	11
22	Harvard-Smithsonian Center for Astrophysics Science Education Department	11
23	Opportunities for Data Exchange	11
24	CREATe Working Paper Series	10
25	American Chemical Science Journal	10
26	JET Preprints and Reports	9
27	Journal of Brief Ideas	9
28	CfA Post-Doc Symposium 2013	9
29	CLOMMUNITY FP-7 Project: A Community networking Cloud in a box	9
30	Solar REU presentations 2013	9
31	Drosophila	8
32	Linked Data	8
33	Sciencedomain international	8
34	Solar REU presentations 2011	8
35	Solar REU presentations 2014	8
36	Wind Energy	8

interesting insight is that the typology of communities is highly heterogeneous, including publishers, journals and events (symposiums or conferences) and projects but also collections with narrow inclusion criteria based on technology elements, as the Linked Data one. This is reflecting a variety of use cases, in which many communities appear as vehicles of organizing contents rather than a first step towards organizing a community around the contents in the repository. This reflects that the repository is currently used basically as a supplementary platform from organizations, publishers or other communities that develop in parallel outside of Zenodo.

Further research in the direction of the present study should more closely examine the types of software and datasets that are being deposited in this kind of repositories, to actually assess the extent to which they are really conveying the promises of reproducible research.

References

1. Bonilla-Calero, A. (2014). Institutional repositories as complementary tools to evaluate the quantity and quality of research outputs. *Library Review*, **63**(1/2), 46-59.
2. Herterich, P., & Dallmeier-Tiessen, S. (2016). Data Citation Services in the High-Energy Physics Community. *D-Lib Magazine*, **22**(1/2).
3. Kelty, C. M., Burrus, C. S., & Baraniuk, R. G. (2008). Peer review anew: Three principles and a case study in postpublication quality assurance. *Proceedings of the IEEE*, **96**(6), 1000-1011.
4. Plant, R. (2004). Online communities. *Technology in society*, **26**(1), 51-65.
5. RodriguezSolano, C., SnchezAlonso, S. & Sicilia, M. A. (2015). Creation of reusable open textbooks: Insights from the Connexions repository. *British Journal of Educational Technology*, 46(6), 1223-1235.
6. Stodden, V., Leisch, F. & Peng, R. D. (Eds.). (2014). **Implementing reproducible research**. CRC Press.
7. Yoon, A. (2014). End users trust in data repositories: definition and influences on trust development. *Archival Science*, **14**(1), pp. 17-34.