swissbib

Clustern von Daten auf der swissbib Plattform

KIM Workshop 2019
Mannheim, 3.4.2019

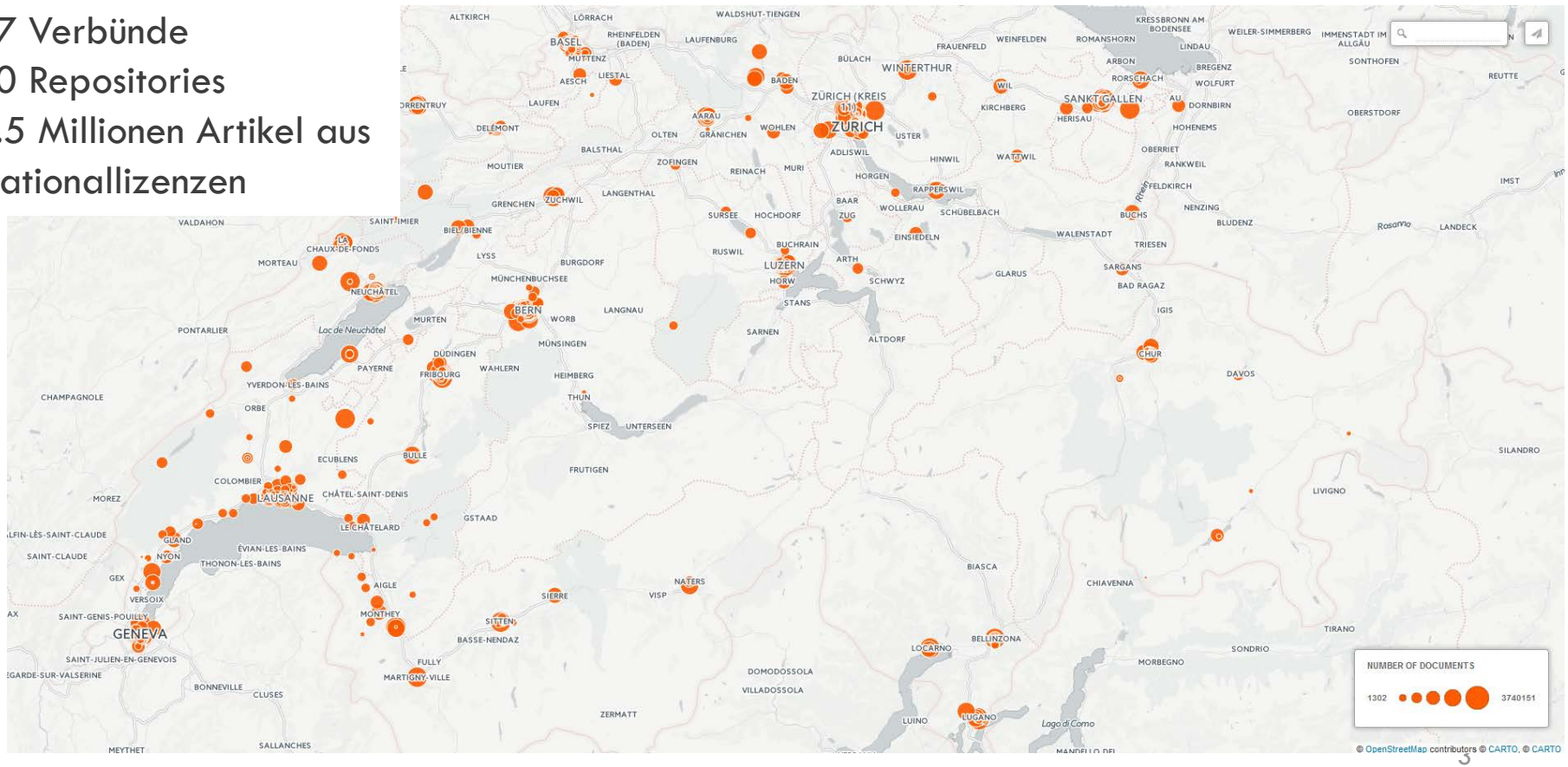
Günter Hipler – Systemarchitekt, swissbib
Silvia Witzig – Metadatenspezialistin, swissbib

swissbib

- Plattform für Daten- und Suchservices
- Wird seit 2008 an der Universitätsbibliothek Basel betrieben
- Enthält Daten von Bibliotheksverbünden, Repositories und Nationallizenzen
- Stellt verschiedene Oberflächen für Menschen und Maschinen zur Verfügung
- Bietet eine Grundlage für verschiedene Projekte in der Schweiz

