The future role of repositories for OA discovery

OAI 11 Workshop on Innovations in Scholarly Communication
“Open Science – its impact and potential as a driver for radical change”, University of Geneva, June 19th-21st 2019
Dirk Pieper, Bielefeld UL
Agenda

1. Introduction
2. BASE
3. DOI pretest
4. Conclusions
Introduction

- Metadata quality has improved in repositories thanks to initiatives like COAR, DINI, OpenAIRE and others
- Need for Plan S compliance will initiate further improvements
- But what is the relevance of IRs for OA discovery?
Crossref has become the primary source for journal articles

The number of Gold OA journal articles has increased because of financing APCs

Do we find those articles on IRs?
Introduction

- Analysing frequency and distribution of DOIs in BASE index
- Pretest: Analysing DOIs of OpenAPC data set (OA journal articles, which have been financed by funders, academic institutions, libraries, …) in BASE index
Agenda

1. Introduction
2. BASE
3. DOI pretest
4. Conclusions
BASE development since 2004
Global heatmap of documents on repositories
BASE background development for data analysis
Example 1: Documents per repository type in Germany

Repository Type Development (Number of Documents) since 2007

- IRs
- Theses
- Publications
- Journals
- Digital Collections
- other

Chart showing the development of documents by repository type from 2007 to 2017.
Example 2: Metadata quality in German repositories (doc lang, doc OA status)
Example 3: Metadata quality in German repositories (doc license, publication type)
First insights

- About 27% of all records in BASE are journal articles (exact number cannot be determined because of lacking metadata)
- Share differs from country to country
- Providing comprehensive and valid information on OA status and license information is still a challenge
1. Introduction
2. BASE
3. DOI pretest
4. Conclusions
Basic questions

- How many DOIs are in the BASE index?
- How many DOIs of the OpenAPC data set can we find there?
- How many OpenAPCs-DOIs do we find on IRs?
First results

- Out of 149,8 million documents and 7,175 sources there are 43,051,773 DOIs indexed in BASE
- Top ten: Data Cite (15,1 mill.), PubMed Central (3,3 mill.), DOAJ (2, 8 mill.), HighWire Press (2,4 mill.), Elsevier SD (1,3 mill., only OA articles via Crossref), ), Zenodo (1 mill.), ArXiv (822 k), ETH e-periodica, HAL, Dara
Distribution of DOIs on repository types in BASE
First results for OpenAPC data set

- Out of 78,264 OpenAPC-DOIs 67,614 are indexed in BASE
- Top ten: PubMed Central (42,3 k), DOAJ (24,8 k.), DataCite (17,8 k), Elsevier SD (11 k), PLOS (8,1 k), Frontiers (5,1 k), MPG Pure (3,9 k), Oxford Univ. Research Archive (3,7 k), Imperial College (3,1 k), CCSD Articles HAL (2,9 k)
Distribution of OpenAPC-DOIs on repository types in BASE
Agenda

1. Introduction
2. BASE
3. DOI pretest
4. Conclusions
Conclusions

- Nearly the half of OpenAPC-DOI sample can be found on IRs
- PMC, DataCite and DOAJ have the greatest shares on the OpenAPC-DOI sample
- The first academic institutions (MPG, Oxford, Imperial College) are becoming relevant for the discovery of Gold OA journal articles
Conclusions

- More academic institutions have to make stronger efforts to list Gold OA journal articles on their repositories / CRIS
- Need for workflows to get article data from transformative agreements into repositories
Thank you!

www.base-search.net

Follow us: @BASEsearch