



CBS in the context of (linked) swissbib

CBS Partner Meeting 2017

Günter Hipler – Systems Architect, Project swissbib

Silvia Witzig – Metadata Specialist, Project swissbib

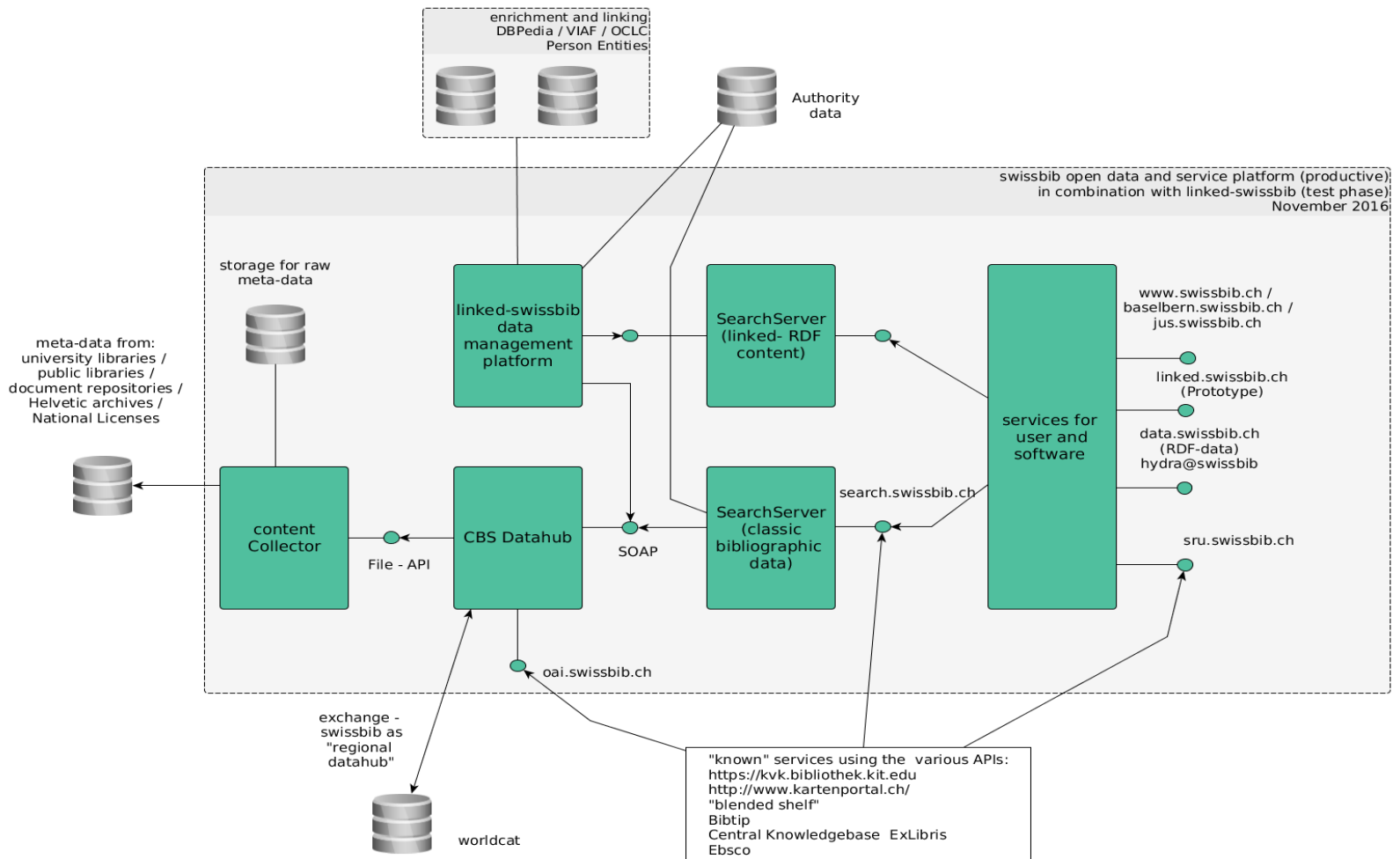
Agenda

- Background
- CBS and the Master Record Model
- linked.swissbib.ch: The Results
- Architecture development
- What we'd like to do in the cooperative

swissbib

- Platform for data and search services
- Since 2008 at the University Library of Basel
- Data from 23 library networks and digital repositories
- Various interfaces for users and software services

Architecture of the swissbib Platform



CBS in the context of swissbib

- Tool for data management
- Clustering and merging
- CBS as a data hub
- Daily data processing and pushing to other components

swissbib and MRM

- What we needed:
A model which takes all updates in source systems into account and updates merged records daily
- The solution:
Master Record Model (MRM)
Developed with OCLC 2010-2013, live in swissbib since 2013

