Universität Bielefeld
INFORMATION.plus!
Universitätsbibliothek

BASE
Bielefeld Academic Search Engine

Data Science Tools zur Anaylse des globalen Repository-Netzwerks

Dirk Pieper – Friedrich Summann
Bielefeld University Library

Kolloquium Wissensinfrastruktur, 31.5.2019

Abstract

- Der Vortrag skizziert technischen Möglichkeiten, das globale Publikationsnetzwerk und seine Entwicklung zu analysieren und zu beschreiben.
  Seit 2004 fallen zahlreiche Metadaten im BASE-Umfeld an: insbesondere Publikationsmetadaten, Metadaten zu den Quellen, Indexdaten und damit verbunden Informationen, die zu den Quellen die Verteilung bei Publikationsstyp, Publikationsdatum, Sprache des Dokuments, OA Status und Lizenzinformationen liefern.
  Skizziert wird das Konzept eines Monitoring-Systems inklusive des Datenflusses und der zu implementierenden Schnittstelle.
  An einigen konkreten Beispielen wird prototypisch gezeigt, wie mit geeigneten Tools eine visuelle Umsetzung implementiert wurde und erweitert erden kann.

Overview

- Introduction
- The Data Basis and its Pre-Processing
- The API Layer
- Concrete Examples
- Summary

Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

BASE
Bielefeld Academic Search Engine

BASE – OA Search Engine

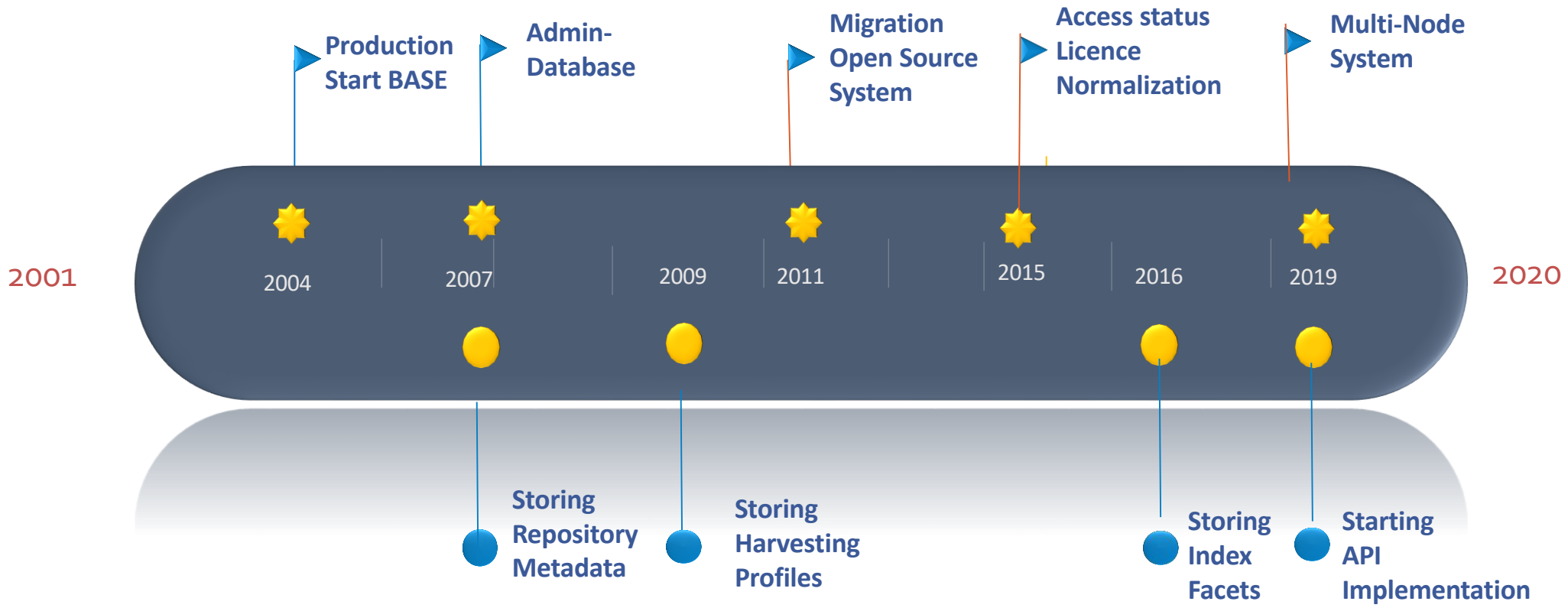
- 7115 Repositories included (mostly via OAI-PMH, some via Crossref, few via propietary …)
- From 124 Countries world-wide
- Ca. 146 Mill. Documents/Objects
- Ca. 70 % Open Accessible
- Internal Scheme: extended Dublin Core Format
- Inst. Repositories, Journal Platforms, Research Data, Digital Collections, Multimedia

Overview

- Introduction
- **The Data Basis and its Pre-Processing**
- The API Layer
- Concrete Examples
- Summary

BASE – Collecting Metadata

- Repository Metadata –
  - Common resource description (country, platform, repository type, OA status, geocodes, name)
- Harvesting Metadata
  - Number of Documents, Number of OA Documents
  - Technical figures (Metadata Formats, Deleting Strategy)
- Index Data
  - - Number of repositories, number of documents
  - - the Index Facets (date, language, publication type, Access status, licences


Production Start BASE
Admin-Database
Migration Open Source System
Access status Licence Normalization
Multi-Node System
2001
2004
2007
2009
2011
2015
2016
2019
2020
Storing Repository Metadata
Storing Harvesting Profiles
Storing Index Facets
Starting API Implementation

