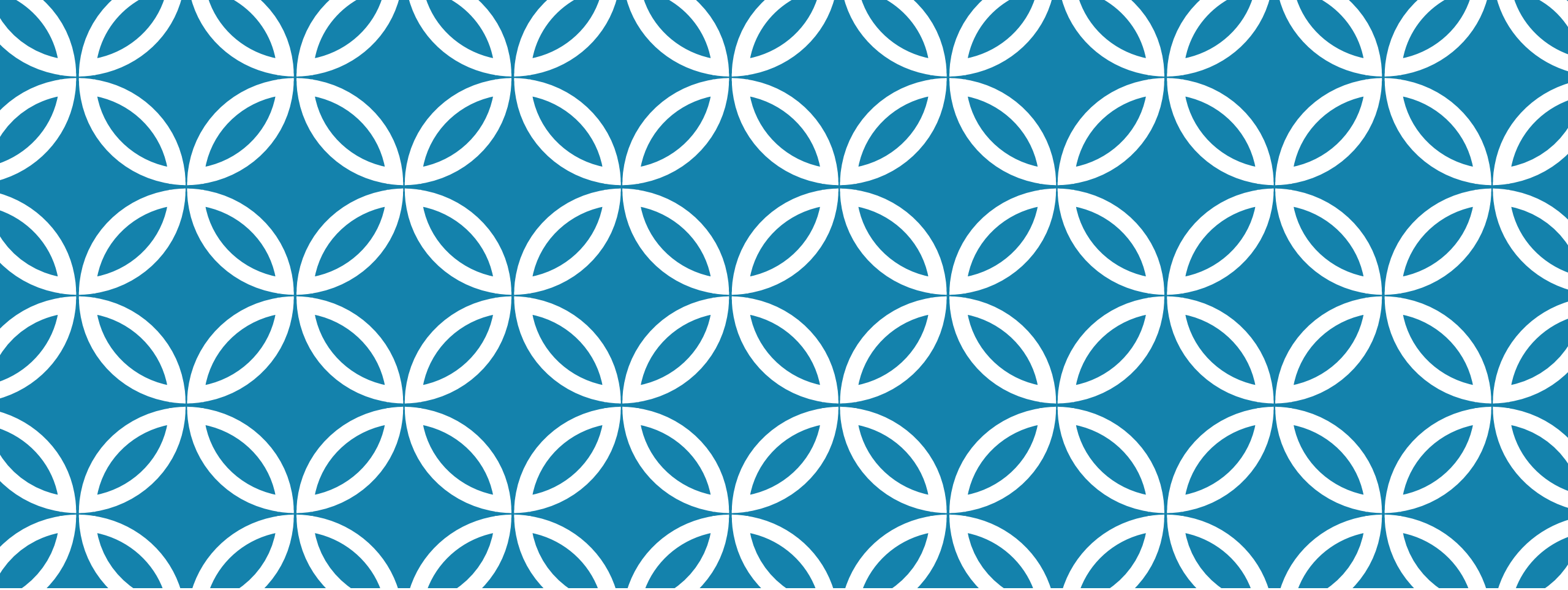


COMMUNITY CURATION IN OPEN DATASET REPOSITORIES: INSIGHTS FROM ZENODO

Miguel-Angel Sicilia, Elena García-Barriocanal, Salvador Sánchez-Alonso
University of Alcalá
msicilia@uah.es

CONTENTS

- Motivation and case
- Approaches to quality control
- The idea of community controlled filters
- Data collection
- Results
- Conclusions and outlook

MOTIVATION

- Proliferation of “data repositories” and “aggregators”.
- Prospects for citable data and recognition of data as a “first-class” actor in research careers.
- Quality of data becomes a major concern.
- No tradition or common practice in filtering relevant datasets.

THE CASE

- Zenodo (https:// zenodo.org/) is an online repository hosted at CERN which allows sharing publications and supporting data.
- Launched May 2013
- to help 'the long tail' of researchers based at smaller institutions to share results.

A DATA REPOSITORY?

Usually included as such.

For example, Asante et al. (2016) in "*Are Scientific Data Repositories Coping with Research Data Publishing?*"

	up to 2010	2011	2012	2013	2014	2015	Total
3TU.Datacentrum	1692	446	379	345	371	296	*3529*
CSIRO DAP	0	46	62	438	454	418	*1418*
Dryad	493	773	1309	1990	2687	2424	*9676*
Figshare	0	16,929	28,224	108,221	94,223	72,818	*320,415*
Zenodo	99	24	68	43	268	1107	*1609*
Total	*2284*	*18,218*	*30,042*	*111,037*	*98,003*	*77,063*	*336,647*

APPROACHES TO QUALITY CONTROL

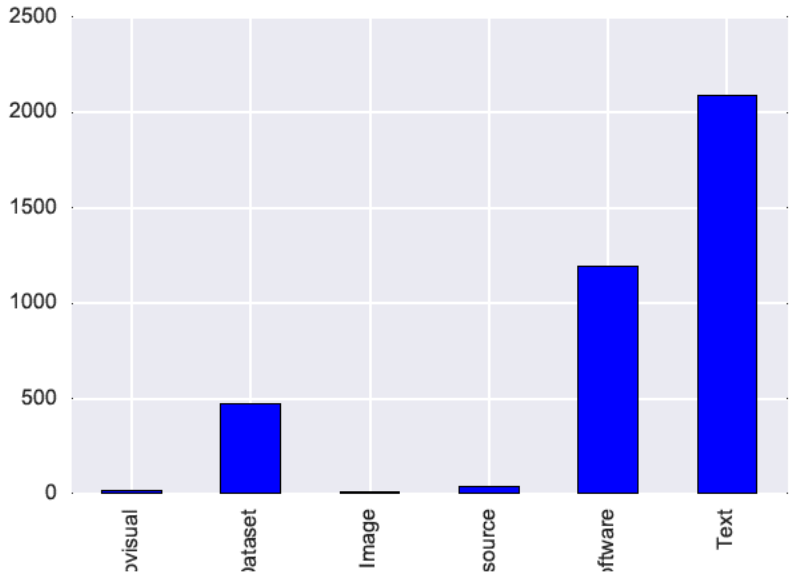
- Different forms of control to publishing (maybe combined):
  - Peer review
  - Formal checks
  - Board check for scope and potential (e.g. arXiv)
- "Community control": no barriers to publishing, social approach.
- Filtering via collections
  - Similar to the one experimented in Connexion's "lenses" for learning materials
  - Relies in a concept of "community of practice" that organizes around the repository.