MRM Process

- Transform and store all records from source systems
- Find possible duplicates (divide)
- Evaluate and decide if the records should be merged
- Choose a master record
- Merge data from slaves to the master record

Divide

Matchkey or ISBN must match, otherwise no evaluation

- Matchkey:
 - Based on title (n letters from word x)
 - Form of item and material type
 - Exact date, if present
 - URL, if published before 1900
- ISBN:
 - Based on ISBN/ISSN
 - + Form and material, exact date, URL if published before 1900

Evaluate

- ISBN (same as for divide)
- Title
- Corporate
- Person
- Year of publication
- Decade of publication
- Century of publication
- Edition
- Part
- Pages (+/- 1)
- Volumes
- Publisher – Initials (for all records)
- Publisher (only for serials)
- Scale
- Coordinates
- Source (only for non-text material)

Decide

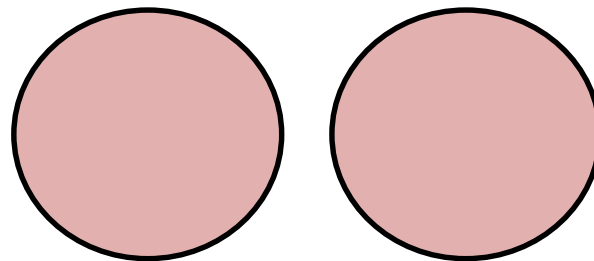
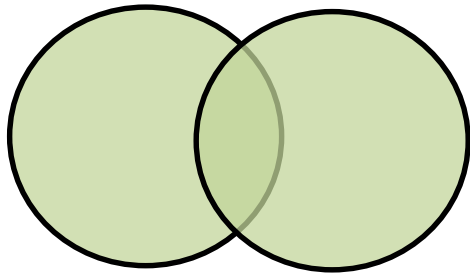
- One case per evaluated index
- Similarity (0 or 1) per case
- Weight for each case is possible
- Calculation of overall similarity of the records

In swissbib:

- Match Limit for merging is 1
- Weight is irrelevant
- 3 main cases which decide about merging

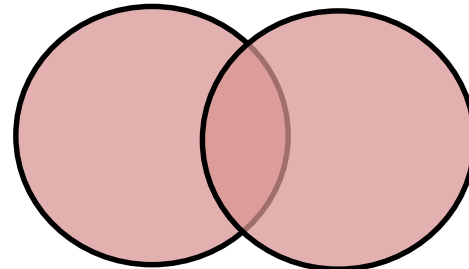
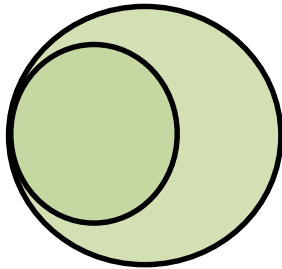
Decide

- If index is present in one of the records but not in the other → merge is possible
- If index is present in both records:
One of them matches → merge is possible
None of them matches → no merge



Decide

- If index is present in both records:
The values of one record are completely contained in the other → merge is possible
Both records have differing values → no merge



