Systemarchitektur der Schweizer Metadatenplattform auf der Basis von swissbib

Technisches Konzeptpapier

Version 0.9

13. Februar 2017

Inhaltsverzeichnis

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Zusammenfassung und Einleitung

Im «Konzeptpapier swissbib - Die Entwicklungsperspektiven der Schweizer Metadatenplattform» (nachfolgend mit KP abgekürzt) werden die Positionierung und die strategischen Perspektiven von swissbib in der Schweizer Bibliothekslandschaft sowie die Entwicklung einer Metadatenplattform und darauf aufbauenden Services für aktuell rund 960 Institutionen in der Schweiz beschrieben und diskutiert. Die offene Schichtenarchitektur der «Lösung swissbib» und ihre austauschbaren Komponenten, die über frei zugängliche Schnittstellen miteinander kommunizieren, wird umrissen (siehe KP, Kapitel 2.3).

Die Idee von swissbib besteht darin, dass es bei einer Metadatenplattform nicht nur um die blosse Speicherung von Metadaten in einer Datenbank geht, sondern auch um die flexible und weitergehende Verarbeitung dieser Metadaten, auch durch die Institutionen selber. Die Ergebnisse solcher Verarbeitungen sind Grundlage für spezifische Benutzerservices der Institutionen oder können anderen Diensten für deren Zwecke weitergegeben werden[1].

Die hier vorgestellte Systemarchitektur sowie die mögliche Roadmap zu ihrer Umsetzung hat ihren wesentlichen Ursprung in den Vorarbeiten mit der FH Bern zum Thema Big-Data. Das «Technische Konzeptpapier» ergänzt das KP und beschreibt, ausgehend von der Problemstellung im Kapitel 2, wie die Architektur einer zukünftigen swissbib Metadatenplattform aufgebaut sein sollte (Kapitel 3).

Neben dieser Integration ist der Einsatz der Bausteine einer Schweizer Metadatenplattform auch in anderen Szenarien gut denkbar. Alle Softwarekomponenten ausser dem aktuellen CBS (*Central Bibliographic System*) Daten-Hub sind offen und frei verwendbar sowie kombinierbar.

Das bereits vorhandene und noch weiter auszubauende Know-How sollte innerhalb der Schweizer Wissenschaftsgemeinschaft ausgetauscht und vernetzt werden.

1 Siehe Kap. 2.3 des KP «Alleinstellungsmerkmale swissbib»

2. Problemstellung

Die Problemstellung, welche eine Schweizer Metadatenplattform zu lösen anstrebt, kann mit der folgenden Abbildung dargestellt werden:

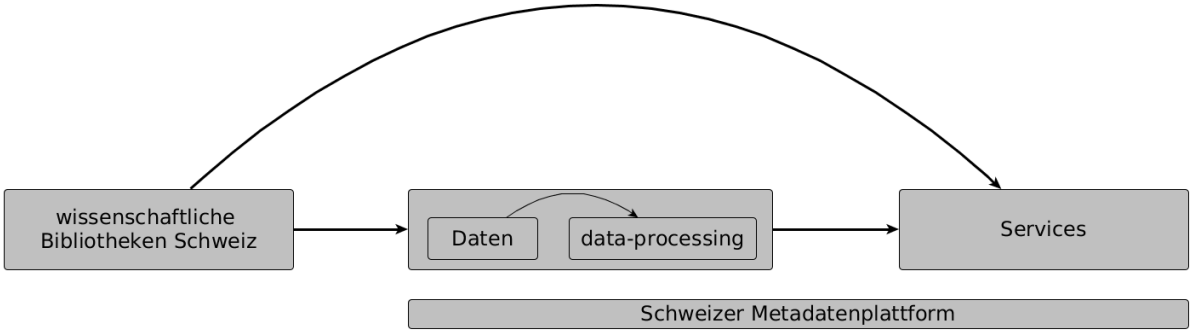

Abbildung 1: Problemstellung «Schweizer Metadatenplattform»

