Metadata on books in BASE

Kolloquium Wissensinfrastruktur, 28.10.2022

Dirk Pieper, Bielefeld UL

BASE = Bielefeld Academic Search Engine

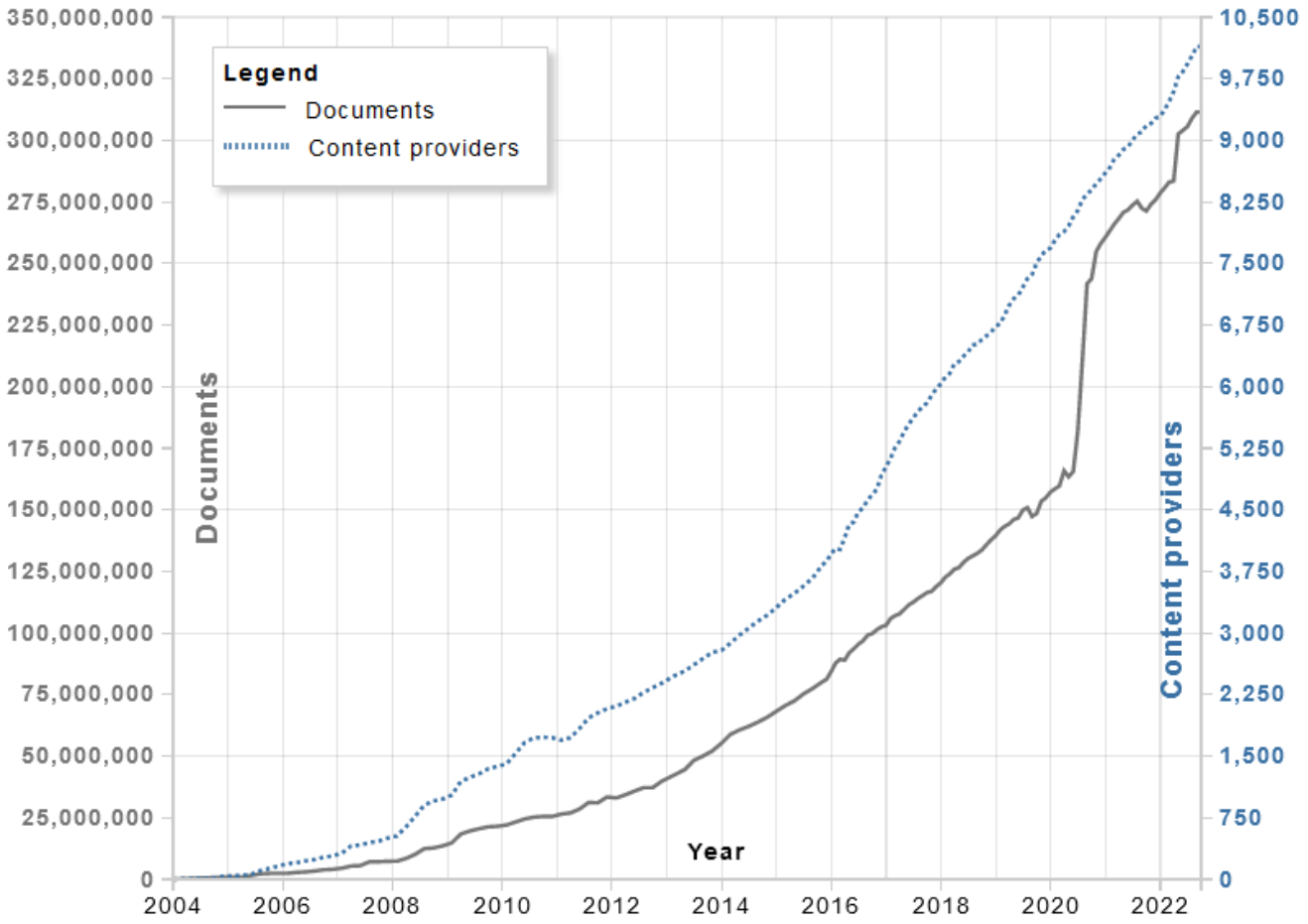
www.base-search.net

Harvesting, aggregation, normalisation, enrichment
of (OAI-DC) metadata

Uniform index and search interface for >10,000 repositories and additional data sources with academic content