Some observations

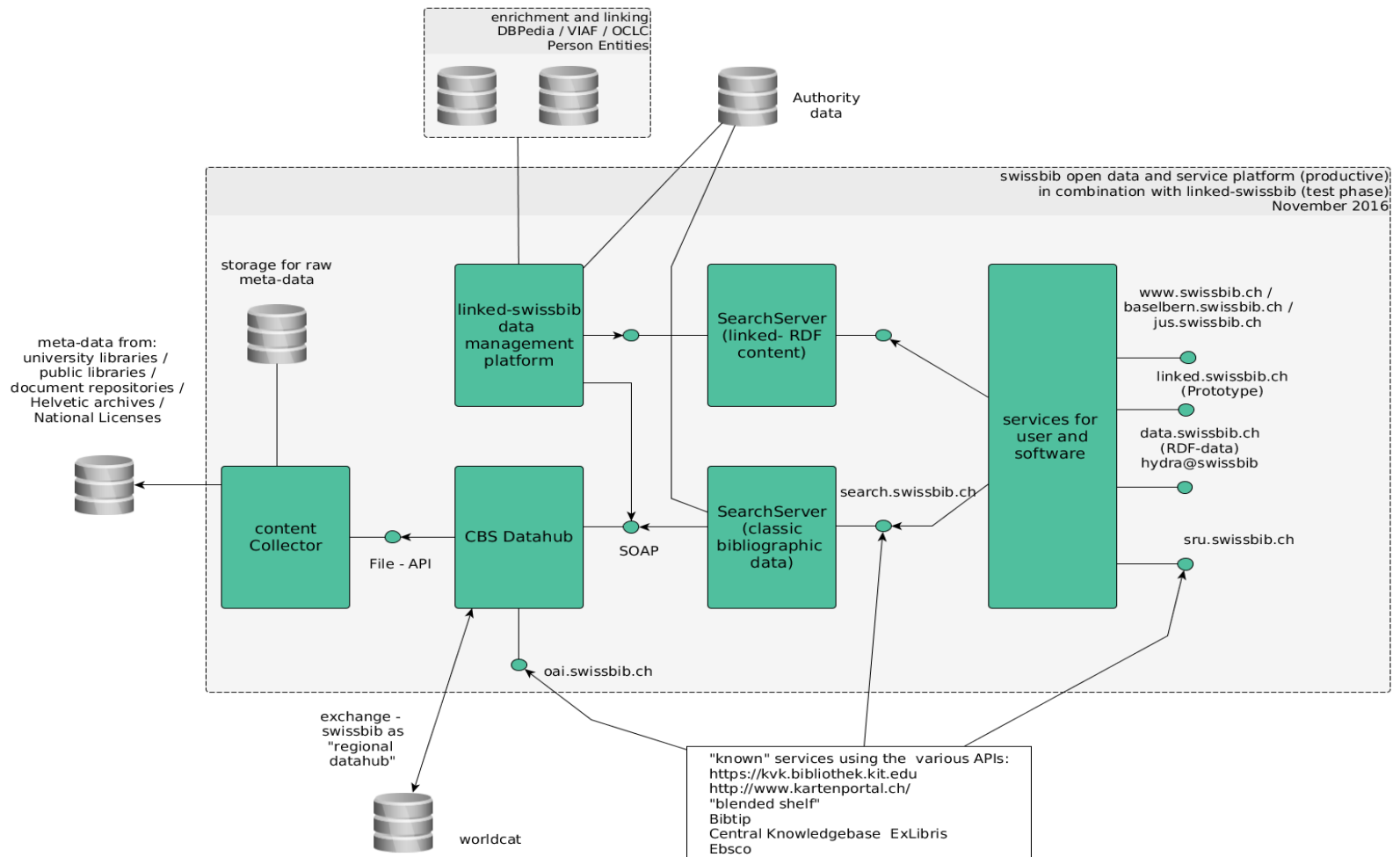
- MRM does what we need
- Not enough merging vs. Too much merging
- Small differences prevent merging
- Records with little information cause wrong merges

Updates through MRM

- Number of updated masterrecords: 29'515
- Number of deleted masterrecords: 634
- Number of inserted masterrecords: 4'977
- Number of records made unique: 23'288
- Number of relinked slaves: 17'917

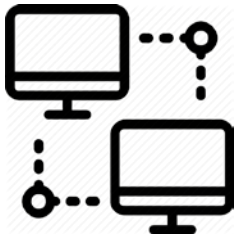
- Records / Items move a lot → no persistent ID

Architecture of the swissbib Platform



Project linked.swissbib.ch

Objective: Make swissbib linked data compatible



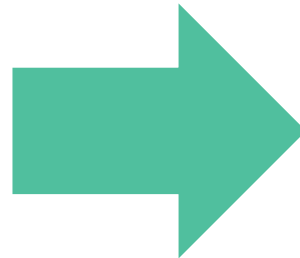
Create an open interface for computer clients (RESTful API)



Create an improved interface with linked data for end users

Results: Data Transformation

~ 29 Mio. MARC
Records




~125 Mio. documents in JSON-LD, divided into 6 bibliographic Concepts

- Bibliographic Resource
- Document
- Item
- Work
- Person
- Organisation

Results: User Interface

Werke Personen

bauer 

Bücher & Co.

Ein Blick in einer mögliche Zukunft der Schweizer **Bauern**

Die **Bauern** in der Tiroler Landschaft vor 1500 : Die politische Aktivität der Gerichte und ihre Repräsentanten auf den Landtagen

Wir entdecken Berge, Meer und **Bauernhof**

Der Toggenburger **Bauern**maler Gottlieb Feurer brachte es als Autodidakt zu hohem Können

Bauer sucht Hahn

AutorInnen

Rudolf **Bauer** (? - ?)

Thomas Wilhelm **Bauer** (1986 - ?)

Günter **Bauer** (1936 - ?)

Horst **Bauer** (? - ?)

Egbert **Bauer** (? - ?)