Universität Bielefeld | INFORMATION.plus! Universitätsbibliothek

Input Form BASE Admin Database

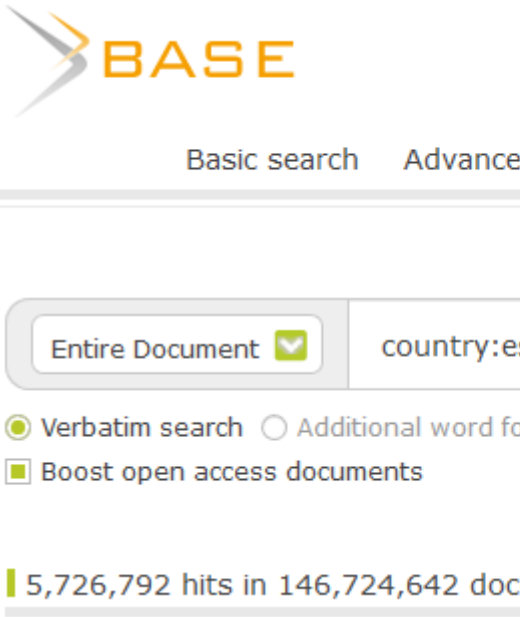
BASE Bielefeld Academic Search Engine

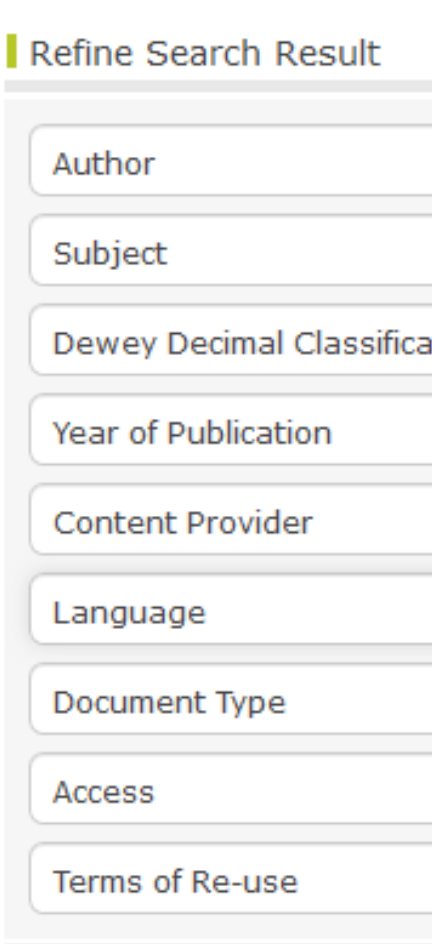
Einstellungsdatum*:	20.03.2006 Datum auswählen
Bearbeiter:	Friedrich Summann
Quellenname:	PUB - Publications at Bielefeld University
Harvestername (OAI)*:	UBBielefeld-PUB
System:	LibreCat Sonstiges (hat Vorrang!):
Klassifikation:	DDC
Herkunftsland (Domain / TLD):	de
Bundesstaat (z.B. denw):	denw
DINI-Zertifikat:	2016
URL (Applikation-Startseite):	http://pub.uni-bielefeld.de/
OAI Basic URL:	http://pub.uni-bielefeld.de/oai
Repository Id:	od_2294
OA:	Unbekannt
OA Set:	open_access
Breitengrad (Finder):	52.037600
Längengrad:	8.493600
Adresse:	---------- From OpenDOAR: Universität Bielefeld 25, D-33615 Bielefeld Germany ----------
Kontakt:	---------- From OpenDOAR: Friedrich Summann Administrator friedrich.summann@uni-bielefeld.de ----------
Typ:	Hochschulpublikationen

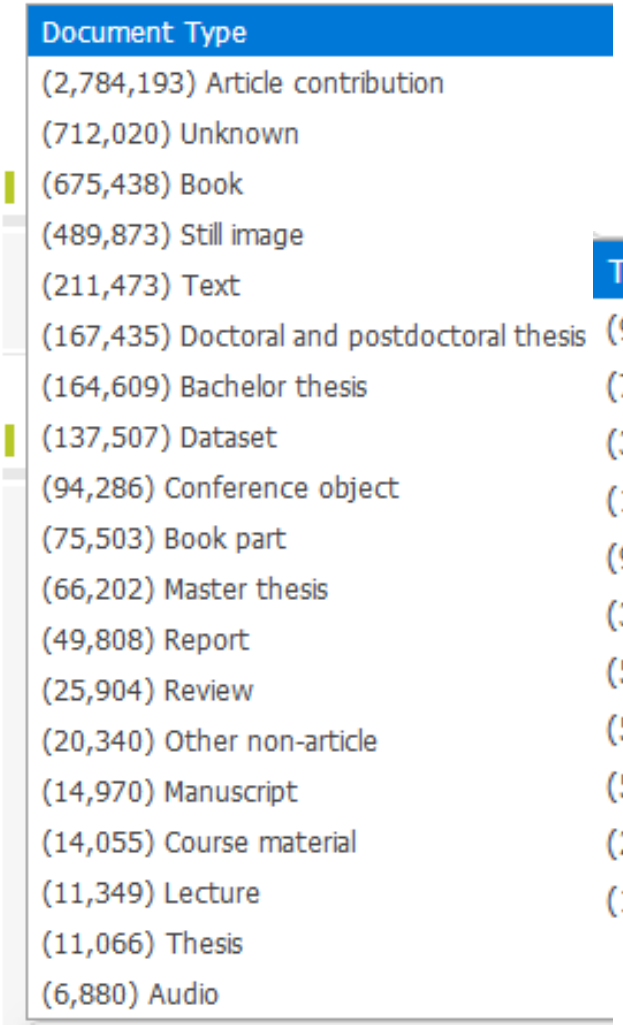
Universität Bielefeld | Universitätsbibliothek