Erschliessen, Verarbeiten, Zusammenführen und Aggregieren von Metadaten sind Aufgaben wissenschaftlicher Bibliotheken. Diese Aufgaben erfolgen idealerweise automatisiert mit Hilfe der Datenverarbeitung. Die verarbeiteten Daten sind die Grundlage für Services wie die Suchoberfläche, Datenanalyse, -visualisierung, -bereinigung und Datenstrukturierung. Beispielsweise durch Visualisierung und Ausspielen von Ergebnissen an Benutzer oder offene, vernetzte Daten (Linked Open Data, LOD). Die Services sind Kundenwünsche aus dem Forschungsumfeld und stets evolutionär oder neuartig. Dies bedeutet für die Gestaltung der Systemarchitektur, dass sie für Daten und Dienste ausgelegt sein muss - auch für solche, die noch kommen werden. Zudem muss die Systemarchitektur grosse Datenmengen in kurzer Zeit verarbeiten können und eine geringe Latenz bei komplexen Abfragen und bei der Analyse unterschiedlicher Informationstypen haben. Die Plattform swissbib erfüllt heute schon die genannten Kriterien. Mehr noch: Die Systemarchitektur kann für die Verarbeitung unterschiedlicher Daten erweitert werden.

Die Schweizer Metadatenplattform besteht aus den Komponenten Datenverarbeitung und Services (siehe Abbildung). Für den Ausbau der Systemarchitektur „swissbib“ zur Schweizer Metadatenplattform sind folgende Punkte wichtig:

1. **Abholen der Daten und Datenvorbereitung**
   Die Metadaten der kulturelle Institutionen und Gedächtnisorganisationen (GLAM) werden regelmässig eingesammelt und aufbereitet. swissbib sammelt und verarbeitet heute Metadaten aus 960 Institutionen. Verarbeitungsprozesse können im Idealfall für neue Institutionen verwendet werden.
2. **Datenverarbeitung**
   Die Metadaten werden bereinigt, zusammengeführt, referenziert und mit zusätzlichen Daten angereichert. Diese Datenverarbeitung fällt je nach Dienst und Empfängertyp unterschiedlich aus. Bei der Datenverarbeitung werden grosse, komplexe, schwach strukturierte und sich schnell ändernde Datenmengen bewältigt. Das Zusammenführen von Daten aus unterschiedlichen Quellen, die Integration,

Datensicherheit, Datenintegrität und Datenqualität sind zusätzliche Anforderungen. Diese Anforderungen werden aktuell mit Big-Data-Techniken gelöst. Zum Einsatz kommen NoSQL, In-Memory und Hadoop. Für die Suchdienste und analytischen Anwendungen (Online Analytical Processing) werden die Daten spaltenorientiert aufbereitet.

3. **Services für Menschen und Maschinen**
   Services können die aufbereiteten Metadaten über Schnittstellen abrufen. Benutzer der Services sind Menschen oder Maschinen.

3. Architektur der Schweizer Metadatenplattform

Seit sieben Jahren verarbeitet swissbib rund 35 Millionen bibliographische Daten. Die Daten werden Diensten in der Wissenschaft, Verwaltung und dem allgemeinem Publikum zur Verfügung gestellt. Der Datenverarbeitungskern besteht aus Komponenten, die miteinander verbunden sind. Die offene Systemarchitektur erlaubt es, die Komponenten auszutauschen und neue Verfahren zur Verarbeitung der Daten einzubinden. Angesichts der im KP aufgezählten Herausforderungen wird man für eine Metadatenplattform und den darauf aufbauenden Services klären, wie die modernen und heute relativ leicht einsetzbare Methoden und Verfahren der Big-Data-Technik sinnvoll in eine zukünftige Architektur zu integrieren sind[2].

Die nachstehende Abbildung gibt einen Überblick über die wichtigsten Bausteine der schweizer Metadatenplattform. Die Grün gekennzeichneten Komponenten sind bereits jetzt im produktiven Einsatz von swissbib. Die gelb gekennzeichneten Komponenten wurden bereits in Prototypen erprobt.