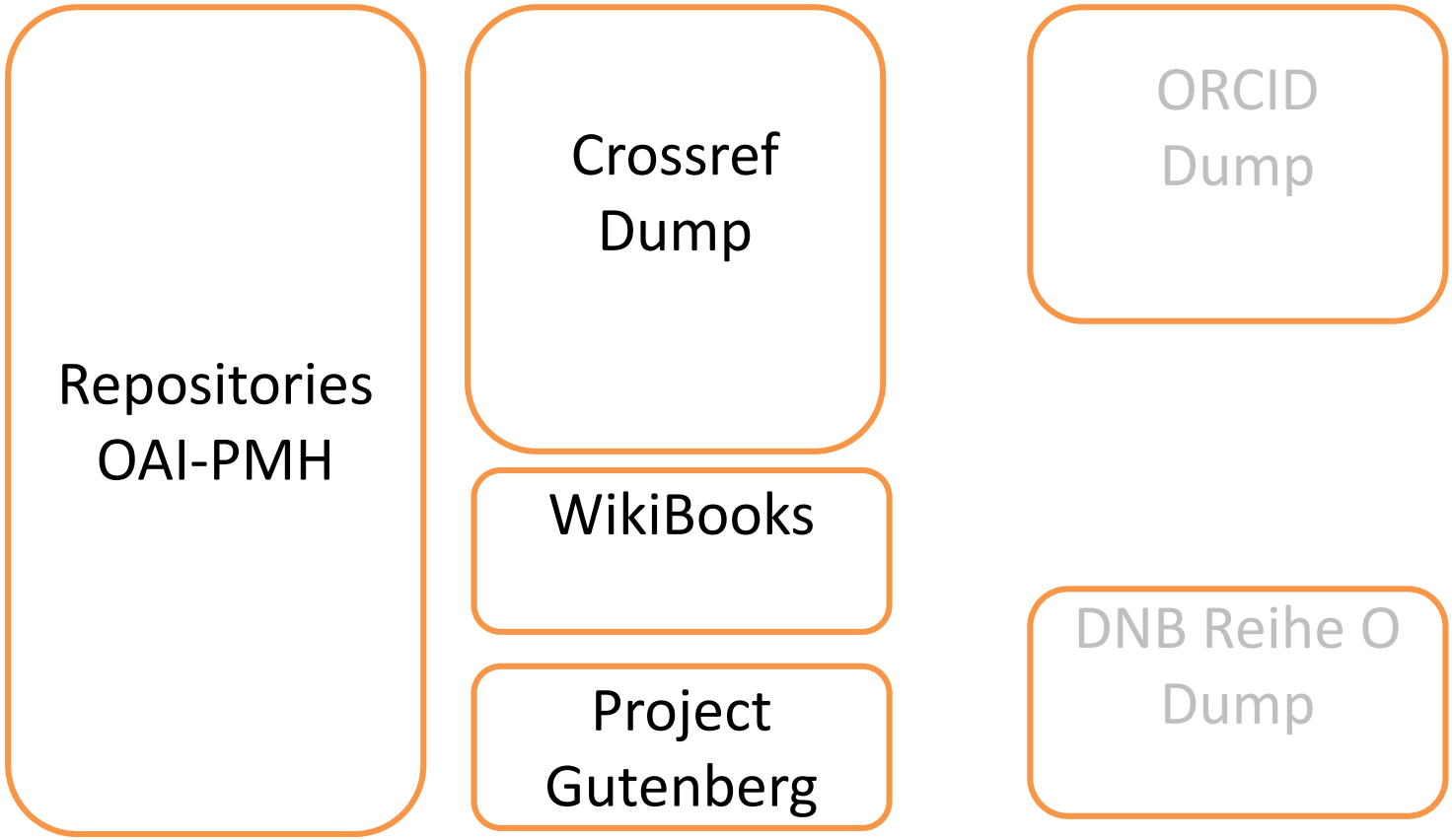
BASE

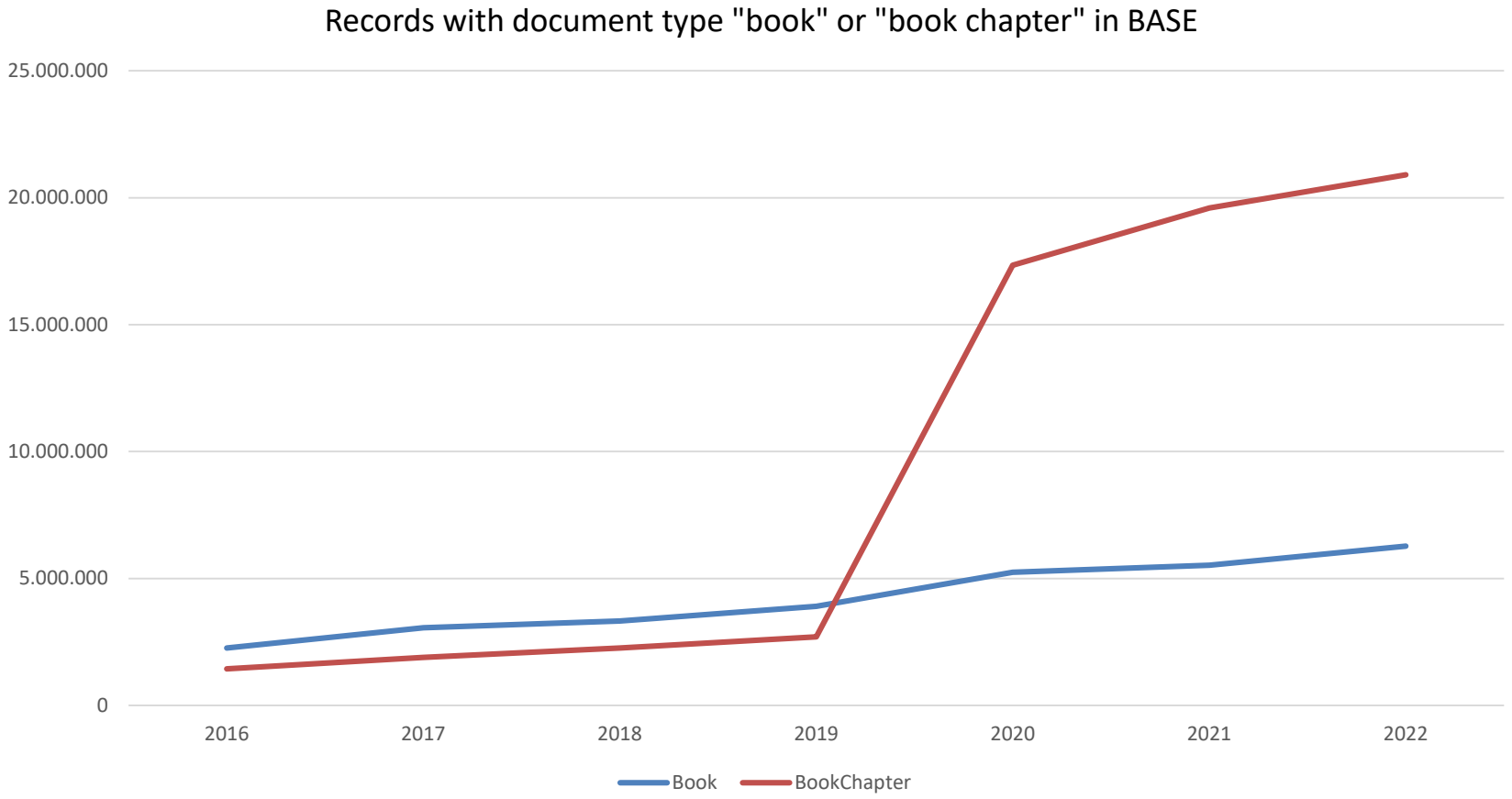
Development of the number of indexed content providers and documents in BASE since September 2004.