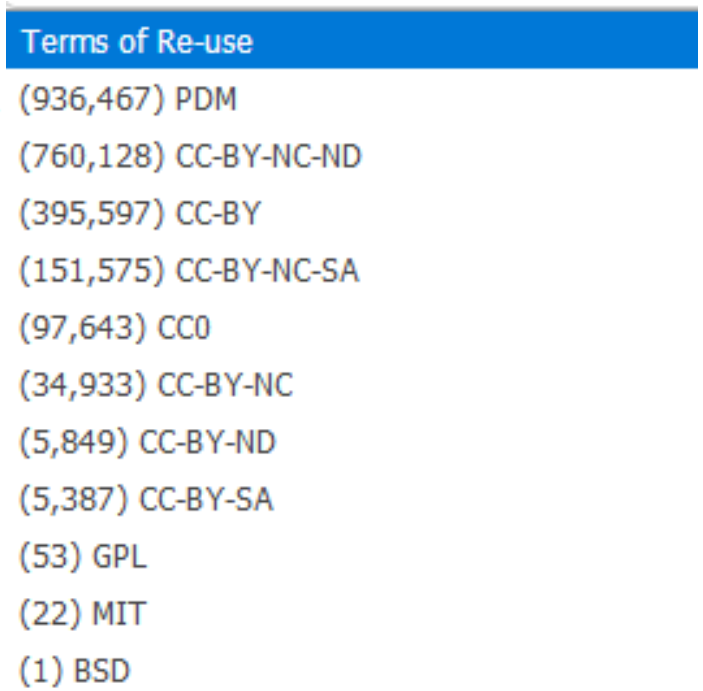
BASE Index Facets

BASE Bielefeld Academic Search Engine







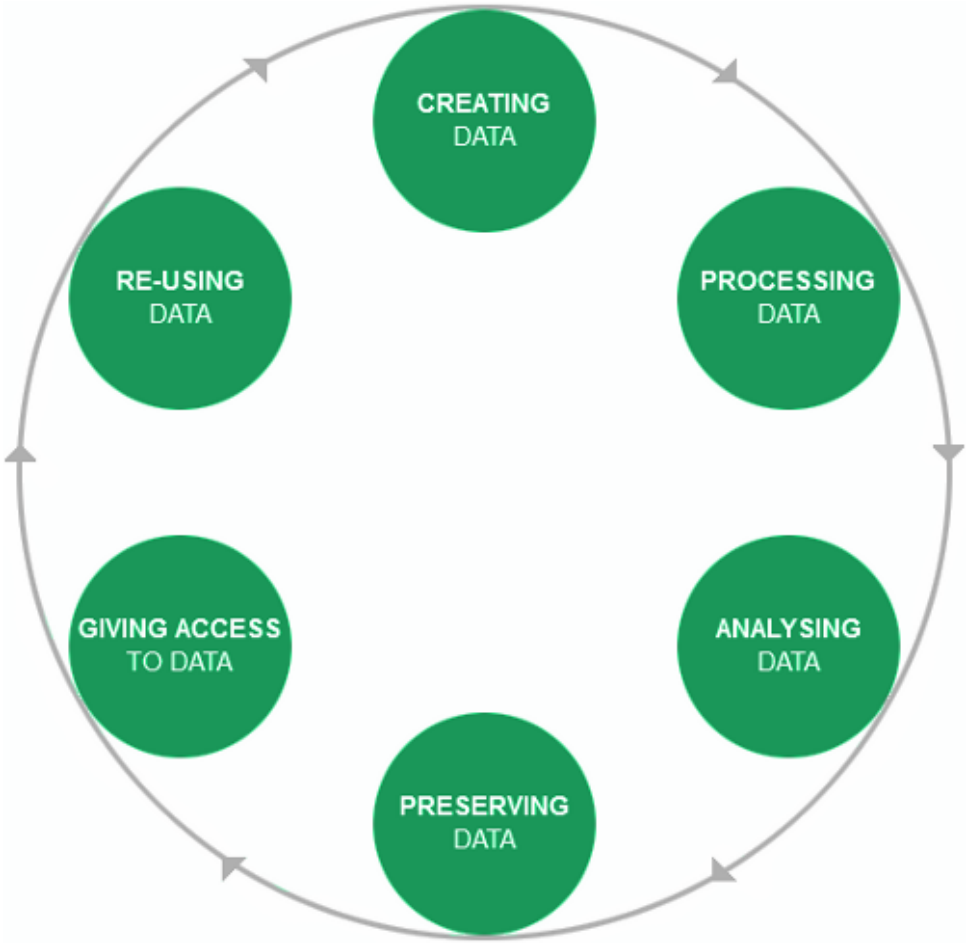


Data Life Cycle

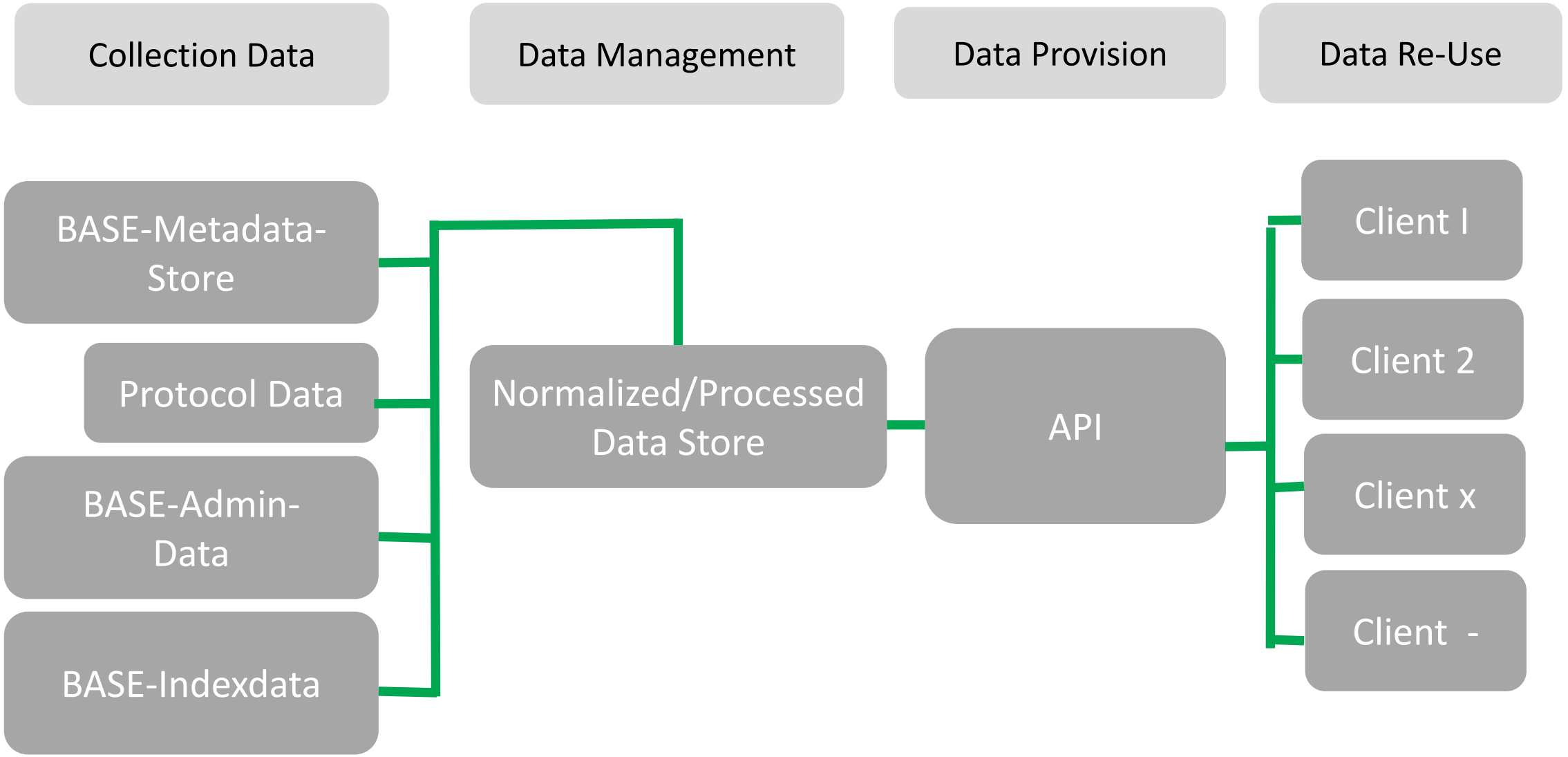
Bielefeld Center for Data Science

Konzeptioneller Rahmen	Einführung in Daten
Datensammlung	Datenerschließung und -sammlung
	Evaluierung und Sicherstellung der Qualität von Datenquellen
Datenmanagement	Datenorganisation
	Datenmanipulation
	Datenkonvertierung
	Erzeugung und Verwendung von Metadaten
	Datenheilung, -sicherheit und -wiederverwendung
	Datenaufbewahrung
Datenevaluation	Datenwerkzeuge
	Grundlegende Datenanalyse
	Dateninterpretation (Datenverständnis)
	Nutzung von Daten zur Identifizierung von Problemen
	Datenvisualisierung
	Datenpräsentation (verbal)
	Datengetriebene Entscheidungsfindung
Datenanwendung	Kritisches Denken
	Datenkultur
	Datenethik
	Datenzitierung
	Datenteilung
	Evaluieren von Entscheidungen basierend auf Daten

[UK Data Archive research data lifecycle model]



Data Infrastructure BASE Monitoring



External data (to compute correlations)

- Number of Scientists, Number of Students per country
- Size of Population per Country
- Publication Output per Country
- Funding Budgets
- Bibliometric figures

Level of Aggregation

- Repository
- *Institutional (more than 1 resource per institution)*
- Country Aggregation
- Repository Type Aggregation
- Continent Aggregation
- Global Aggregation

Overview

- Introduction
- The Data Basis and ist Pre-Processing
- **The API Layer**
- Concrete Examples
- Summary

API (to be Developed)

- Restful API (via http)
- Query Language Definition
- Format Definition Responses
- Export Formats (xml,json,csv)
- Documentation

Flexible search query-like with Multi-Level - Support

Query Examples:

- List number of publications with publication type ‚thesis‘ per country
- List percentage of OA documents per repository/country/continent
- Showing the Increase of Research Data Objects per country per year
- List languages of documents with type books

Timeline support

Overview

- Introduction
- The Data Basis and its Pre-Processing
- The API Layer
- **Concrete Examples**
- Summary

