Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

All about that BASE

presented at PIDapalooza 2018 in Girona

Christian Pietsch https://orcid.org/0000-0001-8778-1273

Bielefeld University Library, Germany

with material borrowed from Friedrich Summann https://orcid.org/0000-0002-6297-3348

January 23, 2018

Overview

We are here

BASE: Bielefeld Academic Search Engine

BASE harvests, aggregates, enriches and exposes OAI metadata, mainly from institutional respositories (IR), subject repositories and open access (OA) journals.
BASE is . . .

- the second largest academic search engine by index size (behind GScholar) [Khabsa & Giles, 2014]
- an OAI-PMH aggregator, metadata mirror of the IR/OA landscape, IR checker (both automatic at `http://oval.base-search.net` and in person via e-mail)
- an OAI-PMH service provider
- a search API provider
- on Twitter: @BASEsearch

Scope

The BASE scope:

- OA Repositories world-wide
- Academically valuable contents
- Focus on Institutional Repositories
- Aggregators (RePEc)
- Subject Repositories (arXiv, CiteSeerX, PubMed etc.)
- Electronic journals (e.g. OA journals from CrossRef)
- Digital collections
- Dataset repositories

History

- 2001 Building a search engine based follow-up to a metasearch system
- 2004 Official Start (using FAST Data Search)
- 2005 Functionalities search history, sorting added
- 2006 Starting participation in EU projects (DRIVER), Search API
- 2007 Introducing multilingual search
- 2008 Index > 10 million records
- 2009–2011 automatic DDC classification of OAI-DC metadata (ta, DFG)
- 2011 Switch to open source platform (Lucene/Solr, VuFind)
- 2012 OAI Interface, data delivery of subject sections
- 2014 OA search rank boosting (on by default)
- 2016 Index > 100 million records
- 2017/05 ORCID claiming in BASE (funded by DFG project ORCID DE)
- 2017/09 ORCID Search & Link wizard integrates BASE as a claiming service

The Present

Through BASE you can search:

- Number of documents: 122,231,419 (ca. 70% open access; 42.5% confirmed OA)
- Number of content sources: 6,103
- Number of source countries: 127

source: `https://www.base-search.net/about/en/about_sources.php?menu=2`

Three ways to access BASE

- web interface: researchers, information specialists
- search API: many website that do not have their own search index, e.g. meta search engines MetaGer, etools.ch, and Searx.
- OAI-PMH API: German National Library (DNB), Discovery Services (EBSCO, Karlsruher Virtueller Katalog), alternative DOI resolvers (DOAI, and, until recently, oaDOI) and other innovative Open Access services such as Dissemin and Open Knowledge Maps.

We are here

OAI-PMH

Open Archives Initiative – Protocol for Metadata Harvesting
since ca. 2000
pros:

- simple, time-tested, widely spread and stable (specification)

cons:

- technical implementation
- misconfigurations
- slow for large repositories
- HTTP-based but no quite RESTful
- not enough transparency regarding availability

Alternatives to OAI-PMH

- Web harvesting (unstructured data unless Semantic Web)
- ResourceSync `https://www.openarchives.org/rs/toc`
- Linked Open Data?

We are here

Dublin Core

Dublin Core (DC) is the only metadata format all OAI repositories must support.
It is simple, flexible and somewhat vague.

Issue at hand: identifiers

- OAI Identifier: often misconfigured, e.g., hostname not set, i.e. <dc:identifier>`http://localhost/9160/`</dc:identifier> instead of <dc:identifier>`http://eprints.an.edu/9160/`</dc:identifier>
- Handle/DOI/ISSN/ISBN/URN/PMCID
- Author/Organization/Funder IDs

We are here

DOIs in the wild

Where to look for DOIs:

- We found DOIs in the following Dublin Core elements: identifier, relation, source, description, rights.
- Those in dc:relation are likely to refer to other documents. Those in dc:description can, too.

How to identify DOIs:

- Not trivial: The suffix can be made up of any Unicode characters, so where does it end?
- Real specimen: 10.1002/(SICI)1521-3765(19990604)5:6<1728::AID-CHEM1728>3.3.CO;2-M
- Most pathologic case: comma as last character of a DOI – or not?

We are here

ORCID iDs in Dublin Core

- source: IFPRI ca. 2017

<dc:creator>http://orcid.org/0000-0002-1179-0189 Ball, Anna</dc:creator>
<dc:creator>http://orcid.org/0000-0001-9794-2026 Van den Bold, Mara;
http://orcid.org/0000-0002-8501-5943 Gillespie, Stuart;
http://orcid.org/0000-0001-5988-2894 Menon, Purnima</dc:creator>

- other creative solutions: (search in BASE)

ORCID-DE and DINI intend to publish joint recommendations this year.