BASE

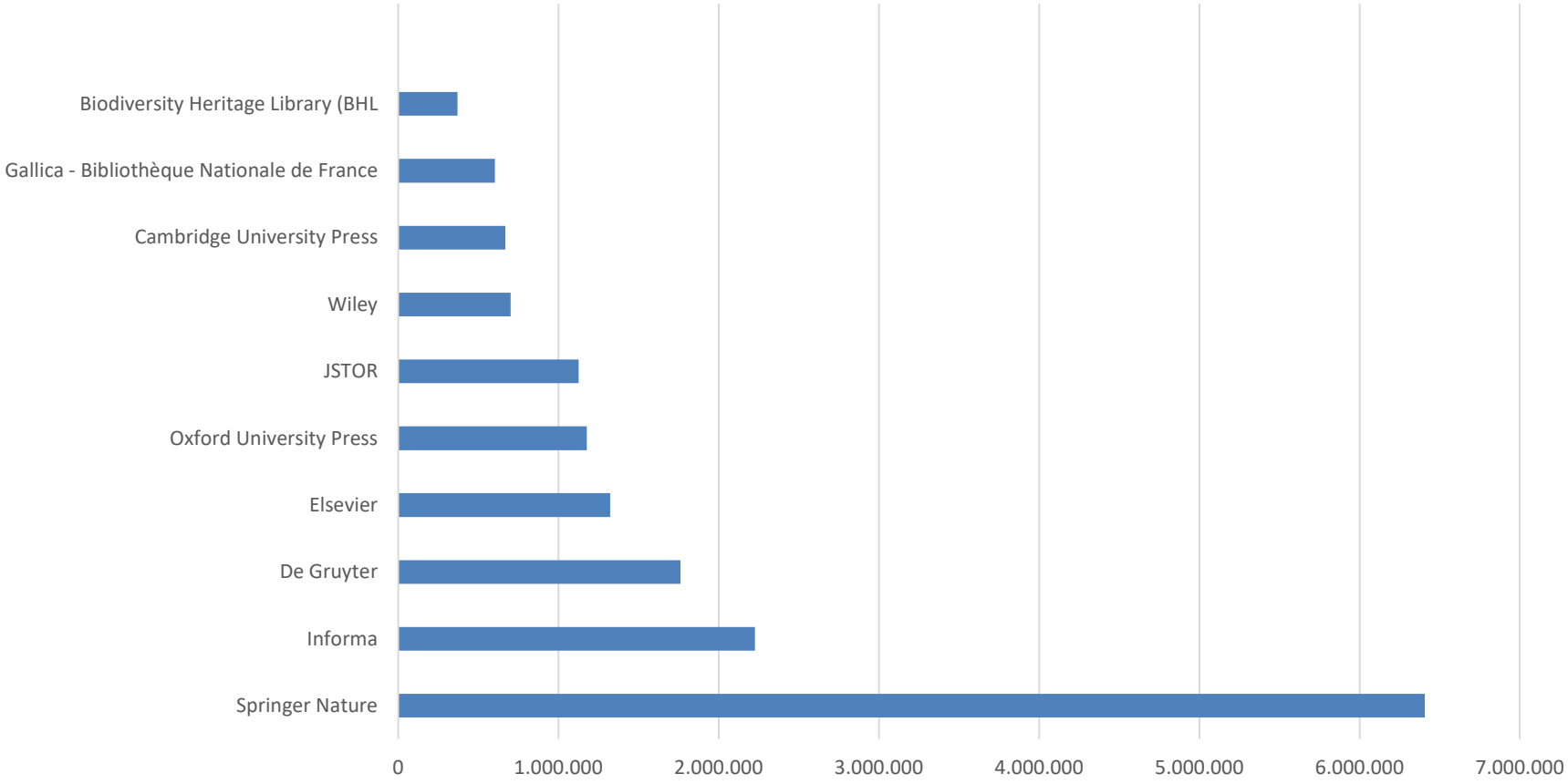
Data sources




Records with document type "book" or "book chapter" in BASE
25.000.000
20.000.000
15.000.000
10.000.000
5.000.000
0
2016
2017
2018
2019
2020
2021
2022
Book
BookChapter