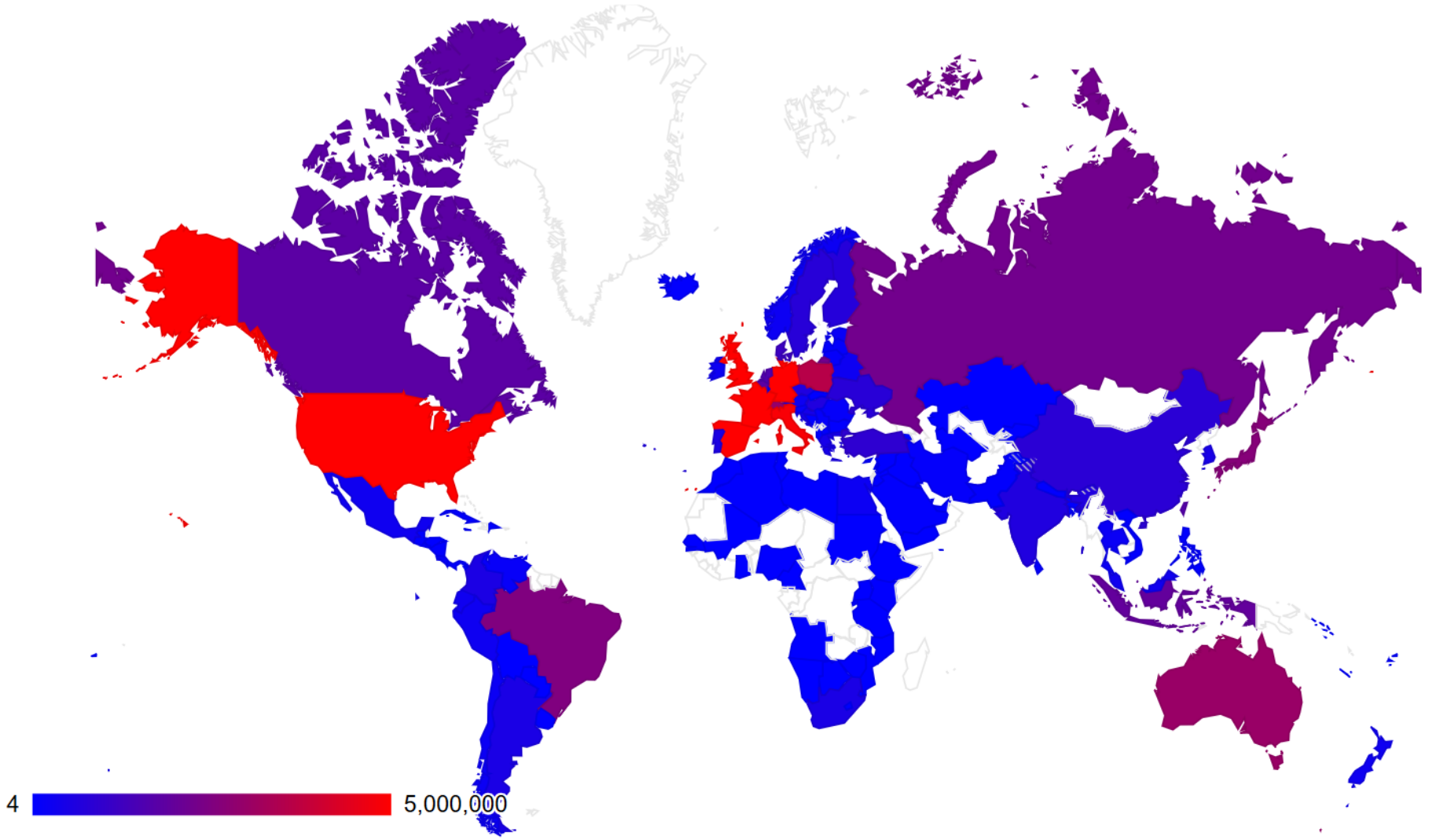
Existing Tools for Visualisation

- World Maps
  - Google Charts
- Repository/Country Evaluation
  - D3js (Javascript Data Visualisation Framework)

Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek

Total Number of Documents per Country (May 2019)

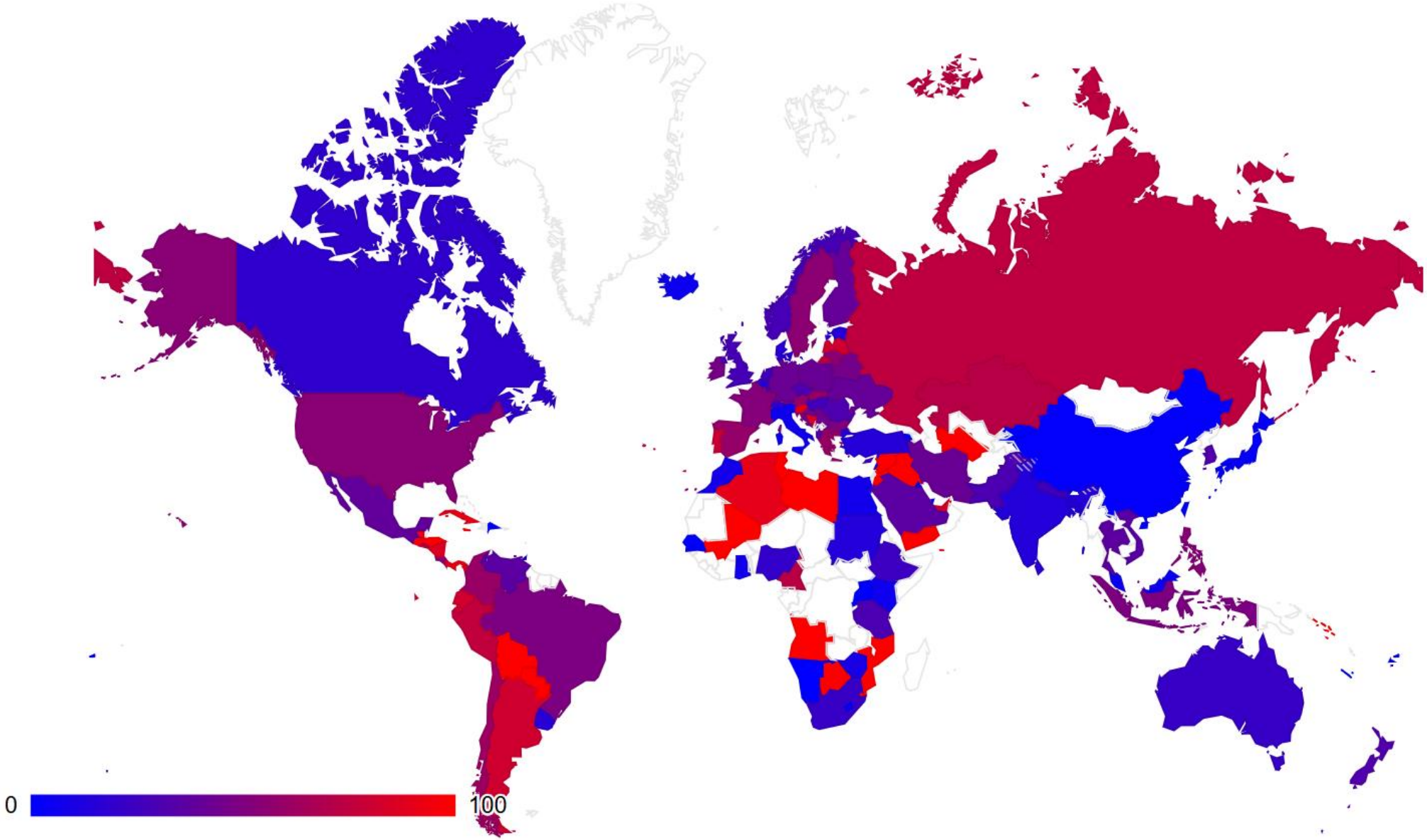
BASE
Bielefeld Academic Search Engine



Universität Bielefeld | INFORMATION.plus! Universitätsbibliothek

Percentage OA Records per Country (May 2019)

