Markus M. Geipel

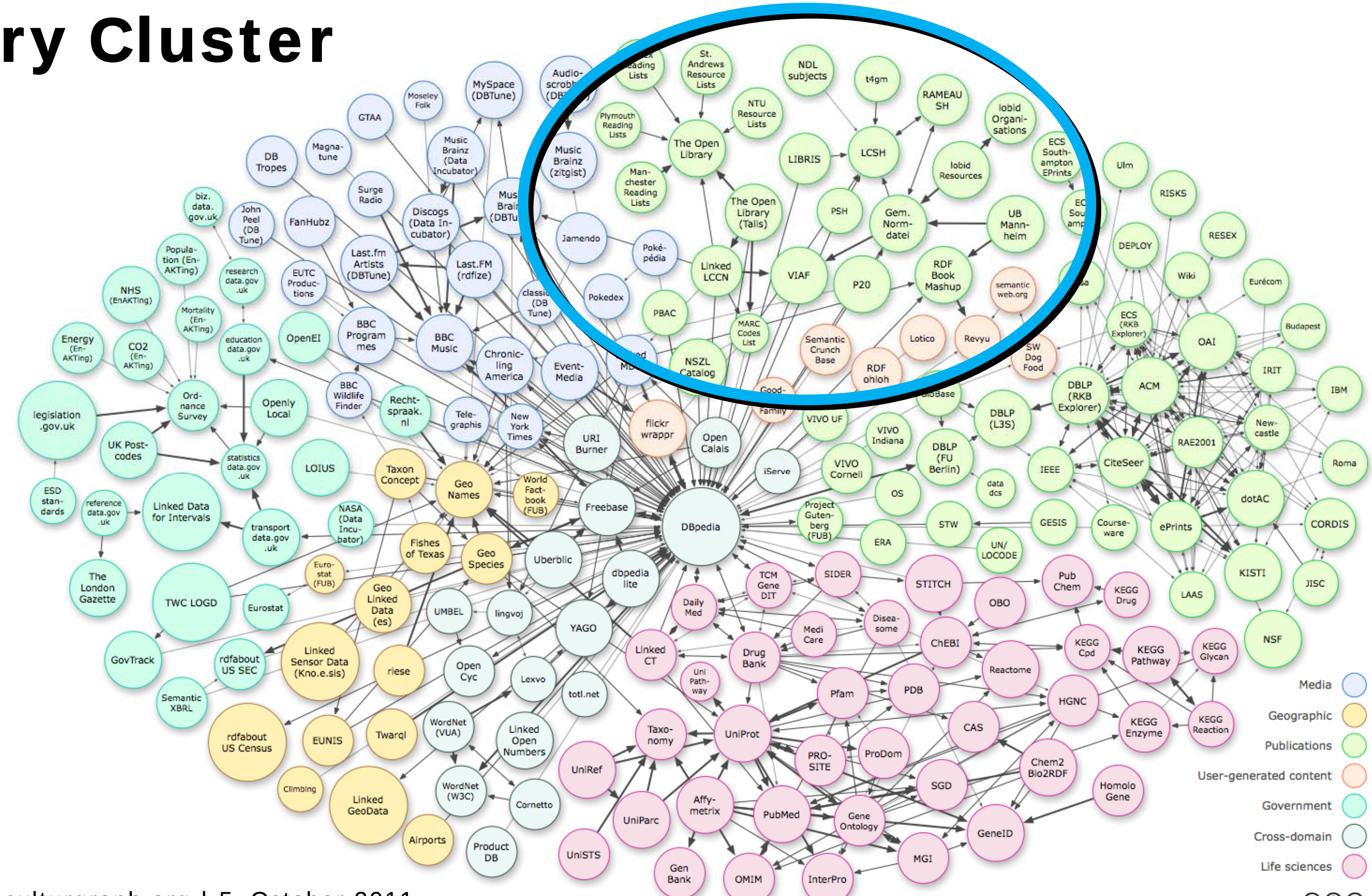
Improving the German Digital Library - Data enrichment with culturgraph.org