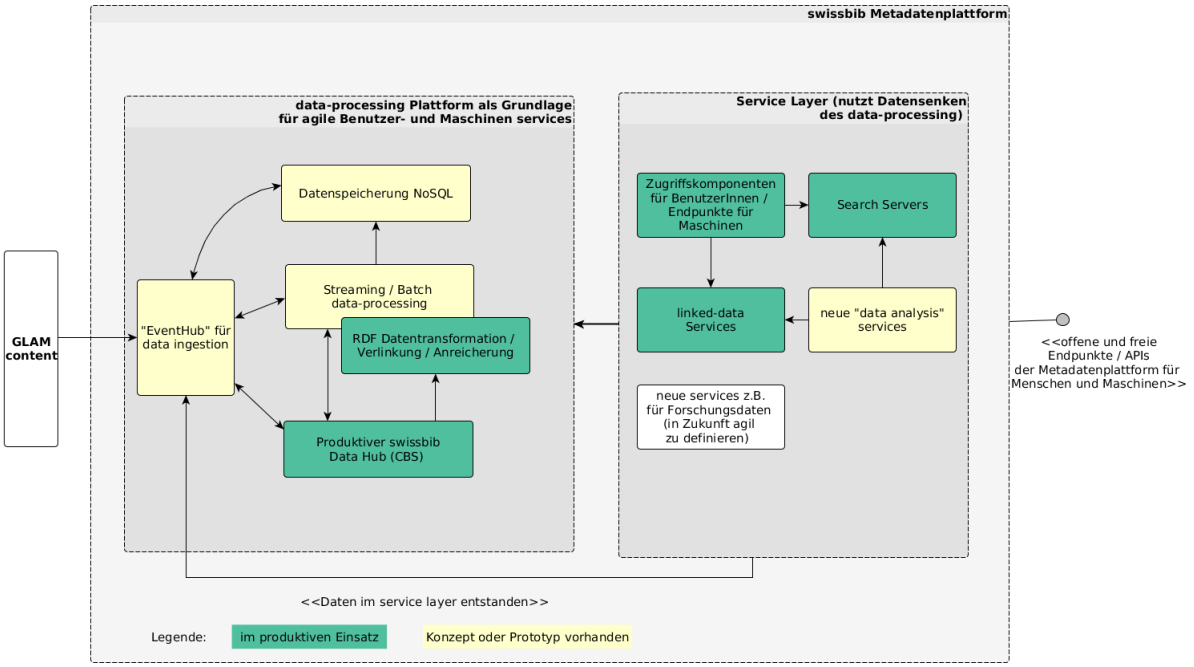

Abbildung 2: Architekturübersicht swissbib Metadatenplattform

Die Schweizer Metadatenplattform basiert auf den Metadaten der Schweizer GLAM-Institutionen (*GLAM Content*) und ist in zwei Bereiche unterteilt: Datenverarbeitung und Service-Schicht. Die Plattform verfügt über offene Endpunkte und Schnittstellen. Für die Bereinigung, Gruppierung und Anreicherung der Daten wird derzeit CBS von OCLC eingesetzt. Dieser Metadaten-Hub soll durch die neuen Verfahren und Techniken aus der Big Data Technologie ergänzt werden. Ziel kann es sein, die in den langen Jahren der Entwicklung dieses Daten-Hubs gesammelten Erfahrungen schrittweise mit neuen

2 Das Thema, wie aus Daten Informationen gewonnen und diese in modernen Services eingesetzt werden können, ist natürlich nicht nur für eine Metadatenplattform aktuell. Siehe dazu z.B. den Bericht zu einer Veranstaltung des ICT Fokus (https://www.switch.ch/de/stories/big_science_data/).

Technologien abzubilden. Am Ende dieser Entwicklung kann ein Ersatz der bestehenden Technolgieen in einem evolutionären Prozess möglich sein.

4. Bausteine der Datenverarbeitung

4.1. Event-Hub

Ein Event-Hub ist ein hyperskalierter Dienst für die Erfassung von Daten, der Millionen von Ereignissen sammelt, transformiert und speichert. Diese Streamingplattform bietet eine niedrige Latenz und konfigurierbare Aufbewahrungszeiten, wodurch riesige Mengen an Daten eingespeist werden können. Über die Semantik „Veröffentlichen/Abonnieren“ können die Daten verschiedener Anwendungen gelesen werden. Der Event-Hub ermöglicht die Integration und Kommunikation von Softwarekomponenten innerhalb eines Systems. Die Kohäsion und Kopplung von Softwarekomponenten der swissbib-Plattform können durch den Einsatz eines Event-Hubs vereinfacht werden. Der Event-Hub wird als Baustein für die Bereitstellung von Rohdaten aus Fremdsystemen eingesetzt. Gleichzeitig ist er als Datengeber oder Datennehmer integraler Bestandteil der Datenverarbeitung.