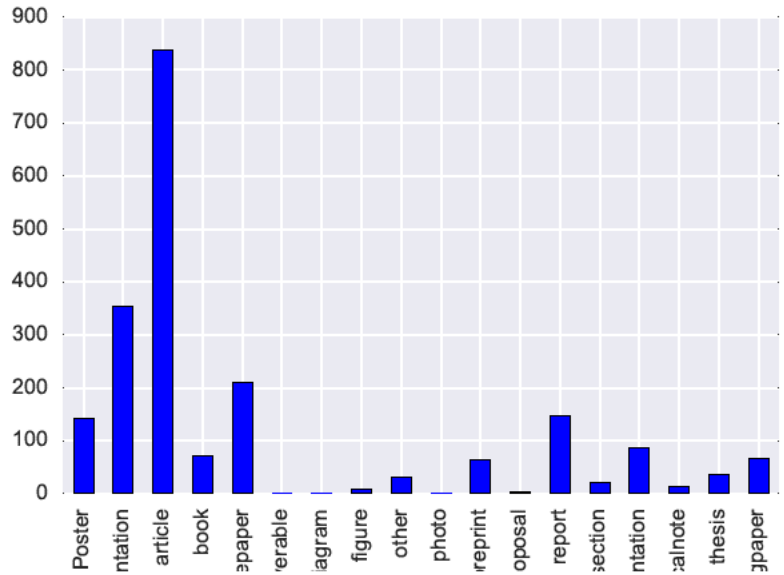
DATA COLLECTION

- OAI-PMH endpoint user-zenodo, and recommended format, oai datacite3
- Additional scraping for data not included in the OAI export.
- 3,828 records at the time of the data extraction.
- DOIs in around 91% of the cases.
- Language in metadata is only present in 15% of the resources, of which 97% corresponds to English.

A DATA REPOSITORY?

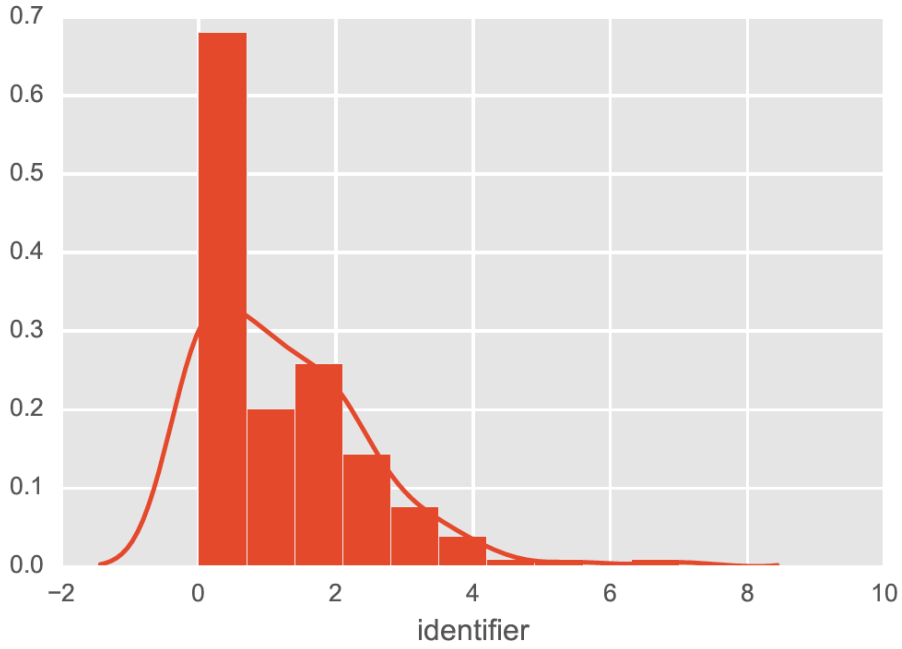
Dominant use is still for publications.





ANALYSIS OF COMMUNITIES

- Skewed distribution in collections.
- Mix of motives for communities:
  - Journals (F1000Research)
  - Projects (FP7 DEVOTES)
  - Topical (Drosophila)



COMMUNITIES AND QUALITY CONTROL

- Non-applicable for some of the communities. e.g.:
  - Journals (that provide their own, external quality controls).
  - Projects (scope is implicit to ptoject reach)
- Unclear for other communities.
- Lack of possibilities of judging selection criteria in many cases
  - This calls for a more informed approach to creating communities, now it is 100% unrestricted.

CONCLUSIONS AND OUTLOOK

- Relatively small growth, compared with the intended audience.
- Not primarily used for data depositing.
- Community curation still in inception, better information needed.
- Re-thinking on the scope of communities:
  - MERLOT as a more structured community approach.
  - Still useful for funders and national agencies (if they eventually use it)