DEUTSCHE
NATIONAL
BIBLIOTHEK

Table of Contents

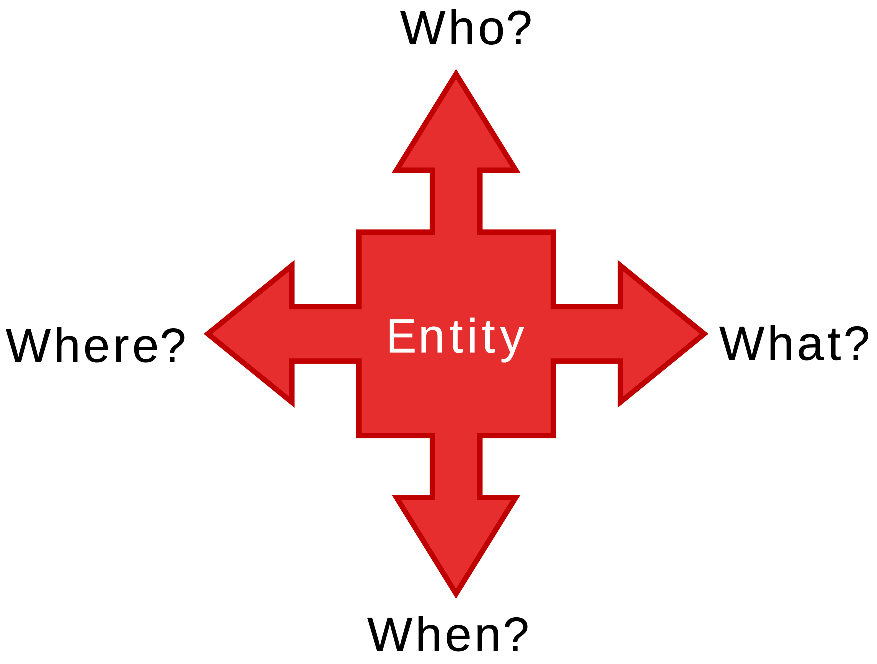
1. **Motivation**
2. **Culturegraph**
3. **Projects**
4. **Technology**
5. **Current State**
6. **Perspectives & Discussion**

Culturegraph is about tightening connections in the Library Cluster

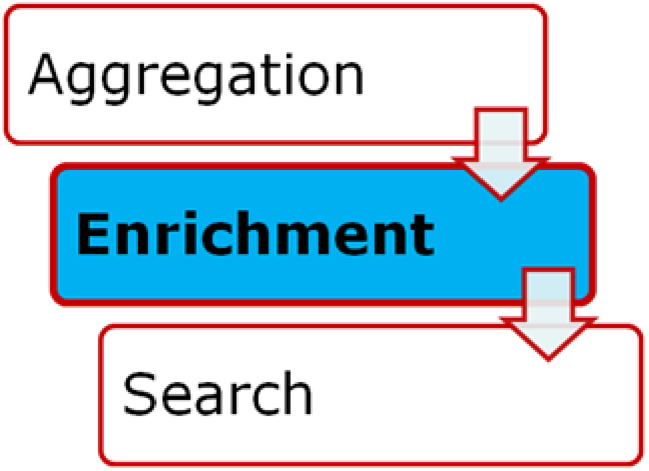


As of September 2010

The Challenge:
Semantic connections between datasets



Aggregation + search is not enough!



The Situation

- Several Bibliographic Catalogues with Overlap
  - 6 Main Catalogues in Germany
  - DDB receives data from ~40.000 cultural heritage organizations
- Different Interfaces
  - No uniform RDF/Linked Data access (SILK/SPARQL not an option)

- Common use of Authority Data
  - Person names, Corporate Bodies, Subject Headings
- Common need for linked Authority Data
  - Geographic Subject Headings – Geo coordinates
  - Person names – Wikipedia Entries for Persons
  - Large amount of redundant work

- **How to offer a coherent view on cultural data?**

DEUTSCHE NATIONAL BIBLIOTHEK

Culturegraph as a Platform

1. **Open Tools**
   - Open algorithms and code; reuse

2. **Integration into existing Workflows**
   - Synchronization of data
   - Integration of results into original data sources

3. **Publication Results**
   - Connections and views, *not* the entire aggregated Data
   - Linked Open Data/RDF