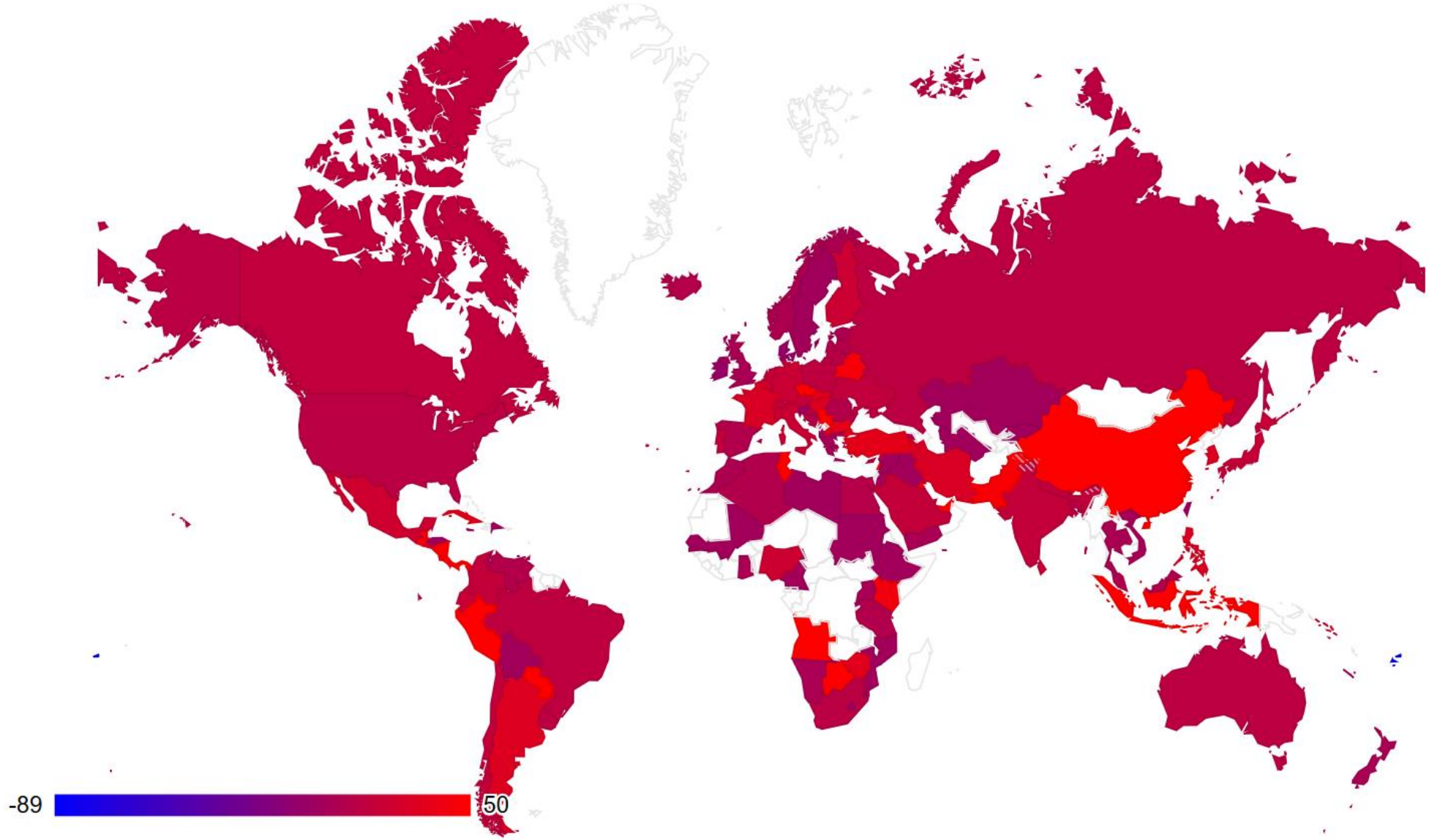
BASE Bielefeld Academic Search Engine



Universität Bielefeld | INFORMATION.plus! Universitätsbibliothek

In/Decrease Number of Documents per Country - 2018

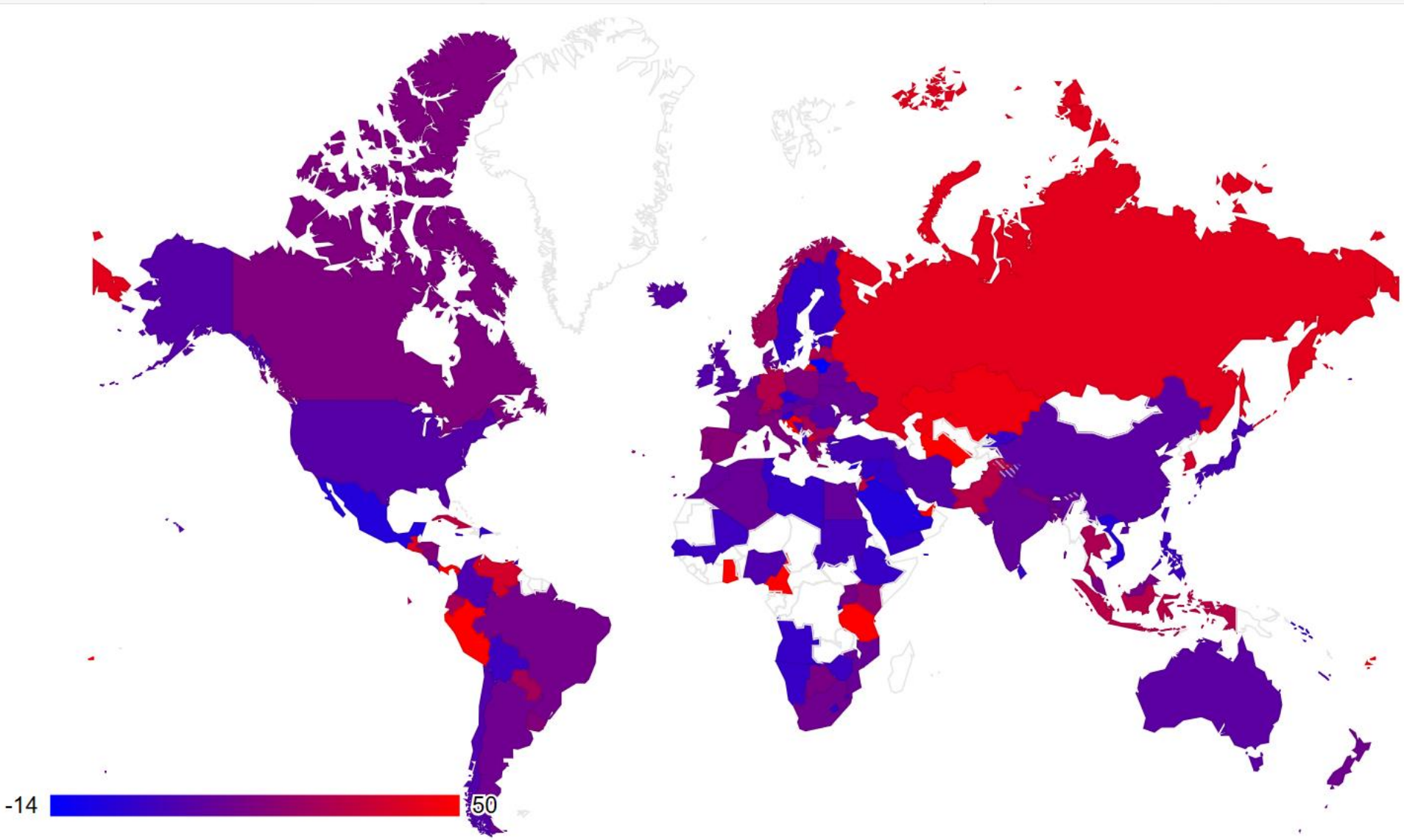
BASE Bielefeld Academic Search Engine