Kommerzielle Event-Hubs werden heutzutage von Anbietern wie Amazon (Kinesis), Microsoft (Azure) oder Oracle (Oracle Cloud) bereitgestellt. Im Open-Source Bereich ist das Apache Projekt Kafka[3] mit Abstand die bekannteste und am meisten genutzte Lösung[4]. Für Kafka existieren Konnektoren zur Anbindung von Streaming- oder Speichersystemen. Ein *Event-Hub* kann neben Metadaten weitere Datentypen (zum Beispiel Transaktionsdaten des Servicelayer) aufnehmen, so dass diese in der Datenverarbeitung mit den Metadaten verknüpft werden können. Diese Verknüpfung ermöglicht die Bereitstellung weiterer Dienste.

Für den Event-Hub gilt:

- Der Event-Hub ist ein zentraler Bestandteil der Schweizer Metadatenplattform.
- Die Integration in swissbib ist machbar. Teile des bestehenden Datensammlers können als Datengeber oder Datennehmer wiederverwendet werden.
- Die bestehenden Arbeitsabläufe lassen sich durch den Einsatz des Event-Hubs vereinfachen und ermöglicht neue Verarbeitungsverfahren.
- Auch der bestehende swissbib Daten-Hub kann an einen Event-Hub angebunden werden.

4.2. Datenspeicherung

Die neuen Konzepte der Datenverarbeitung setzten voraus, dass eingehende Daten in ihrer Rohform abgelegt werden. Die Rohdaten dienen als Datenquelle und werden erst bei Gebrauch verarbeitet. Diese Vorgehensweise unterscheidet sich vom Data-Warehouse-Modell, bei welchem die Daten bereits vor dem ersten Schreibvorgang verarbeitet werden. Diese Festlegung auf eine Struktur erschwert oder macht Dienste unmöglich, die erst zu einem späteren Zeitpunkt entwickelt werden[5].

swissbib speichert alle vom Baustein „Content Collection“ gesammelten Rohdaten in einer NoSQL-Datenbank[6] (vgl. Abbildung 2 des KP), bevor sie weitergereicht werden.

3 Vgl. https://kafka.apache.org und https://www.confluent.io

4 Im neuen Logservice der IT-Dienste der Uni Basel beispielsweise wird Kafka eine prominente Rolle einnehmen.

5 Dieses Modell wird häufig auch mit den Begriffen «schema on write» versus «schema on read» umschrieben

6 Vgl. https://de.wikipedia.org/wiki/NoSQL

In Zukunft soll der Ladeprozess von Rohdaten mit Hilfe eines Event-Hubs in *HBase*[7] erfolgen. Parallel dazu werden die Daten strukturiert und für ein späteres Clustering gespeichert. Das Struktur wird ebenfalls in Hbase abgelegt. Dabei kommen Methoden zum Einsatz, die im Projekt *CultureGraph* der Deutschen Nationalbibliothek auf Basis von *Metafacture* erstmals im Jahre 2013 entwickelt und eingesetzt worden sind.[8] Diese Methode ist ein Beispiel dafür, wie Daten in unterschiedlichen Formen abgelegt werden können, wodurch die Wiederverwendung der Information für diverse Aufgabenstellungen zu einem späteren Zeitpunkten erleichtert oder erst möglich wird.

4.3. RDF-Transformation, Verlinkung und Anreicherung

Im Projekt linked.swissbib.ch wurde ein Baustein entwickelt, der die Daten in RDF transformiert und ebenso für die Vernetzung und die Anreicherung der Daten sorgt. Dieser Baustein wurde in swissbib integriert. Einen detaillierteren Überblick der swissbib-Architektur nach Integration der linked-swissbib-Komponenten und Aufnahme des Testbetriebs im November 2016 gibt die nachfolgende Abbildung.