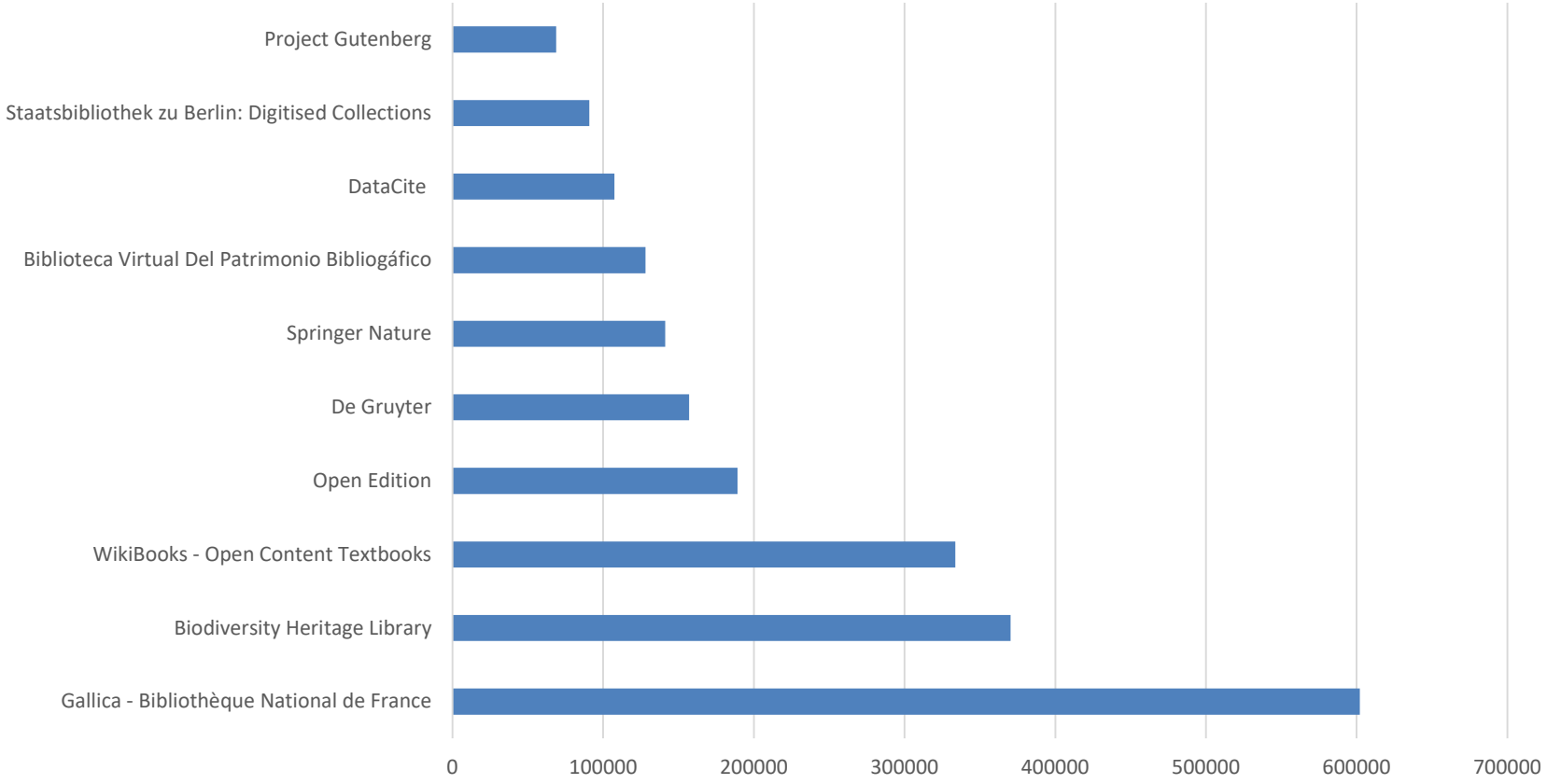
BASE

Top 10 data providers for books and book chapters in BASE 2022-09-06

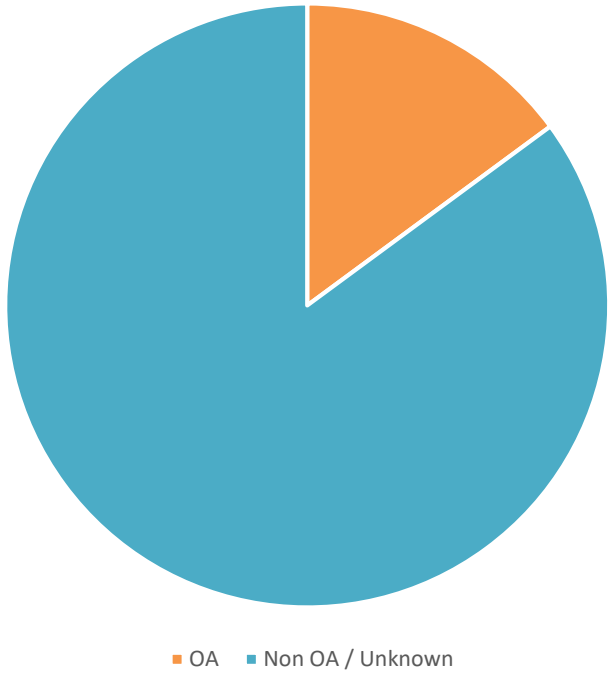


BASE

Top 10 data providers for OA books and book chapters in BASE 2022-09-06



Distribution of OA and Non OA / Unknown books and book chapters in BASE
2022-09-06