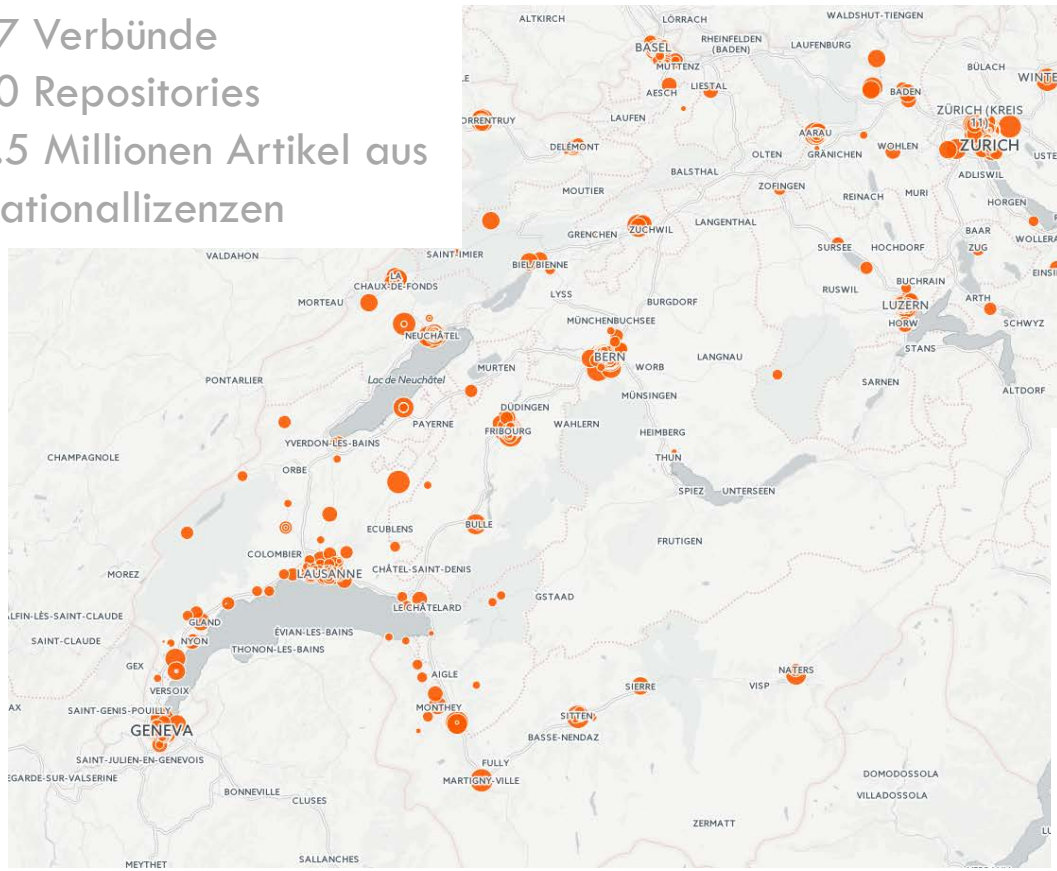
Bibliothekslandschaft – die swissbib-Sicht

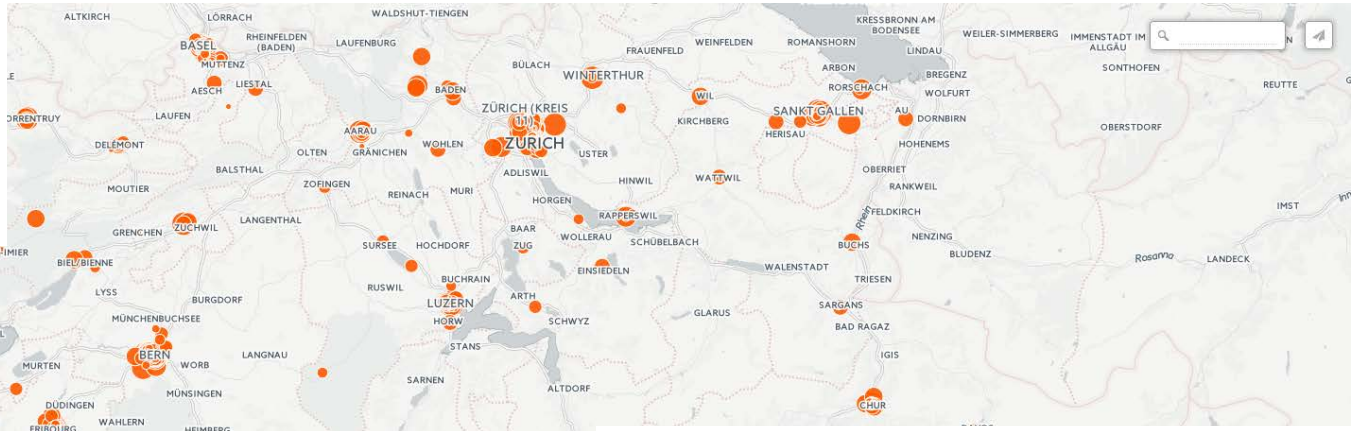
In swissbib:
30 Millionen Aufnahmen
970 Bibliotheken
17 Verbünde
10 Repositories
6.5 Millionen Artikel aus Nationallizenzen