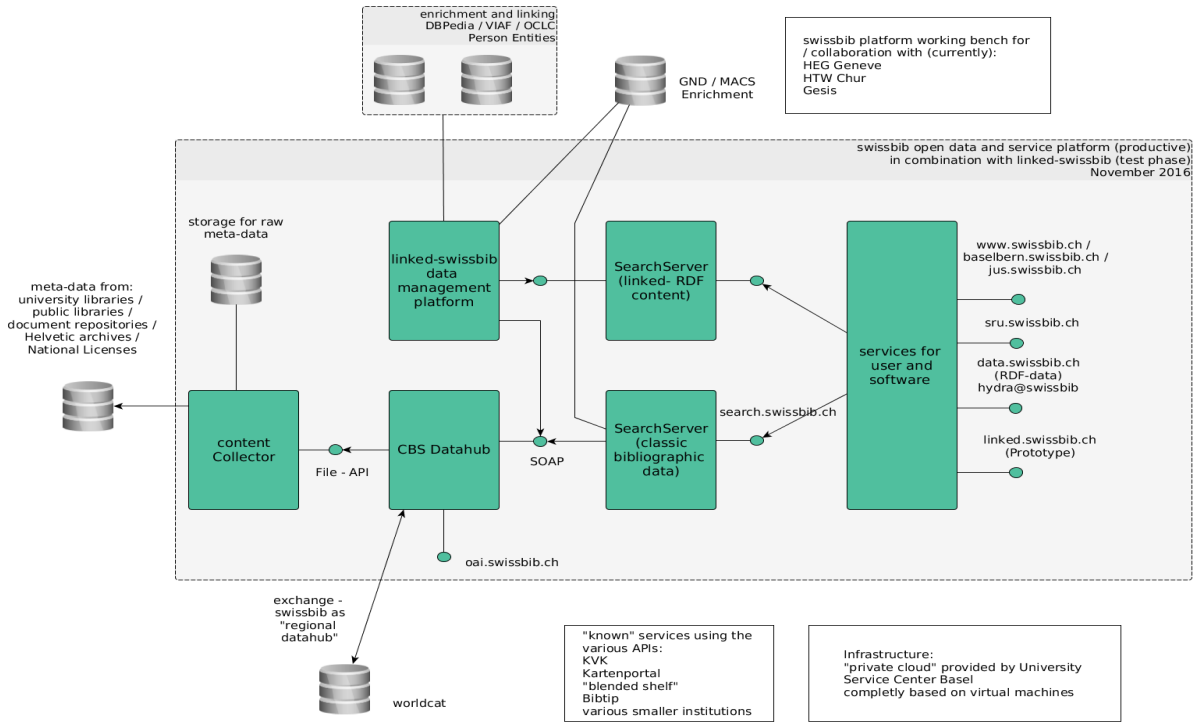

Abbildung 3: Integration der linked-swissbib Komponenten Ende 2016

7 https://hbase.apache.org/

8 https://github.com/culturegraph/metafacture-cluster

linked-swissbib.ch wurde mit dem Ziel entwickelt, eine skalierbare Infrastruktur zur Metadatenverarbeitung für swissbib zur Verfügung zu stellen. Die in RDF[9] überführten bibliographischen Daten aus CBS sowie die zusammengeführten Daten aus anderen Quellen werden über Schnittstellen bereitgestellt[10].

Die Skalierbarkeit wird durch die Software *Metafacture* der Deutschen Nationalbibliothek[11] garantiert. Für Verlinkungen zu VIAF[12] oder DBPedia[13] wird eine von der Gesis Köln[14] entwickelte Software verwendet[15]. ElasticSearch stellt die RDF-Daten über eine Schnittstelle bereit. Die HTW Chur entwickelte eine webbasierte Benutzerschnittstelle als integrierbares Modul innerhalb der bestehenden swissbib Präsentationskomponente.

Die Verwendung des RDF-Modells ist in der Schweizer Bibliothekswelt eine neue Form der Informationszusammenführung. Bislang wurden hierfür MARC-Daten verwendet. Weiterentwicklungs- und Verbesserungsmöglichkeiten sind:

- Mit der Einführung der GND[16] im Informationsverbund Deutschschweiz zu Beginn 2016 im Rahmen der Formalkatalogisierung wird die Qualität der Zusammenführung gesteigert.
- Im Projekt *linked.swissbib.ch* war es möglich, den Prototyp der neuen unveröffentlichten Schnittstelle „OCLC Person Entity“ zu evaluieren. Dieses Datenset sollte möglichst nach der Veröffentlichung für die Informationszusammenführung genutzt werden.
- Verbesserung der Arbeitsabläufe liegen in der Verzahnung mit neuartigen Streaming- und Batchverfahren aus dem Big-Data-Ecosystem (siehe Abbildung).