Universität Bielefeld | INFORMATION.plus! Universitätsbibliothek

In/Decrease Number of Documents per Country - 2019

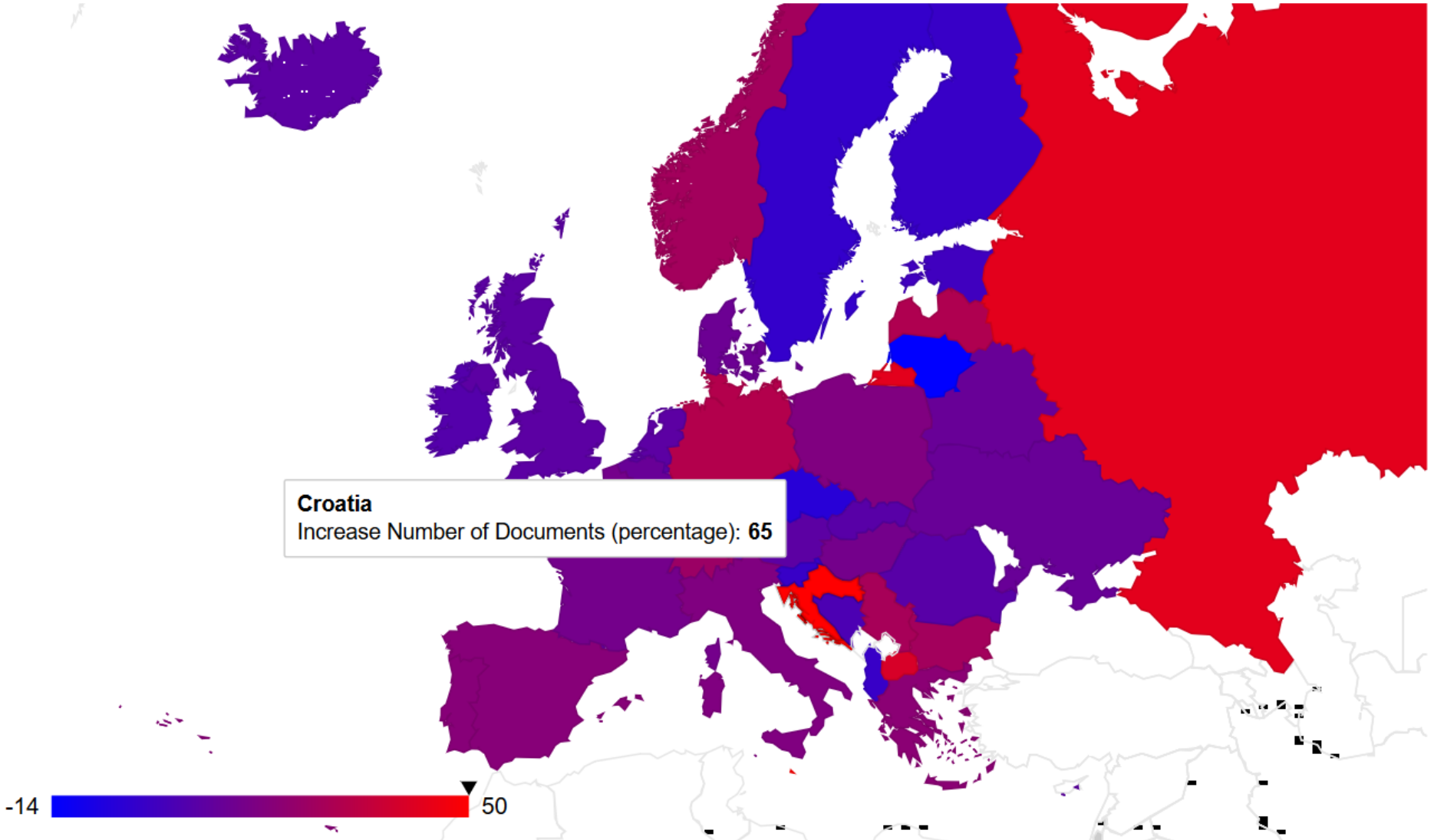
BASE Bielefeld Academic Search Engine



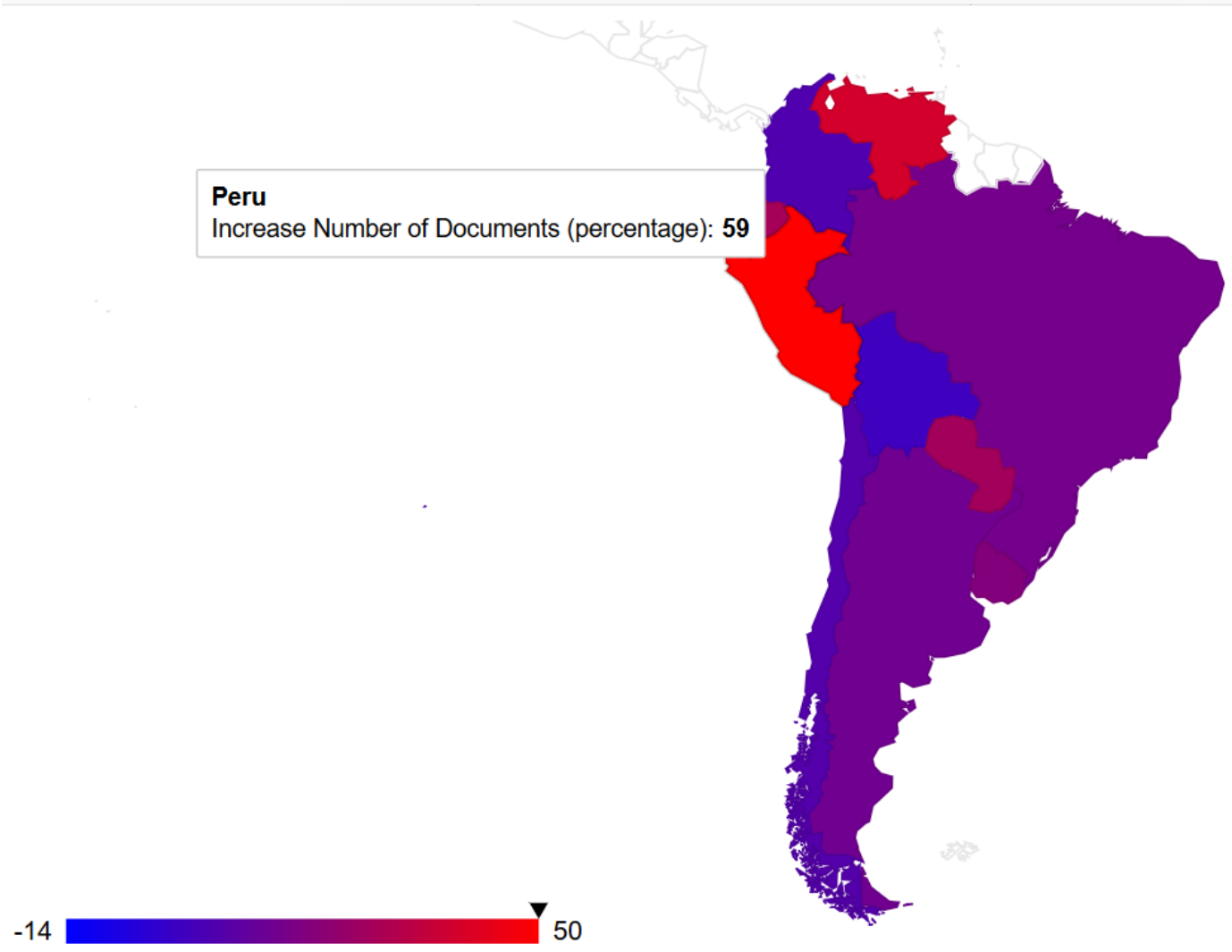
Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek

