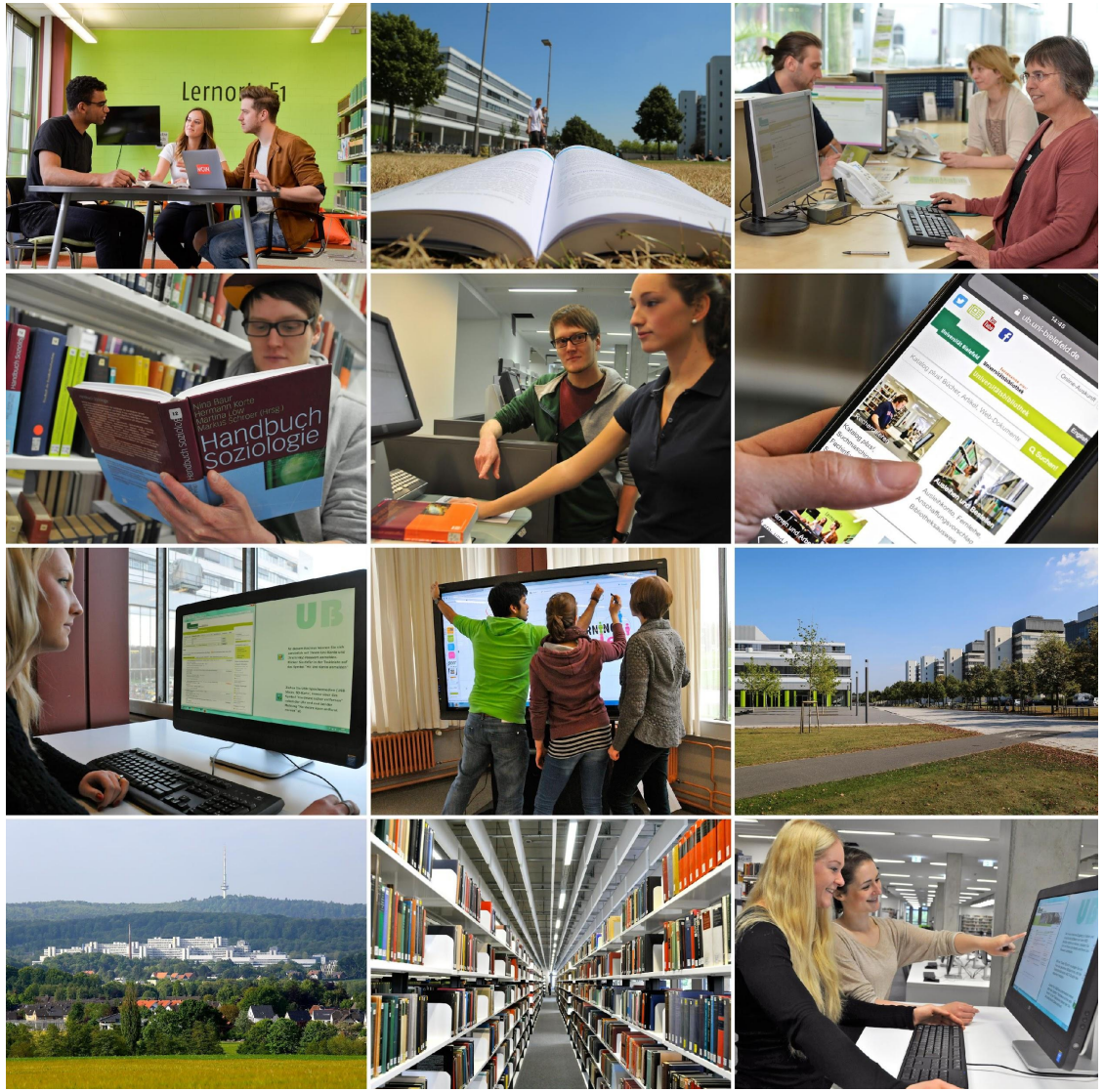
UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Harvesting-Strategien