Themen

Bauer (**Bauernschaft**; **Bauernstand**; **Bauerntum**; **Bauern**)

Bauer

Vergifteter **Bauer**

Results: User Interface



Werke Personen

Erweiterte Suche

/ [Suche](#) / **Personenseite: Robert Walser**

Springe zu: Bücher & Co. von Robert Walser | Co-AutorInnen von Robert Walser | Bücher & Co. von Robert Walsers Co-AutorInnen | Mit Robert Walser verwandte Themen | Bücher & Co. mit ähnlichen Themen wie Robert Walser | AutorInnen von Büchern & Co. mit ähnlichen Themen wie Robert Walser



Robert Walser

Beruf/Tätigkeit: Keine Inhalte vorhanden
Geboren: 15.04.1878, Biel/Bienne, Schweiz
Gestorben: 25.12.1956, Herisau, Schweiz
Nationalität: Schweiz

► [Mehr Details zu Robert Walser](#)

Literatur und Medien

FAQ

Was ist eine Personenseite? ▲

Die Personenseite enthält durch [Linked Open Data \(LOD\)](#) angereicherte Angaben zu Personen, welche als VerfasserInnen oder Beitragende in den Medien von swissbib aufgeführt werden. Als Datenquellen zur Anreicherung werden die [DBpedia](#), die [GND](#) sowie [VIAF](#) genutzt.

Was ist linked.swissbib? ▲

Results: User Interface

Johann Sebastian Bach (1685 - 1750)

close



Beruf/Tätigkeit: Komponist, Thomaskantor

Geboren: 31.03.1685, Eisenach, Heiliges Römisches Reich, Sachsen-Eisenach

Gestorben: 28.07.1750, Leipzig, Stammesherzogtum Sachsen

Nationalität: Keine Inhalte vorhanden

Mit Johann Sebastian Bach verwandte Themen: Keine Inhalte vorhanden

Bücher & Co. von Johann Sebastian Bach

➔ Widerstehe doch der Sünde : BWV 54 : Kantate zu Oculi für Alt solo, 2 Violinen, 2 Violen und Basso continuo = Stand firm against the lure of sinning : cantata for Oculi for alto solo, 2 violins, 2 violas and basso continuo

» **Mehr Bücher & Co. von Johann Sebastian Bach**

» **Zur Personenseite von Johann Sebastian Bach**

Results: RESTful API

<http://data.swissbib.ch>

- CC0-Data available
- Hydra vocabulary

swissbib 

Hypermedia driven REST API for linked bibliographic resources.

Response

```
@context: http://data.swissbib.ch/contexts/BibliographicResource
@id: http://data.swissbib.ch/bibliographicResource
@type: "hydra:Collection"
hydra:member:
  @id: http://data.swissbib.ch/bibliographicResource/016831268
  @type: http://purl.org/dc/terms/BibliographicResource
  id: "016831268"
  title: "Nella foresta del vocabolario : storia di parole"
  language: http://lexvo.org/id/iso639-3/ita
  instanceOf: http://data.swissbib.ch/work/016831268
  format: "235 p ; 19 cm"
  edition: "[3a ed.]"
  isbn10: "8804360259"
  isbn13: "9788804360254"
  bibliographicCitation: "Oscar guide ; 101. "
  contributor: http://data.swissbib.ch/person/2ca568b8-3e85-3c44-9a5e-aa3cbb7a190c
  issued: "1992"
  isDefinedBy: http://data.swissbib.ch/bibliographicResource/016831268/about

  @id: http://data.swissbib.ch/bibliographicResource/01683142X
  @type: http://purl.org/dc/terms/BibliographicResource
  id: "01683142X"
  title: "Sui miti delle acque"
  language: http://lexvo.org/id/iso639-3/ita
  format: "264 S ; 19 cm"
  contributor: http://data.swissbib.ch/person/8c233b51-62be-3103-895c-361b09bb98bb
  issued: "1895"
  isDefinedBy: http://data.swissbib.ch/bibliographicResource/01683142X/about
```


Software and Architecture Requirements for the linked.swissbib.ch artefacts

(<http://swissbib.blogspot.ch/2014/06/considerations-for-development-of.html>)

- ➡ The software has to transform any kind of meta-data structure into another (meta-data) structure.
- ➡ The software has to be Open Source – otherwise it is often difficult or nearly impossible for other parties or services to (re)-use it.
- ➡ The software shouldn't be in a phase of its infancy. Because of **our really restricted resources (core swissbib team: 2.5 people)** we can't afford to build a new software right from scratch. It should be demonstrated that it is already used by institutions in a similar way that we intend to use it with linked.swissbib.ch.
- ➡ The software tools should be very scalable.
- ➡ The software for meta-data transformation should be usable not only by software developers.
People related to meta-data transformations in the library world (scripting knowledge is helpful but not necessary) should be able to define work-flows for their own purposes