Guidelines for resource / document types

OpenAIRE Guidelines for institutional and thematic repository managers 4.0

COAR controlled vocabularies for repositories. Resource types

DINI: Gemeinsames Vokabular für Publikations- und Dokumenttypen Version 2.0

COAR Controlled Vocabularies for Repositories

Resource Types: book

Definition

A non-serial publication that is complete in one volume or a designated finite number of volumes. [Source: Adapted from http://purl.org/eprint/type/Book]

URI

http://purl.org/coar/resource_type/c_2f33

Preferred Labels

- Buch (Deutsch)
- boek (Nederlands)
- book (English)
- kirja (Suomi)
- kitap (Türkçe)
- kniha (Čeština)
- libro (Español)
- libro (Italiano)
- livro (Português)
- llibre (Català)
- monografija (Slovenščina)
- ouvrage (Français)
- книга (Русский)
- كتاب (العربية)
- 书 (中文)
- 図書 (日本語)

Alternate Labels

- Monographie (Deutsch)
- audiobook (English)
- books (English)
- e-book (English)
- knjiga (Slovenščina)
- livre (Français)
- monografia (Português)
- monografía (Español)
- monografia (Català)
- monografia (Italiano)
- monografie (Čeština)
- monografía (Español)
- monograph (English)
- monographie (Français)
- obra monografica (Español)
- obra monográfica (Español)
- افروده (العربية)
- 专著 (中文)
- 图书 (中文)
- 圖書 (中文)
- 專著 (中文)
- 書 (中文)

Narrower Concepts

- book part

Broader Concepts

- text

Related terms (external)

Broad Match: http://purl.org/dc/dcmitype/Text
Exact Match: http://purl.org/eprint/type/Book
Exact Match: http://purl.org/spar/fabio/Book
Narrow Match: http://purl.org/eprint/type/BookReview
Related Match: https://schema.org/Book

Challenges

Different guidelines, different rules for cataloging

Many languages

Missing information about OA status or license

Metadata provided via machine readable interfaces sometimes differ from what you can see on web interfaces

BASE approach

Analysing used vocabularies and terms in data sources

Mapping the most used terms (more than 1,000 incidents) to an own hierarchical meta vocabulary

Creating an own index field for normalised document types (typenorm)

Current top 10 findings in dc:type

String in <dc:type>	Number of incidents	Number of data sources
info:eu-repo/semantics/publishedVersion	19762970	5651
printed serial	14593295	3
info:eu-repo/semantics/other	5590188	686
Other	1862512	1506
ARTIGO DE PERIODICO	1310311	2
Collection	1231913	202
PhysicalObject	1090644	42
fénykép	1055534	4
PGRFA Material	987413	3
TRABALHO DE EVENTO-RESUMO	983451	2

Examples for „book“ and „book chapter“ findings in dc:type

String in <dc:type>	Number of incidents	Number of data sources
Chapter in Book, Report or Conference volume	483652	1
eReadings Item	371313	1
model:monograph	214420	1
Books and book chapters	76129	1
Book, Whole	73119	5
Digital document	71484	1
Document électronique	71484	1
Chapter, Part Of Book	68346	1
Original research	68129	5
issue	64751	23
Book/Report	59146	1
Publication	51871	22
CHAPTER	51537	3
Documento Completo	50289	2
book part	49084	19
BOOK	27582	10
yearbook	26902	11
eBook	24552	8
dokument elektroniczny	24131	16
Texte	21616	10