- BASE
- OpenAIRE



Vitali Peil
0000-0002-6477-8992

Andreas Czerniak
0000-0003-3883-4169

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Agenda

- Einleitung
- Harvesting Strategien bei
  - BASE
  - OpenAIRE
- Ausblick

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Einleitung

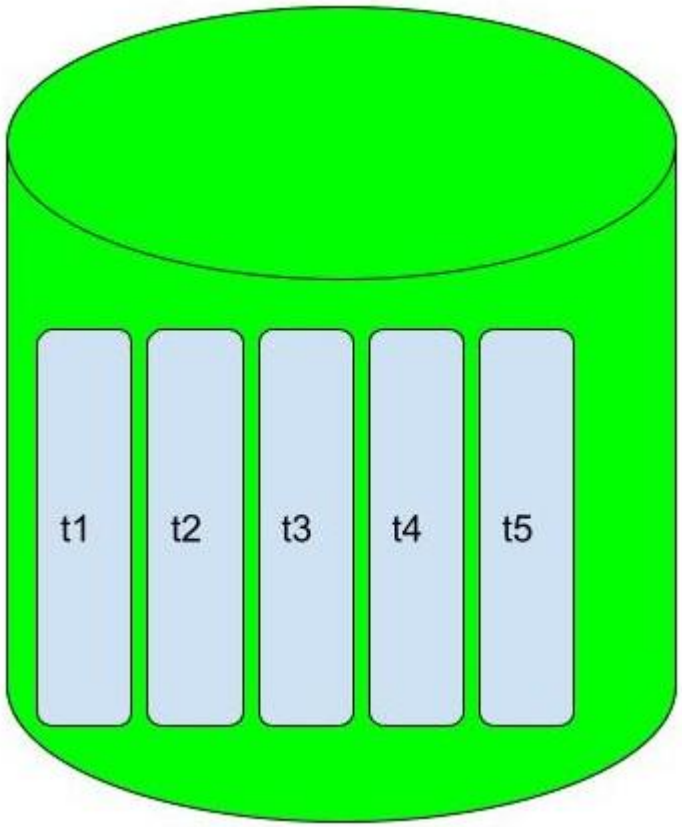
- Von den ersten E-Print-Servern (z.B arXiv) bis zum ersten Standard für deren Interoperabilität verging ein Jahrzehnt (2001/2002)
- Definition des Protokolls OAI-PMH und des Formats Dublin Core
- Grundlage für weltweite Aggregator-Services
- Bis heute keine “Konkurrenz”, was die Verbreitung angeht

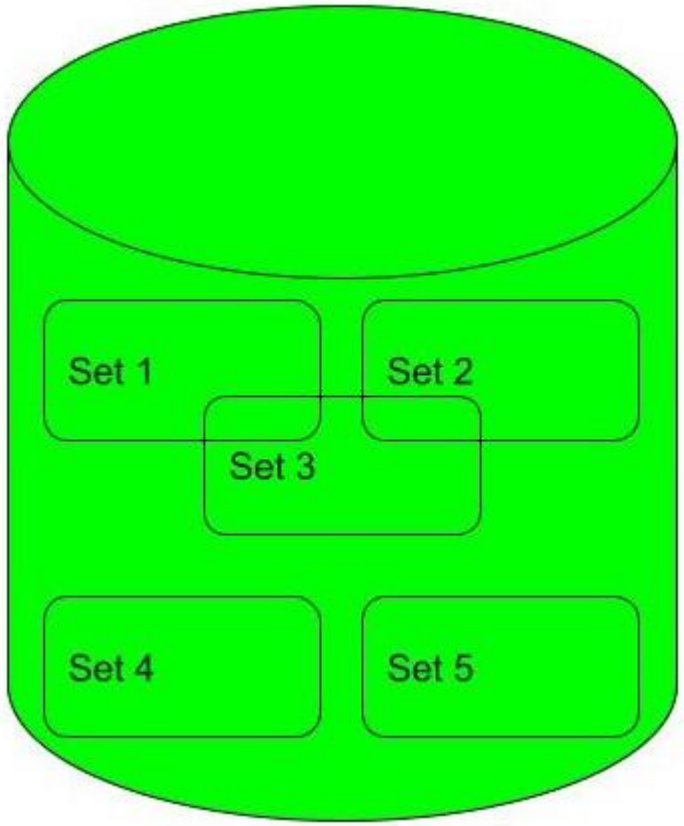
UNIVERSITÄT BIELEFELD
Universitätsbibliothek