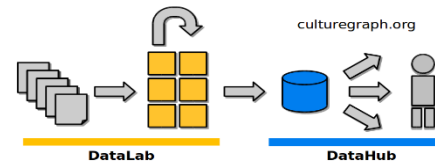
CBS partner-meeting 2012, Amsterdam

(Markus Geipel: Presentation about the «Culturegraph Initiative»)

https://wiki-cbs.oclc.org/wiki/images/Culture_Graph.ppsx

Markus Michael Geipel
Culturegraph

General Architecture



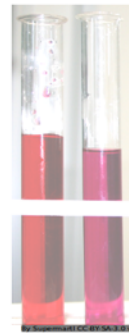
1 Markus Geipel | Culturegraph | CBS partner meeting 18.09.2012

DataLab

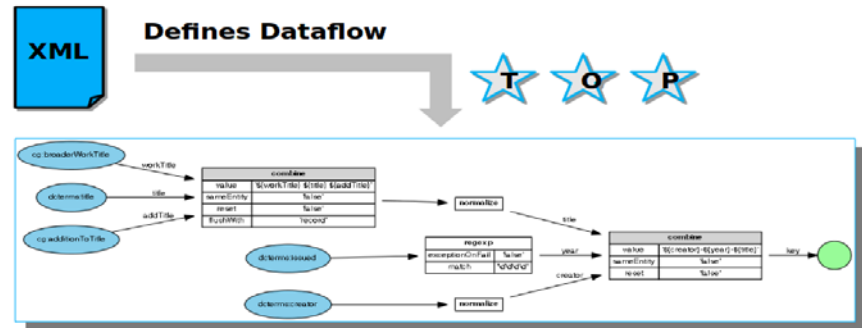
- **Purpose:** Data Processing

- **Implementation:**

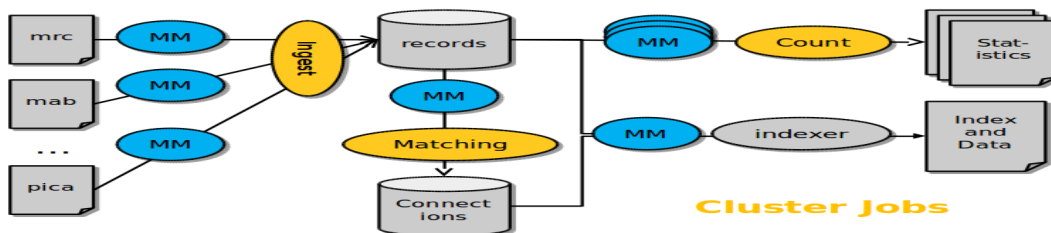
- P** 1. Hadoop Cluster
 - 2. HBase Database
 - 3. Jobs for
 1. Ingest
 2. Statistics
 3. Indexing
 4. Match
 - 4. Java Libraries for bibliographic data processing
- Open Source by Apache Foundation
- O** Open Source by DNB