In/Decrease Number of Documents per Country - 2019

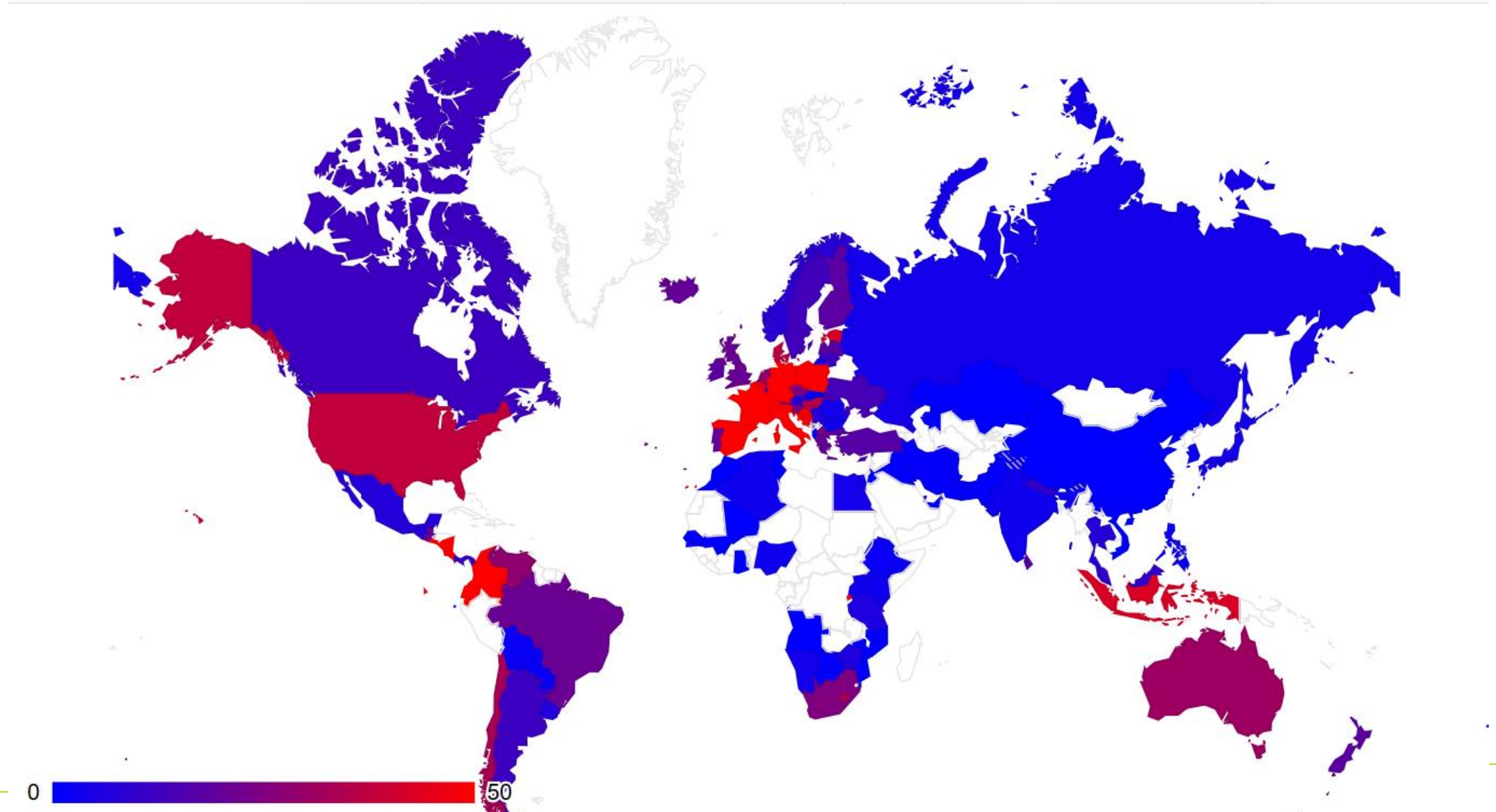
BASE
Bielefeld Academic Search Engine



In/Decrease Number of Documents per Country - 2019



Number of Documents per Researcher (May 2019)

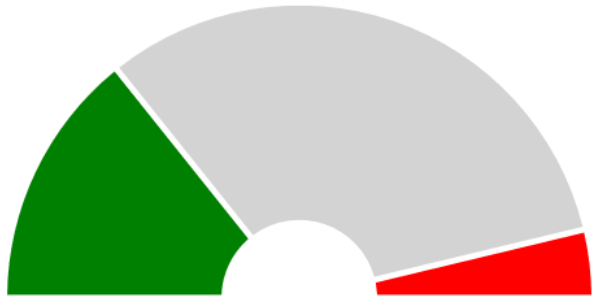


BASE Statistics: - Content Sources

United Kingdom

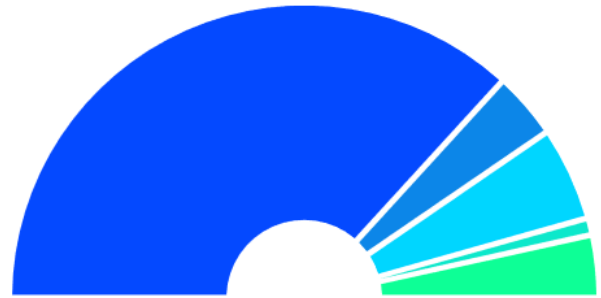
5284718 Documents from **338** Repositories

4 % of Documents world-wide, **5 %** of Repositories world-wide



OA Status Distribution (Number of Documents)

Open Access | Unknown | Non-OA

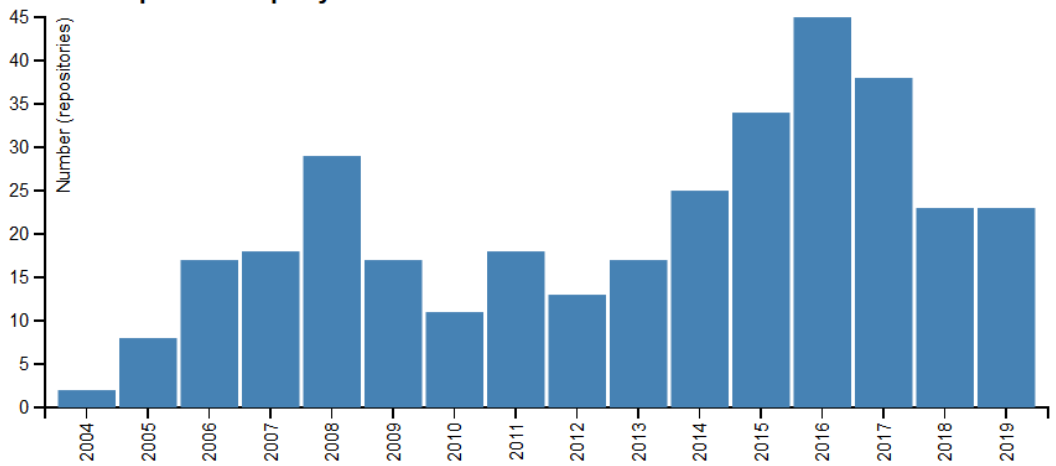


Repository Type Distribution (Number Documents)