Alternatives to Dublin Core (1/2)

Qualified Dublin Core

Dublin Core can be extended. However, none of the extensions is popular. QDC is so rare in the wild, it is not worth looking at.

MARC

source: Hindawi in 2015

<datafield tag=“700” ind1=" " ind2=" “> <subfield code=”a“>Awais, Muhammad</subfield> <subfield code=”u“>Department of Industrial Engineering, King Abdulaziz University, P.O. Box 344, Rabigh 21911, Saudi Arabia</subfield> <subfield code=”j“>ORCID-0000-0002-5229-8780</subfield> </datafield>

Alternatives to Dublin Core (2/2)

MODS

source: `https://pub.uni-bielefeld.de/oai?verb=GetRecord&identifier=2907546&metadataPrefix=mods`

<mods version="3.3">

...

<name type="personal">
<namePart type="given">Amelie</namePart>
<namePart type="family">Bäcker</namePart>
<identifier type="unibi">53852847</identifier>
<description xsi:type="identifierDefinition"
type="orcid">0000-0001-6015-2063</description>
</name>

ORCID in the wild

Repositories we found ORCID iDs in used the following software:

- OJS
- Diva
- DSpace
- DSpace XOAI
- DigitalCommons / BEPress
- Invenio
- ContentDM
- LibreCat
- (maybe more)

ORCID handling

- Detecting ORCIDs in metadata
- Publication claiming
- Feeding to ORCID publication list
- Feeding back to repositories?

ORCID claiming in BASE

Goals:

- assigning more ORCID iDs to records from the OAI document space
- one-stop shop for claiming ORCID iDs in >6000 manually curated sources
- re-use of thusly enriched records by third parties – e.g. ORCID

Two entry points:

- integrated in BASE's web front-end (VuFind, PHP, MySQL)
- integrated in ORCID.org's Search & Link wizard

We are here

ORCID claiming

https://orcid.org/blog/2017/09/06/announcing-base-orcid-search-link-wizard

We are here

Literature

- Vierkant, P. (2017). Announcing the BASE ORCID Search & Link Wizard. ORCID blog. https://orcid.org/blog/2017/09/06/announcing-base-orcid-search-link-wizard
- Abad-García, M.-F., González-Teruel, A., & González-Llinares, J. (2017). Effectiveness of OpenAIRE, BASE, Recolecta, and Google Scholar at finding Spanish articles in repositories. Journal of the Association for Information Science and Technology. https://doi.org/10.1002/asi.23975
- Summann, F. (2017, November). Aktuelle Entwicklungen im globalen Repository und Open-Access-Netzwerk: Relevante Themenbereiche aus der technischen Sicht eines Service Providers. Zenodo. https://doi.org/10.5281/zenodo.1048845
- BASE (Bielefeld Academic Search Engine). Eine Suchmaschinenlösung zur Indexierung wissenschaftlicher Metadaten. In: Datenbank-Spektrum. 2017. https://doi.org/10.1007/s13222-017-0246-9
- Summann, F. (2016). Die Verwendung von Autorenidentifikatoren in wissenschaftlichen Repositorien : Ansätze, konkrete Umsetzungen und Herausforderungen. Presented at the 105. Deutscher Bibliothekartag in Leipzig 2016 = 6. Bibliothekskongress, Leipzig.
- Fenner, M. et al. (2016). Autorenidentifikation für wissenschaftliche Publikationen. Bericht über den Workshop der DINI-AG Elektronisches Publizieren auf dem 6. Bibliothekskongress. o-bib, 3(4), 286-293. doi:10.5282/o-bib/2016H4S286-293
- McMurry., J. et al. (2015). 10 Simple rules for design, provision, and reuse of identifiers for web-based life science data. https://doi.org/10.5281/zenodo.31765
- Summann, F. (2016). Establishing Open Access Services in the Global Repository NetworkExperiences and Challenges from the Service Provider Perspective (BASE). Presented at the 2016 Chinese Institutional Repository Conference, Chongqing.
- Summann, F., & Shearer, K. (2015). COAR Roadmap Future Directions for Repository Interoperability. Göttingen: COAR (Confederation of Open Access Repositories).
- Pieper, D., & Summann, F. (2015). 10 years of “Bielefeld Academic Search Engine” (BASE): Looking at the past and future of the world wide repository landscape from a service providers perspective. Presented at the OR2015. 10th International Conference on Open Repositories, Indianapolis. http://ir.las.ac.cn/handle/12502/8749
- Khabsa, M., & Giles, C.L. (2014). The Number of Scholarly Documents on the Public Web. PLOS ONE 9 (5) https://doi.org/10.1371/journal.pone.0093949
- Pieper, D., & Summann, F. (2006). Bielefeld Academic Search Engine (BASE) An end-user oriented institutional repository search service. Library Hi Tech, 24(4), 614-619. https://doi.org/10.1108/07378830610715473