Data Transformation with Metamorph



DataLab Flow Chart



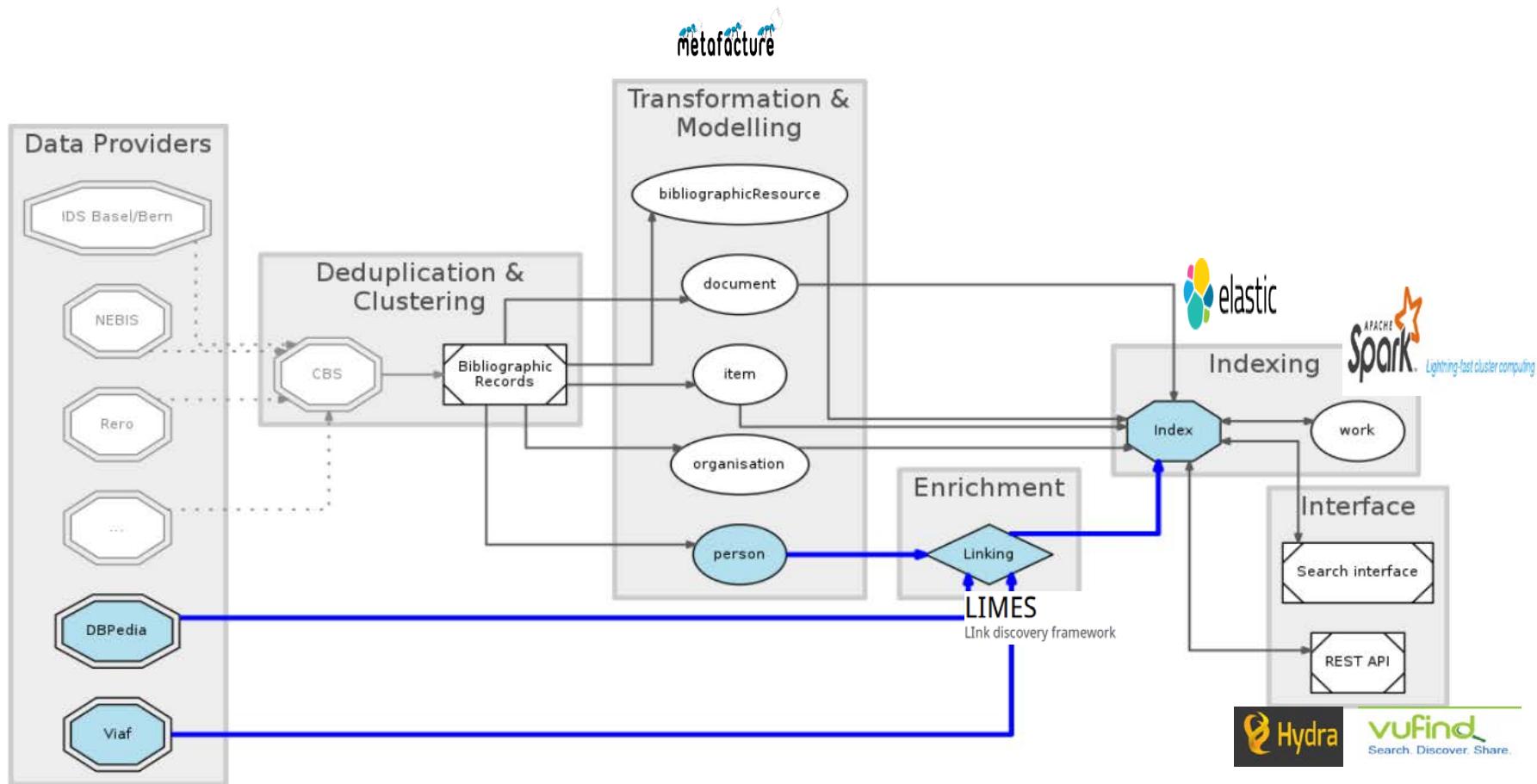
Cluster Jobs

Data Transformations



still complicated workflows

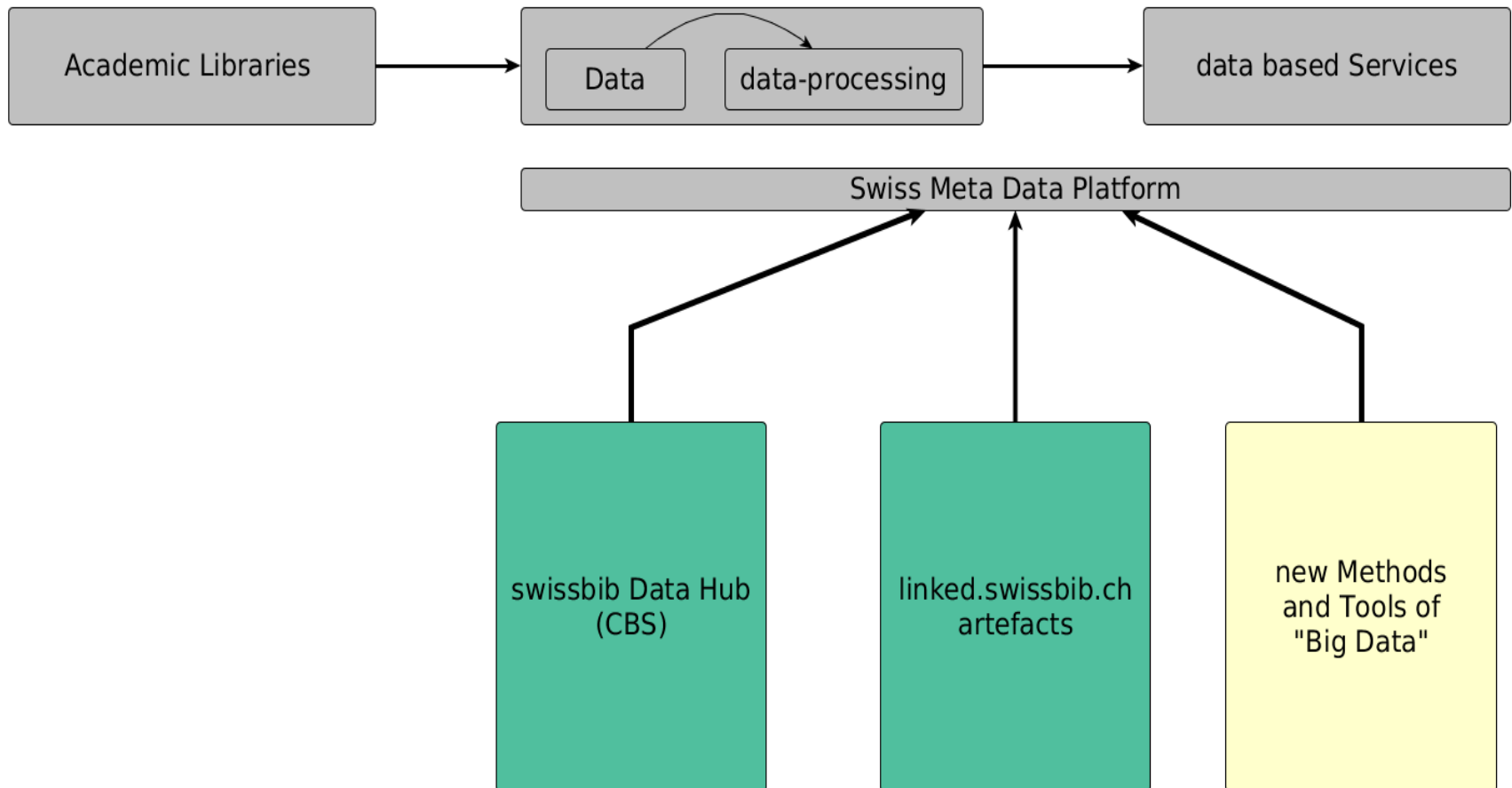
Workflows and components used for linked.swissbib