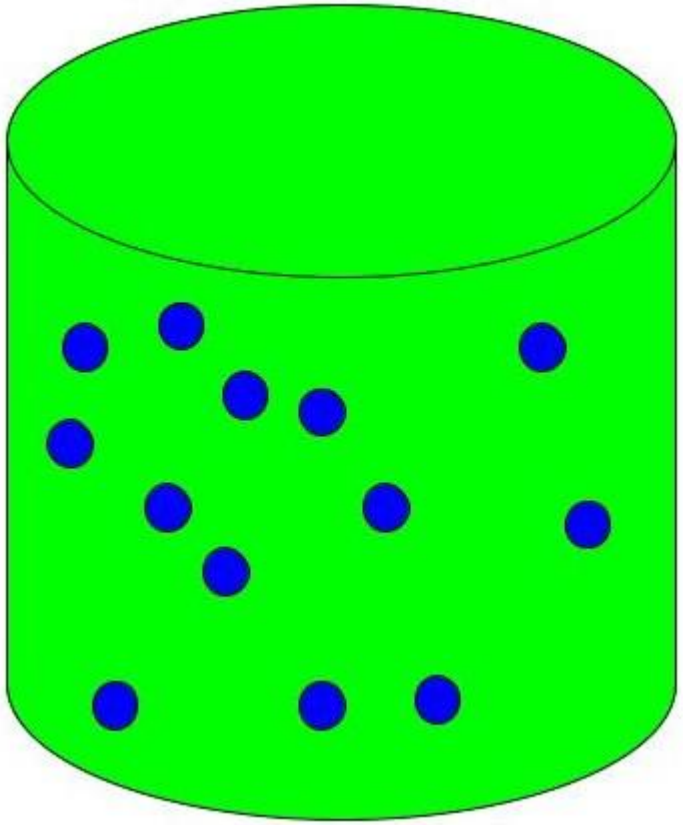
Harvesting :: allgemeine Strategien

- Wie sammelt man große Datenmengen mit Hilfe von OAI-PMH ein?
  - “ungeordnet” alles auf einmal
  - nach den Sets
  - nach Zeitscheiben (z.B. monatsweise)
  - Liste der einzelnen Identifier und dann jeden Record einzeln
  - wie verfährt man, wenn es bei einem dieser Verfahren zum Abbruch kommt?

Harvesting :: allgemeine Strategien







UNIVERSITÄT BIELEFELD
Universitätsbibliothek

BASE

Das Aggregationssystem

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Harvesting :: derzeitige Infrastruktur