Bibliothekslandschaft – die swissbib-Sicht

In swissbib:
30 Millionen Aufnahmen
970 Bibliotheken
17 Verbünde
10 Repositories
6.5 Millionen Artikel aus Nationallizenzen





Drei Katalogisierungssprachen
Drei Regelwerke für die Katalogisierung
Drei Datenformate, n lokale Flavours

Zusammenführen was zusammengehört

- In CBS von OCLC (mit Master Record Model)
- Bei swissbib so im Einsatz seit 2013
- Clustering verläuft Index-basiert
- Alle Veränderungen in den Quell-Aufnahmen werden berücksichtigt
- Täglich aktualisiert
- Zwei Formen des Clusterings:
  Dubletten und FRBR

Zusammeführen was zusammengehört

- Indexiere und normalisiere die Kriterien
- Finde mögliche Dubletten (divide)
- Evaluiere und entscheide, ob die Aufnahmen gemergt werden sollen (evaluate)
- Wähle einen “preferred Slave”
- Erstelle die Master-Aufnahme und füge die Informationen aus allen Slaves hinzu

Divide

Unterteilt die Daten in Gruppen möglicher Dubletten

Matchkey oder ISBN/ISSN müssen übereinstimmen

- Matchkey-Index:
  Titel (n Buchstaben aus Wort x)
  + Format- und Trägerinformation
- ISBN-Index:
  ISBN/ISSN
  + Format- und Trägerinformation

Evaluate

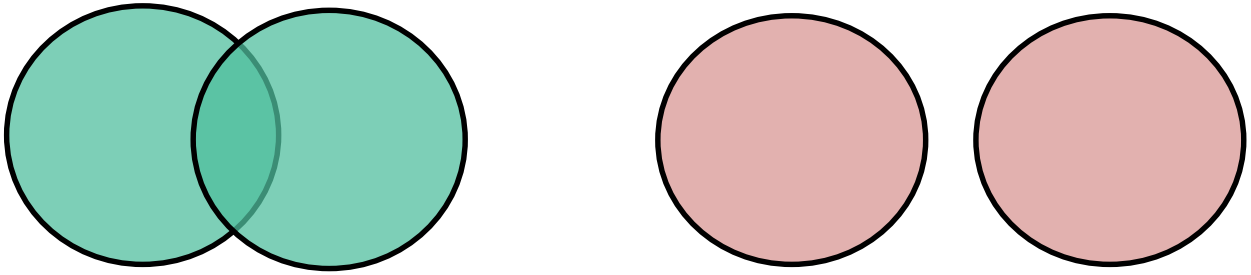
- ISBN/ISSN
- DOI
- ISMN
- Musik Verlagsnummern
- Titel
- Körperschaft
- Person
- Publikationsjahr
- Dekade
- Jahrhundert
- Auflage
- Zählung
- Seiten (+/- 1)
- Anzahl Bände
- Initialen des Verlags (für alle Aufnahmen)
- Verlag (nur für fortlaufende Publikationen)
- Massstab
- Koordinaten
- Quelle (nur für nicht-Text Material)

Evaluate

- Für jeden Index werden die Einträge verglichen
- Pro Index tritt ein Fall ein
- Pro Fall ist eine Gewichtung möglich (bei swissbib nicht eingesetzt)
- Die Ähnlichkeit wird basierend darauf berechnet
- Grundsätzlich setzen wir produktiv drei Fälle ein

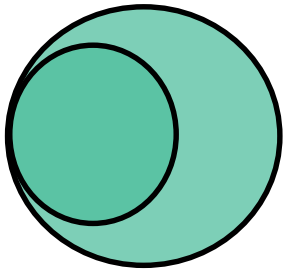
Evaluate

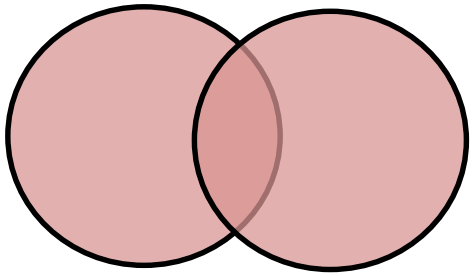
- Nur eine der Aufnahmen hat Index-Einträge
  → merge ist möglich
- Beide Aufnahmen haben Index-Einträge:
  Einer der Einträge stimmt überein → merge ist möglich
  Keine Eintrag stimmt überein → kein merge