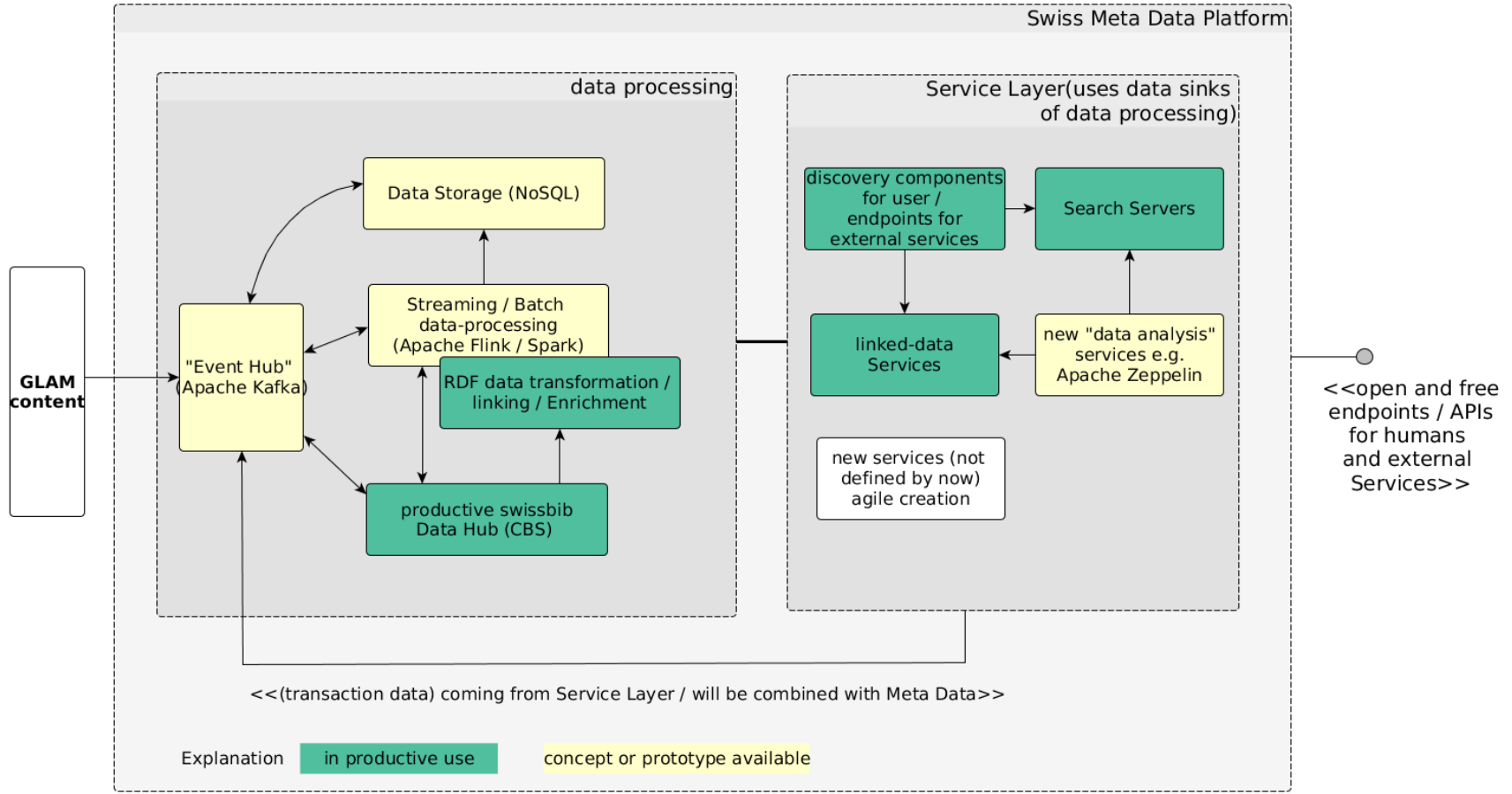
<https://www.elastic.co/videos/elasticsearch-as-hub-for-linked-bibliographic-metadata-zurich-meetup-august-2016>