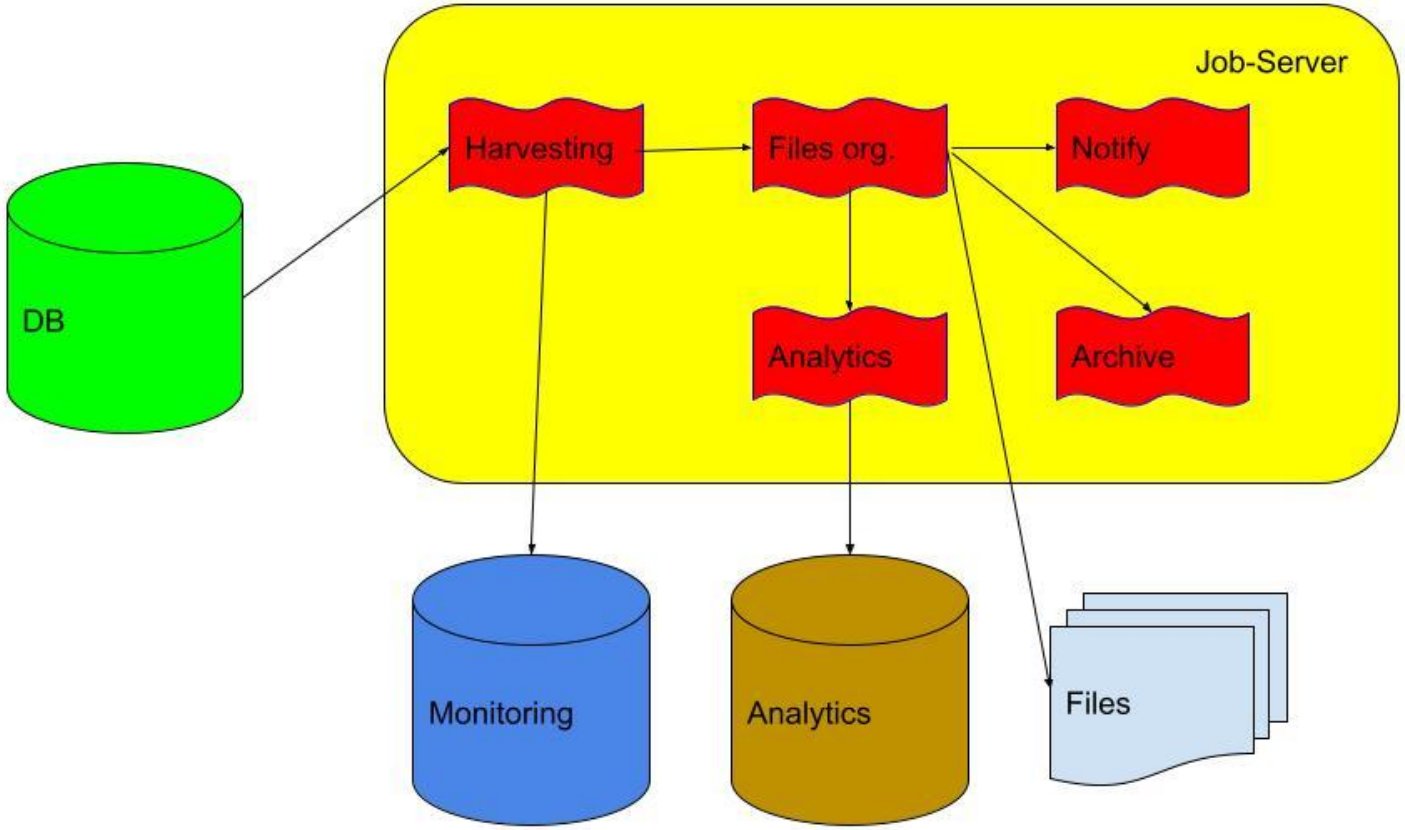
- Eine zentrale Konfigurationsdatei
- Start des Harvestingvorgangs
- Cronjobs
- Nachteil:
  - Bearbeitung der Datei nur von einer Person gleichzeitig
  - Zentrale Ablage der Konfigurationsdatei notwendig

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Harvesting :: mögliche Infrastruktur

- Zentrale Konfigurationsdatenbank mit Weboberfläche
- individuelle (unabhängige) Harvestingvorgänge mit Hilfe eines Job-Servers
- Abhängigkeiten von Jobs definieren
- Vorteil:
  - Bearbeitung/Pflege der Konfiguration von mehreren Personen gleichzeitig möglich
  - parallele Data-Pipelines

Harvesting :: mögliche Infrastruktur



Harvesting :: Herausforderungen

- Speicherung der geharvesteten Dokumente
  - Anzahl der Dokumente ~ 10 Mio.
  - Speicherbedarf ~ 10 TB (nicht komprimiert)
- Lösungsansätze:
  - verteiltes Speichersystem
  - Verschmelzen der einzelnen Dateien zu größeren Einheiten (z.B. bis 1000 Records)
  - Dateien archivieren und komprimieren

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Harvesting :: alltägl. Probleme

- Quellen antworten nicht, oder nur sporadisch oder nur sehr langsam
- Quellen liefern keine validen XML-Daten
- Quellen ändern URL, oder nur OAI-URL
- https-Umstellung
- Quellen ändern ihren Zuschnitt (Zusammenlegung/Aufsplittung bei OJS-Zeitschriften)
- uvm.

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

(Echtzeit-)Monitoring

- Harvesting-Vorgänge:
  - Welche Vorgänge laufen derzeit?
  - Wie sind die Antwortzeiten?
  - Wo treten Fehler auf und wie häufig?

(Echtzeit-)Analytics

- Rohdaten:
  - Inhalte potenziell normierbarer Felder (Publikationstyp, Rechte, Sprache)?
  - Welche Systeme liefern welche Felder welchen Inhalts?
  - Wo kann man die Normierung und Indexierung verbessern?