Example: filter for books

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=base_dc&set=typenorm:11

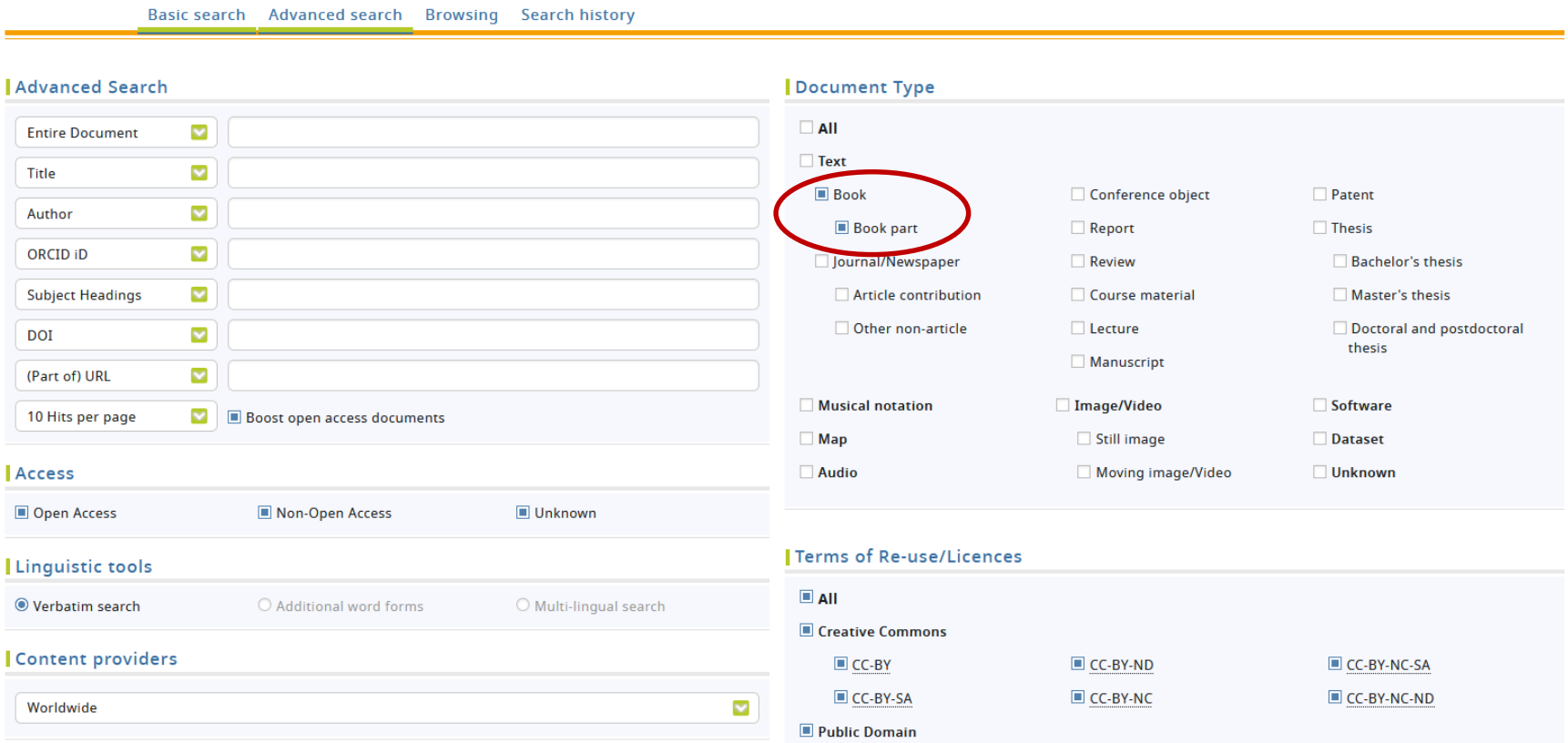
Document Types

As the categorization of document types is very heterogenous across repositories, BASE normalizes them by mapping types into consistent categories which are identified by a numerical code. The table below lists normalized document types, which can be queried by using the field `typenorm`.

Numeric codes for normalized document types.

document type	numeric code
text	1
book	11
book part	111
journal/newspaper	12
article in journal/newspaper	121
other non-article part of journal/newspaper	122
conference object	13
report	14
review	15
course material	16
lecture	17
thesis	18
bachelor thesis	181
master thesis	182
doctoral or postdoctoral thesis	183
manuscript	19
patent	1A
musical notation	2
map	3
audio	4
image or video	5
still image	51
moving image (video)	52
software	6
dataset	7
other/unknown material	F

Example advanced search



Conclusions

BASE vocabulary (not only for books) and normalisation workflows to be updated

Mapping the most used terms will probably work for about 80% of the records in BASE

Guidelines help!