4.4. Streaming- und Batchverfahren

Die Schweizer Metadatenplattform muss für folgende Punkte ausgelegt sein (vgl. KP, Kapitel 2.4):

- Verarbeiten neuer Formen von Metadatentypen (weniger strukturiert bis unstrukturiert) ,
- aufnehmen von künftigen Informationsressourcen (z.B. Zeitschriftenartikel) und neuen Institutionen des GLAM Bereichs sowie
- versorgen der Benutzer mit zielgerichteten Informationen, was unter anderem die stärkere Verknüpfung der Metadaten mit Transaktions- oder Profildaten von Benutzern erfordert.

9 RDF ist ein Akronym für « Resource Description Framework », vgl. http://www.w3.org/RDF

10 Vgl. Projektziele der SUK-P2 Projektanträge «linked-swissbib.ch: the linked Swiss meta catalogue» sowie «swissbib – Schweizer Metakatalog (Betrieb und Ausbau) » im März 2014

11 Detailinformationen: swissbib.blogspot.ch/2014/06/considerations-for-development-of.html

12 VIAF ist ein Akronym für « Virtual International Authority File », Vgl. https://viaf.org

13 Vgl. http://wiki.dbpedia.org

14 Vgl. http://www.gesis.org

15 Detailinformationen: http://swissbib.blogspot.ch/2016/04/swissbib-data-goes-linked-teil-2.html

16 http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html

Die Datenmenge, aus der Informationen gewonnen werden, wird auch in Zukunft zunehmen und die Möglichkeiten der Datenanalyse müssen im Vergleich zu heute stark erweitert werden.

Die Skalierbarkeit des Clusterings grösserer Datenmengen wurde im Bibliotheksumfeld im Projekt *CultureGraph* der Deutschen Nationalbibliothek erstmals 2013 untersucht. Es wurde versucht, die Daten aller deutschsprachigen Verbünde zu clustern. Dieses Ziel entspricht in Teilen den Verfahren des bisherigen swissbib Daten-Hubs. Hierzu wird die Softwarekomponente *Metafacture* eingesetzt, welche auf das MapReduce-Verfahren von *Hadoop* aufsetzt. Der Durchsatz kann mit Hilfe des Streaming-Verfahrens gesteigert werden. Als Ergebnis lässt sich zusammenfassen:

- Die Verfahren und Algorithmen von 2013 sind heute noch gültig.
- Für die Datenanalyse sollte das Streaming-Verfahren eingesetzt werden[17].

Im Projekt *linked.swissbib.ch* wurde dies mit Hilfe eines Spark-Clusters getestet. Der Spark-Cluster stellt die täglich im CBS-Daten-Hub neu zusammengestellten Gruppen von ähnlichen Aufnahmen als RDF-Struktur zusammen. Einmal täglich wird der Bestand mit rund 25 Millionen Aufnahmen auf Basis der Gruppennummern aggregiert. Anschliessend wird der Suchmaschinenindex aktualisiert. Dieser Vorgang benötigt auf einem kleinen Cluster weniger als 5 Minuten[18]. Neben dem Potential dieser Methode verdeutlich dies, dass in der Schweizer Metadatenplattform immer wieder alte Verfahren durch neue Techniken ersetzt werden und die Architektur für einen Austausch alter Komponenten durch neue ausgelegt sein muss.

4.5. Der Daten-Hub

Der seit Februar 2010 im produktiven Einsatz stehende Daten-Hub von swissbib beruht auf CBS von OCLC. Hier findet die Datenverarbeitung statt, welche wiederum die Grundlage der von swissbib bereitgestellten Dienstleistungen ist. Der Daten-Hub aggregiert die Datenbestände der Schweizer Bibliothekswelt, de-dupliziert und fasst das Ergebnis anschliessend zu Bündeln ähnlicher Aufnahmen zusammen. Ein Teil der aggregierten Daten (diejenigen des IDS) wird in den Datenpool *Worldcat* geladen.