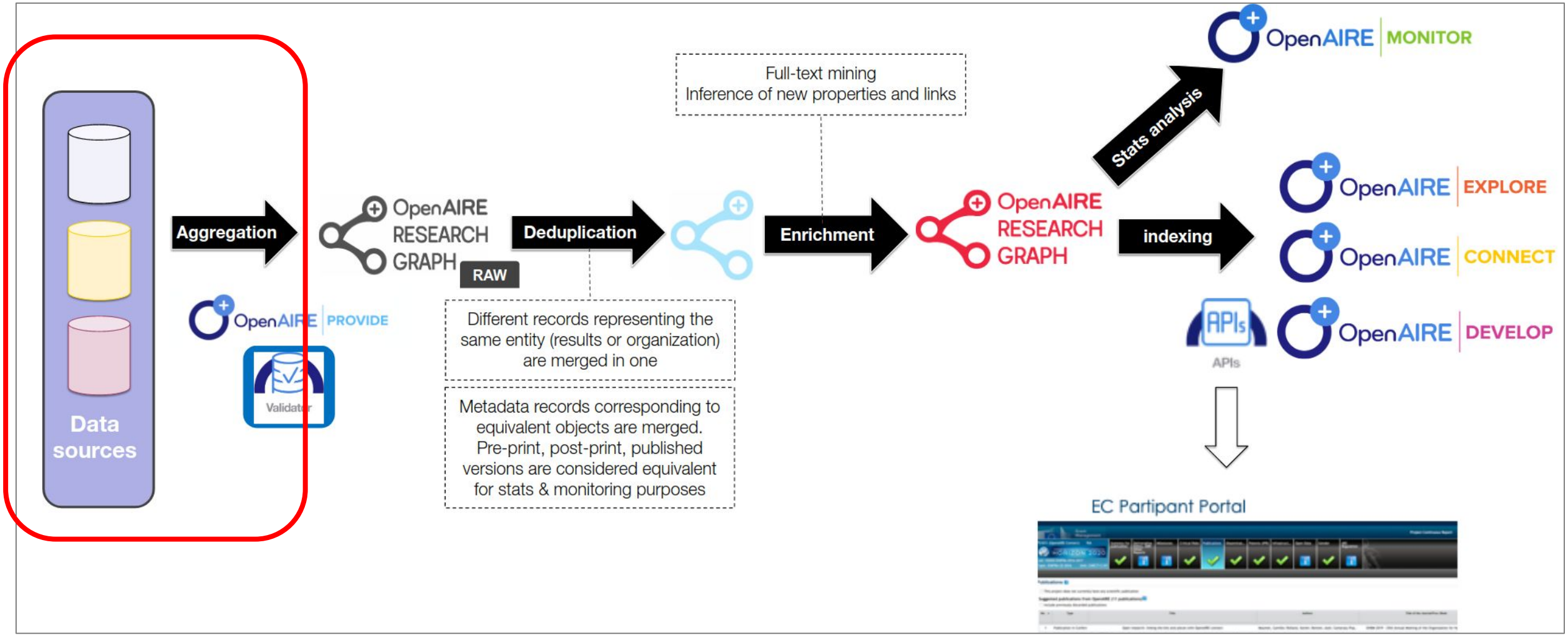
UNIVERSITÄT BIELEFELD
Universitätsbibliothek