4. **Persistence of Results**
   - Integration into URN resolving infrastructure

5. **Tracking provenance**

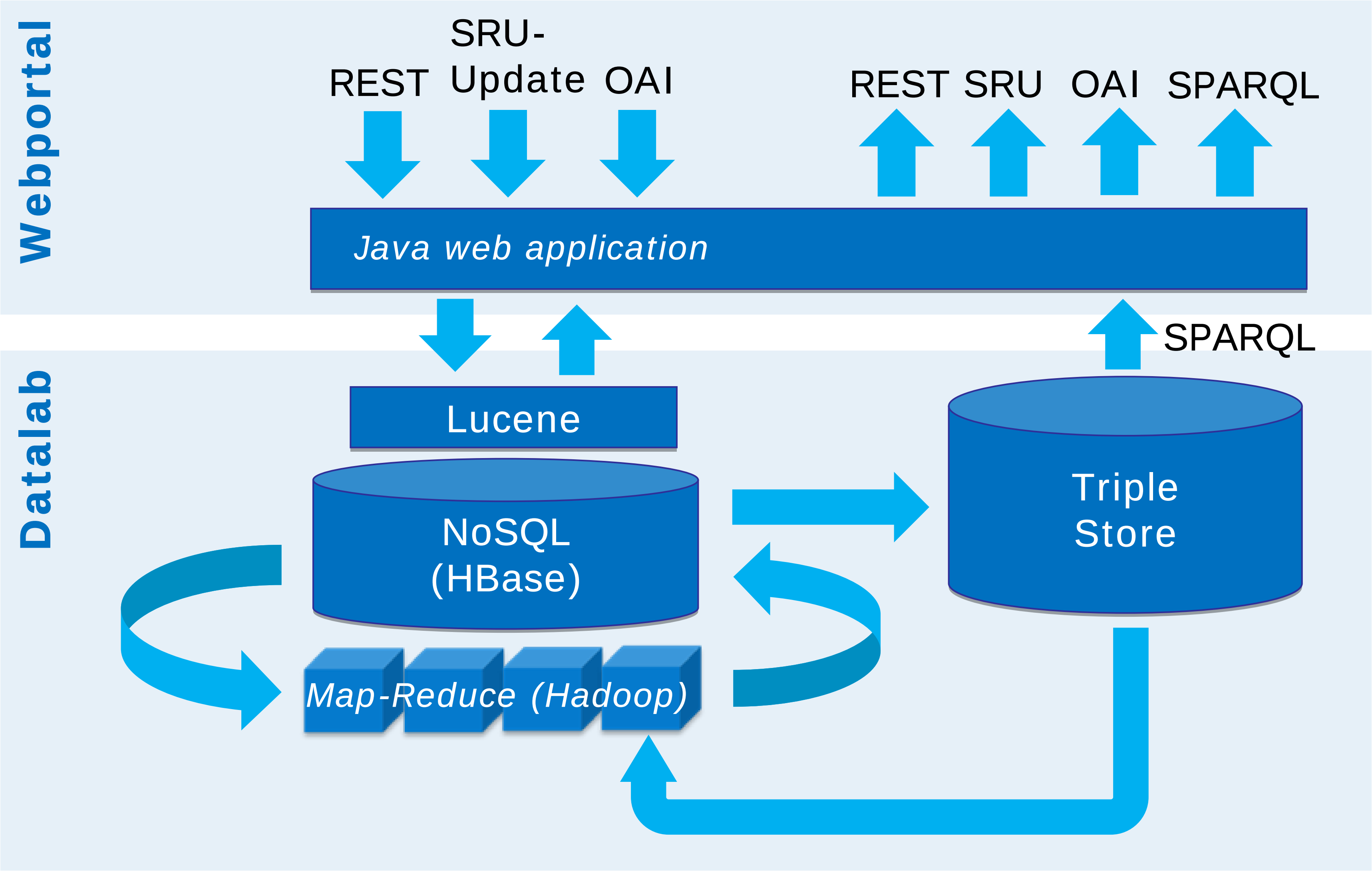
Specific Project: Resolving & Lookup in German Library Data

- **Input:**
  6 main German bibliographic catalogues

- **Objective:**
  Bundling of manifestations

- **Service:**
  - Publication of bundles
  - Minting of URNs for approved bundles
  - Search bundles using established identifiers

- **Part of the DDB Eco-System:**
  - Data Registration
  - Support for Data Aggregation
  - Data Lab


Webportal
SRU-
REST
Update
OAI
REST
SRU
OAI
SPARQL
Java web application
SPARQL
Datalab
Lucene
NoSQL
(HBase)
Triple
Store
Map-Reduce (Hadoop)

Current State

- Focus on Bibliographic Data Matching in Germany
- First Demonstrator at www.culturegraph.org
  - Second planned for December
- Code on Sourceforge
- Technology
  - Preliminary Hadoop Cluster up and running
  - Testdata: ~25 Million bibliographic records in marc21 and mab2
  - First Matching results
  - Good performance (<1h)
- **Soon to come**
  - Publication of results

Perspectives & Discussion

- Integration with DDB
  - DDB as aggregator for Europeana
- Broaden scope
  - Geographical
  - Topical
- Open for cooperation