BASE

… but somehow it feels like this …

(Foto “Used Books 02” von “linmtheu” @ Flickr)

Outlook: PALOMERA

PALOMERA - Policy Alignment of Open access Monographs in the European Research Area

Activity: HORIZON-WIDERA-2022-ERA-01-42, 01/2023-12/2024

16 partners (coordinators: OPERAS & OAPEN)

Initial observation: books are only rarely mandated to be published OA by research funders and institutions

Outlook: PALOMERA

PALOMERA will investigate the reasons for this situation across geographies, languages, economies, and disciplines within the European Research Area (ERA)

PALOMERA will provide actionable recommendations and concrete resources to support and coordinate aligned funder and institutional policies for OA books, with the overall objective of speeding up the transition to open access for books to further promote open science

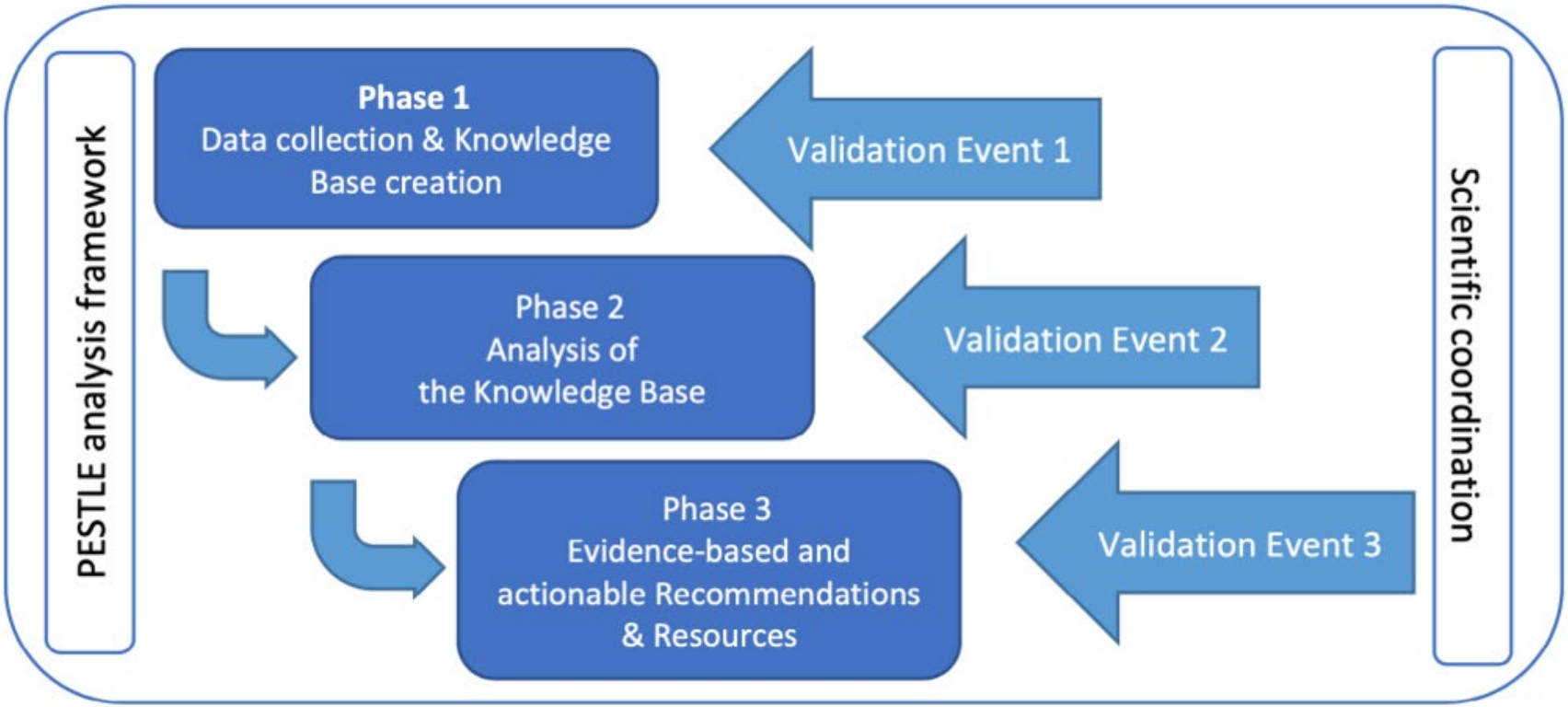


Fig.1 PALOMERA methodology

Outlook: PALOMERA

UNIBI: contribute to data collection & knowledge base creation with data from BASE and OpenAPC

Thank you!

@BASEsearch