OpenAIRE

Das Aggregationssystem

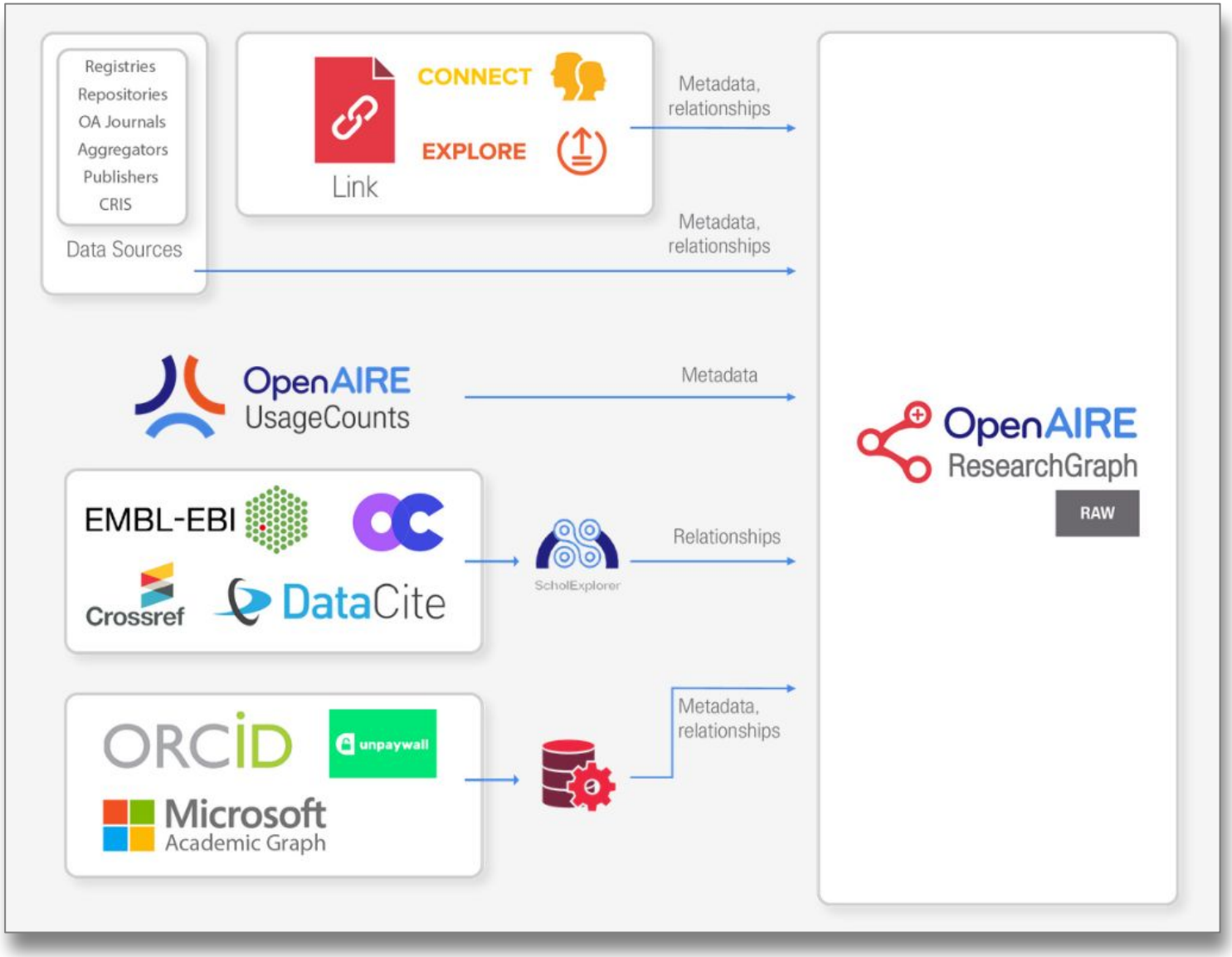
UNIVERSITÄT BIELEFELD
Universitätsbibliothek