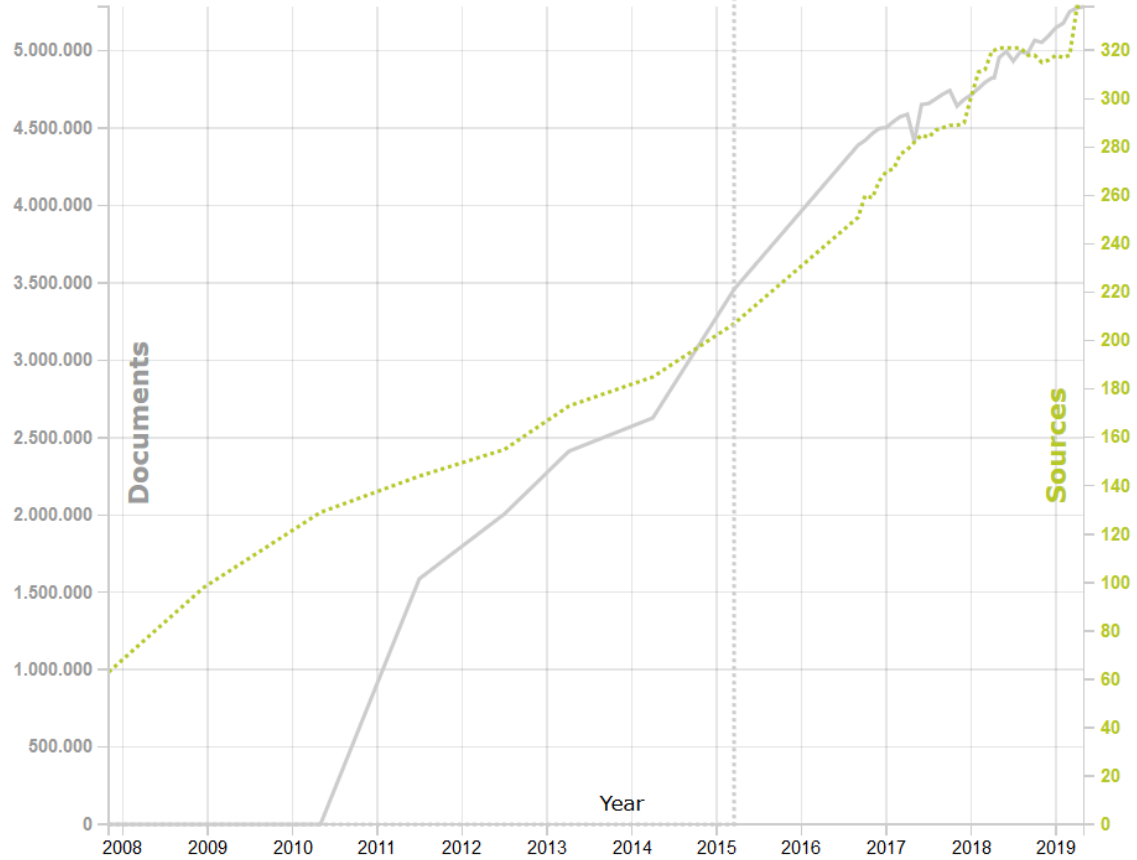
IRs (incl. Thesis) | Journals | Digital Collections | Research Data | Other

Repository Development

New Repositories per year



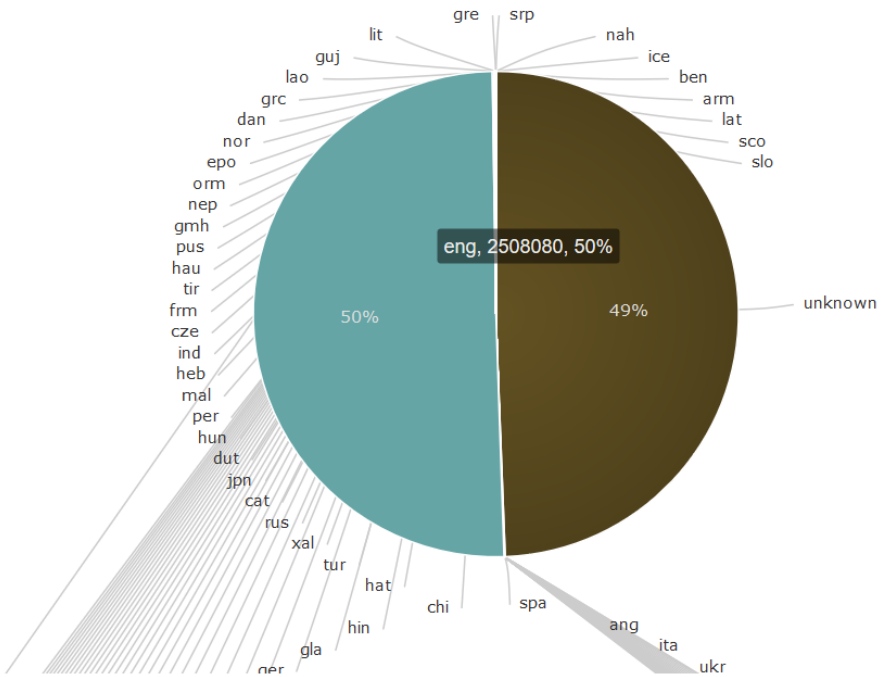
Number of indexed Documente und Sources in BASE for United Kingdom



Repository Contents Analysis

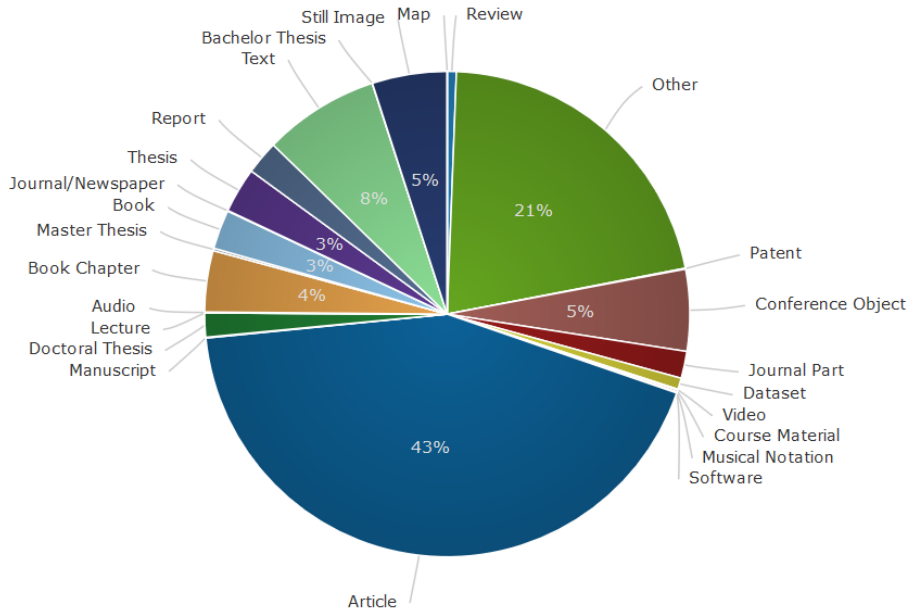
Language Distribution

(number of documents)



Publication Type Distribution

(number of documents)



Project and Funding Context

- Presentation Slides
- Data Provision: Thesis OA Development
- DINI List of Repositories
- Bibliometrics
- ORCID2-DE

Danke!