Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 106 (2017) 54 – 60

Procedia Computer Science

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, Plaza de San Diego s/n, 28805 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.

Peer-review under responsibility of the Organizing Committee of CRIS2016

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 feature 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

Peer-review under responsibility of the Organizing Committee of CRIS2016
doi:10.1016/j.procs.2017.03.009

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 any 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 either explicitly or implicitly carry out some form of quality control, thus becoming *de facto* the quality delegates 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. A good example is *Connexions*, a learning material repository that implemented a similar approach by means of the so-called *lenses*. The concept of lens is, in the Connexions repository, a mechanism that facilitates to focus on the content of the repository that is good or useful to a given user or community. Therefore, lenses enable both organizations and individuals to give their "stamps of approval" to content in the repository, allowing for user-driven quality control of modules and collections. 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. Launched in May 2013, the Zenodo repository was specifically designed 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 one of its creators, Lars Holm Nielsen, "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 good 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`.