<http://files.meetup.com/7646592/20160831%20Elasticsearch%20as%20Hub%20for%20Linked%20Bibliographic%20Metadata.pdf>

«Meta Data Platform» for data-based Services



Architectural Blueprint: «Swiss Meta Data Platform»



Current status of our ideas and concepts (part 1)

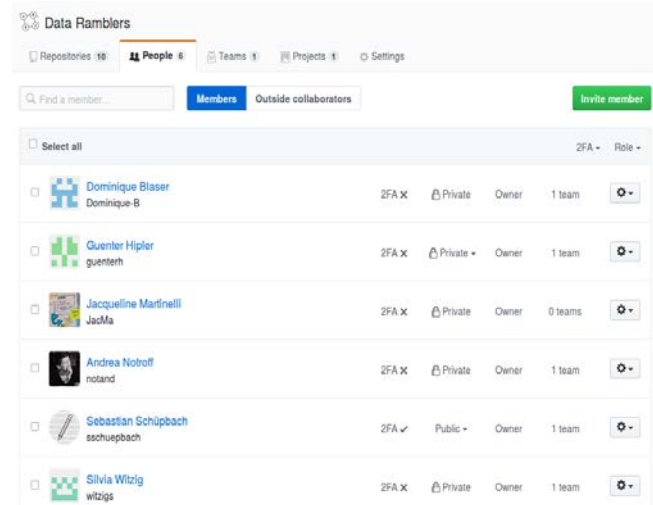
2016 / 2017

- writing documents / discussions with people / continuing qualification and education / arguing for new project funding
(<https://github.com/guenterh/casBigDataBern/blob/master/bern/cas.bigdata.2016.guenter.hipler.pdf>)
- Implementation of very first versions for EventHub and Flink/Metafactory integrations
<https://github.com/swissbib> <https://github.com/linked-swissbib>
<https://github.com/guenterh/dataIngestion>
- Enhancement of our server infrastructure
(Flink and Kafka cluster available)

Current status of our ideas and concepts (part 2)

2017 ... 2020 and more

- «Kafka Event Hub» is going to be the first productive component on the next swissbib platform (end 2017 / beginning 2018)
- Convince more people of the idea



15 / 16. September, Lausanne

What we'd like to do in the cooperative on the CBS field

- Exchange experiences with the Master Record Model
- Implement a solution for Persistent Identifiers in the master record model
- Establish a model / an organisation / a culture where we can learn more from each other
 - ➔ better utilisation of the CBS potential
 - ➔ to get a better knowledge of the system in general

More about swissbib

- GitHub
<https://github.com/swissbib>
<https://github.com/linked-swissbib>
- linked.swissbib.ch: Prototype User Interface
<http://linked.swissbib.ch/>
- linked.swissbib.ch: Beta RESTful API
<http://data.swissbib.ch/>
- Blog series «swissbib data goes linked»
<http://swissbib.blogspot.ch/2016/04/swissbib-data-goes-linked-teil-1.html>
- Interlinking Large Scale Library Data with Authority Records / Felix Bensmann, Benjamin Zapilko and Philipp Mayr
<http://journal.frontiersin.org/article/10.3389/fdigh.2017.00005/full>

Thank you

Günter Hipler

Systems Architect

Project swissbib

Universitätsbibliothek Basel

guenter.hipler@unibas.ch

Silvia Witzig

Metadata Specialist

Project swissbib

Universitätsbibliothek Basel

silvia.witzig@unibas.ch