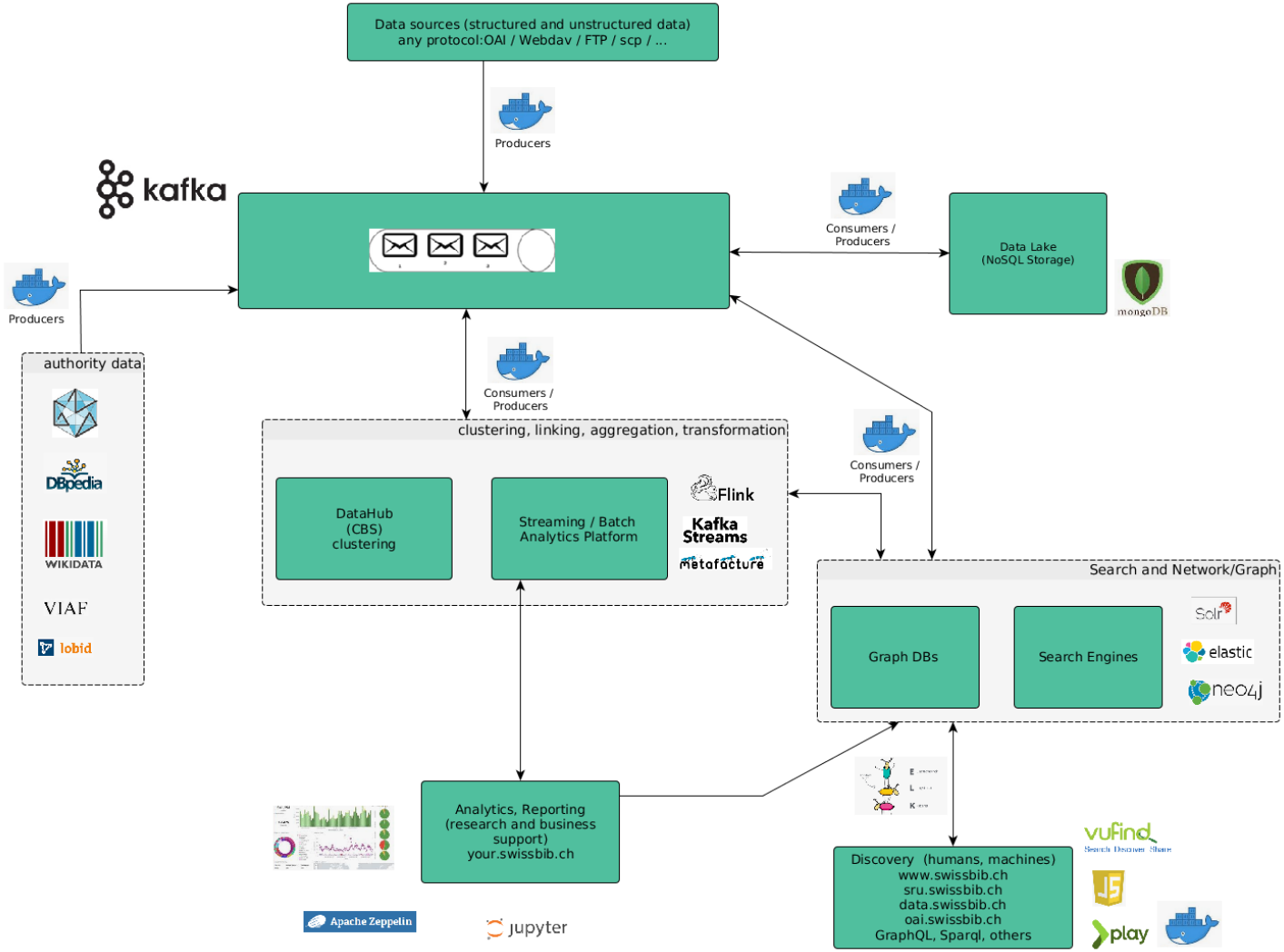
Evaluate

- Beide Aufnahmen haben Index-Einträge:
  Die Einträge der einen Aufnahme sind in denen der anderen Aufnahme komplett enthalten → merge ist möglich
  Beide Aufnahmen enthalten abweichende Einträge → kein merge





swissbib hinter den Kulissen



Fragen und Diskussion

Vielen Dank!

Günter Hipler
Systemarchitekt
swissbib
Universitätsbibliothek Basel

guenter.hipler@unibas.ch

Silvia Witzig
Metadatenspezialistin
swissbib
Universitätbibliothek Basel

silvia.witzig@unibas.ch

Mehr über swissbib

- www.swissbib.ch
- GitHub
  https://github.com/swissbib
  https://github.com/linked-swissbib
- linked.swissbib.ch: Beta RESTful API
  http://data.swissbib.ch/
- Blog
  https://swissbib.blogspot.com/

swissbib hinter den Kulissen