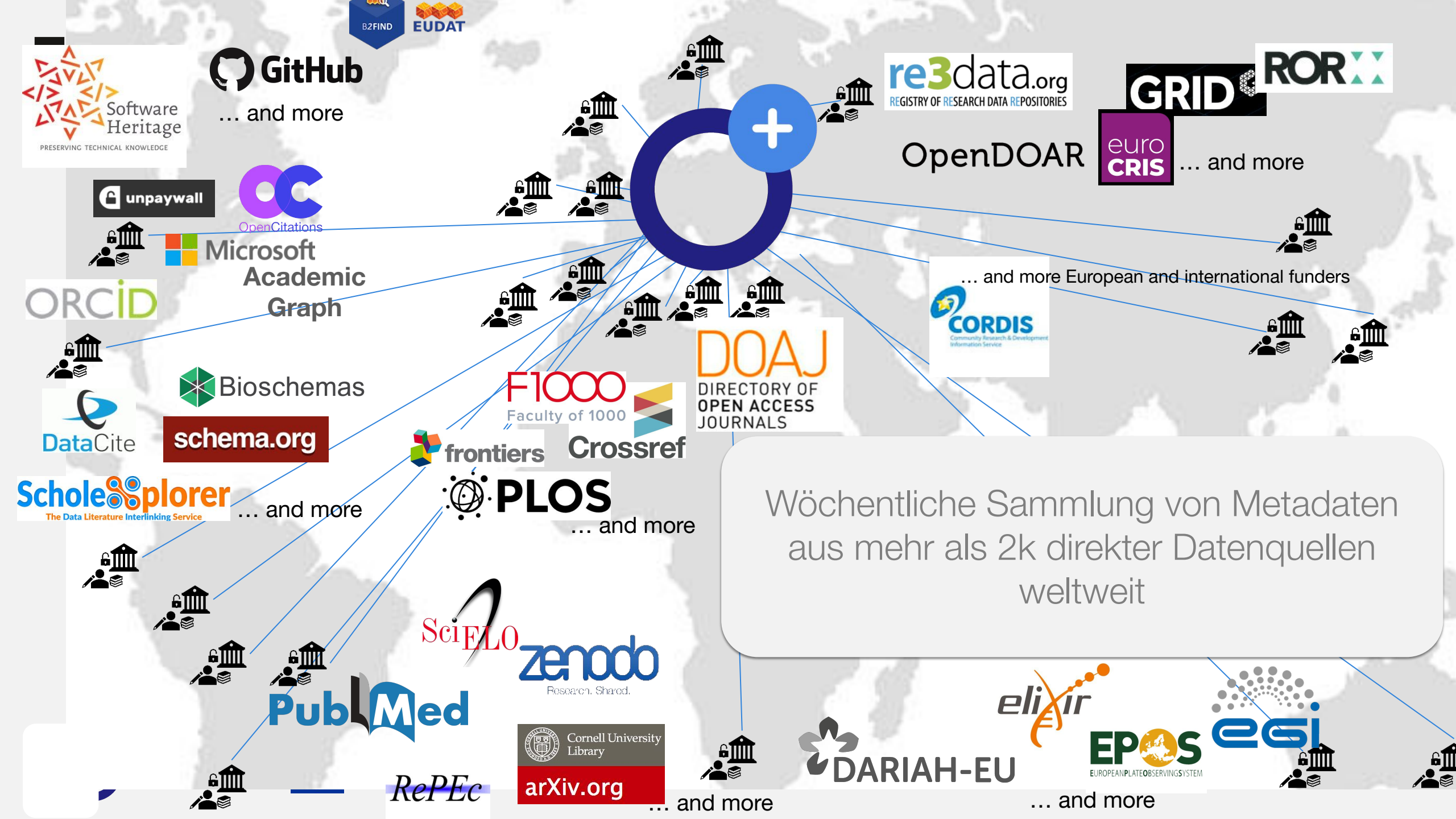
OpenAIRE Research Graph Prozesskette



UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Aggregations-system

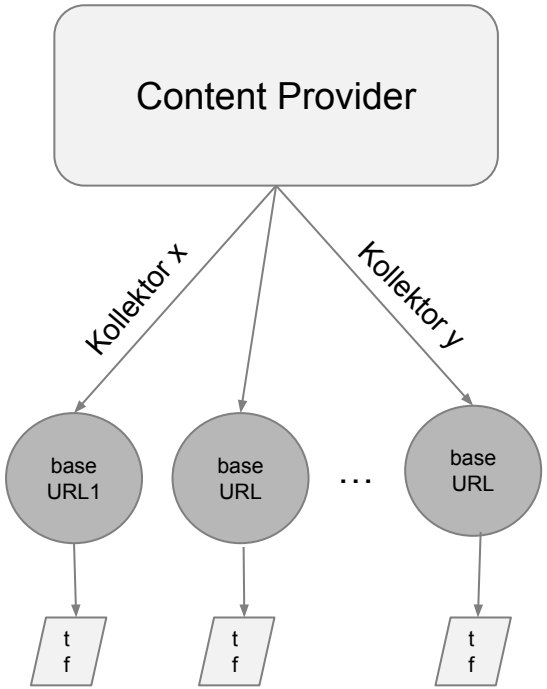



B2FIND
EUDAT
Software Heritage
PRESERVING TECHNICAL KNOWLEDGE
GitHub
… and more
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
GRID
ROR
OpenDOAR
euro CRIS
… and more
unpaywall
OpenCitations
Microsoft Academic Graph
ORCID
… and more European and international funders
CORDIS
Community Research & Development Information Service
Bioschemas
schema.org
DataCite
DOAJ
DIRECTORY OF OPEN ACCESS JOURNALS
F1000
Faculty of 1000
frontiers
Crossref
PLOS
… and more
ScholeXplorer
The Data Literature Interlinking Service
… and more
Wöchentliche Sammlung von Metadaten aus mehr als 2k direkter Datenquellen weltweit
SciELO
zenodo
Research. Shared.
PubMed
Cornell University Library
arXiv.org
RePEc
… and more
DARIAH-EU
elixir
EPOS
EUROPEAN PLATE OBSERVING SYSTEM
… and more

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Design Datenquelle

- *Eine Datenquelle (Content Provider)* kann *einen oder mehrere Endpunkte* besitzen, administrierbar via PROVIDE & DSM-UI.
- Jeder einzelne Endpunkt hat
  - einen Kollektor
  - ein spezielles Transformation-Skript zur Harmonisierung und Normalisierung zur Vermeidung von Inkonsistenzen und Anomalien (post-processing):
    - dc_cleaning
    - XSLT



Design Harvesting

- Harvesting Prozeß je Endpunkt
  - Start des Harvesting Ablaufes via Zeitsteuerung zu festgeleten Zeitpunkten oder Variablen mit einem Mindestzeitraum zwischen den Prozessen,
  - limitiert auf 20 parallele Prozesse in der *nicht-verteilten* und *verteilten* Umgebung,
  - max. 5 Versuche einen Endpunkt zu kontaktieren.
- Kollektor
  - beim Harvesting zwei Modi: REFRESH (no interval), INCREMENTAL
  - Speicherung in MongoDB oder HDFS

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Kollektoren 1/2

- Protokoll: **OAI-PMH**
  - Metadaten Formate gemäß den OpenAIRE Richtlinien, d.h. *oai_dc* , *oai_dc* mit set: *openaire* , *oai_openaire* , *oai_openaire_jats* , *oai_cerif_openaire*

- Protokoll: **FileGzip**
  - Metadaten Formate gemäß den OpenAIRE Richtlinien, d.h. *oai_dc* , *oai_openaire* , *oai_openaire_jats* , *oai_cerif_openaire*

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Kollektoren 2/2

- Protokoll: **REST**
  - Metadaten Formate nicht spezifiziert, individuelle Anpassung an den Endpunkt notwendig, zB. OpenDOAR,

- vollständige Liste der Kollektoren unter: https://support.openaire.eu/projects/openaire/wiki/API_protocols

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Problematik beim OAI-Harvesting

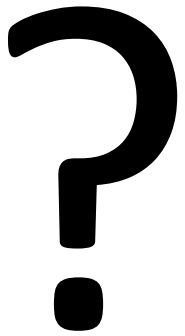
- Strategie bei einer Störung:
  - Verwerfung **aller erfolgreich** geharvestenten Metadaten und Abbruch des Harvestings
- Inkompatibilität zu Endpunkten mit neueren SSL/TLS Protokoll-Varianten,
- keine resourcen-abhängiger Start von Harvesting Prozessen
- Kein frei wählbarer OAI-Kollektor oder Strategie-Änderung für den Endpunkt möglich, mit dem auf eine Störung reagiert werden kann.

Problematik nach dem Harvesting

- Analyse der **original** Metadaten nicht möglich
  - Aufspaltung eines jeden Rekords, für MongoDB bzw. HDFS, in einen *nativen* Metadaten-Rekord,
  - Modifizierung des Original-Metadaten-Rekords.
- Generell
  - keine integrierte Auswertungs-Komponente, welche über die Zeit eine Datenquelle betrachten kann,
  - keine historische Analyse von Metadaten möglich,
  - Schwierigkeit der Analyse von Logdateien, in denen parallele Prozesse einliefern.

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Ausblick I



Ausblick II

- neuere Endpunkte
  - SignPosting, u.a. FAIR Assessment
  - graph-basiert, selektive Abfrage von Metadaten
    - GraphQL, SPARQL
      - Notwendigkeit der Standardisierung von Anfragen
- Link data notifications

UNIVERSITÄT BIELEFELD
Universitätsbibliothek

Vielen Dank für Ihre/Eure Aufmerksamkeit

und seid FAIR.