Beim CBS handelt es sich um einen Verbundkatalog, der wird in grossen Verbünden für den Benutzerdialog eingesetzt wird. swissbib verwendet das CBS lediglich im Batchmodus. Dieser neue Modus wurde mit Interesse vom Gemeinsamen Bibliotheksverbund Göttingen (GBV) begutachtet und soll in Zukunft dafür eingesetzt werden, die Bestände öffentlicher Bibliotheken seines Verbundes zusammenzufassen. Das swissbib-Team hat hier Pionierarbeit geleistet, was zu hervorragenden Arbeitsbeziehungen mit dem GBV und mit der Deutschen Nationalbibliothek oder dem Bibliothekszentrum Baden-Württemberg (BSZ) geführt hat. Die Flexibilität des Systems erlaubt eine problemlose Umstellung des De-Duplizierens und des Clustering auf den Produktionsservern.

Das vom swissbib-Team entwickelte Verfahren zum Einsatz eines Daten-Hubs auf Basis von CBS wurde erst kürzlich von JISC, einer Non-Profit-Organisation des Vereinigten

17 Hier sind an erster Stelle Apache Flink (https://flink.apache.org) und Apache Spark (http://spark.apache.org) zu nennen

18 https://github.com/linked-swissbib/workConceptGenerator/blob/master/src/main/scala-2.11/org.swissbib.linked/Application.scala

Königreichs, mit dem Ziel gewählt, eine landesweite bibliographische Wissensdatenbank zu erstellen[19].

Der GBV hat einen weiteren interessanten Aspekt in der zukünftigen Zusammenarbeit und Integration von kommerziellen Cloudsystemen entwickelt. Institutionen des GBV, die für ihre Bibliotheksprozesse das System Alma von ExLibris einsetzen, katalogisieren ihre Bestände zwar mit einem Client dieses Systems übertragen die bibliographische Beschreibung der Ressource («Katalogisat») jedoch in jedem Fall 1:1 in das Metadatensystem (CBS) des Verbundes. Damit ist das Kernkonzept der unveränderten Rohdaten, wie es im Abschnitt 3.1.2 (Datenspeicherung NoSQL) beschrieben wurde, vollständig umgesetzt.

19 Vgl. https://www.jisc.ac.uk, https://www.oclc.org/en/news/releases/2017/201702sheffield.html, https://www.youtube.com/watch?v=XJ-v_GqIffw

5. Bausteine des Service-Layer

Der Service-Layer einer Schweizer Metadatenplattform enthält die Dienste, welche Institutionen auf Basis Ihrer Daten und deren Verarbeitung anbieten. Möchte die Schweiz im Forschungsumfeld auch in Zukunft Dienste anbieten können, benötigt sie Daten und eine Datenverarbeitung, welche Big-Data-Techniken verwendet. Neben der Datenverarbeitung sind offene Schnittstellen notwendig. Dadurch lässt sich eine lose Kopplung zwischen den Bausteinen des Service-Layers erreichen. Beispielsweise funktioniert der Zugriff auf Komponenten von Suchmaschinen nur durch die Verwendung offener Schnittstellen, so wie es von swissbib bereits heute umgesetzt wird.

In Zukunft werden vermehrt die im SUK-P2 Projekt aufgebaute Infrastruktur zur Informationsvernetzung auch in den Benutzerdiensten integriert werden. Neue Services zur Untersuchung von Daten oder explorativen Suche können neben den direkten Ergebnissen der Datenverarbeitung auch die Ergebnisse von Suchmaschinen oder Dienste für vernetzte Daten mit einbeziehen.

Neue Bausteine eines Dienstleistungsportfolios können durch deren Unabhängigkeit leicht hinzugefügt werden.

Es ist selbstverständlich, dass die Schnittstellen oder Endpunkte auch externen Diensten zur Verfügung stehen. In der Übersicht der Abbildung ist zur Vereinfachung deshalb nur ein einziger Zugriffspunkt angedeutet und mit «offene und freie Endpunkte» bezeichnet.

5.1. Zugriffskomponenten für Benutzerinnen / Endpunkte für Maschinen

Details zu Zugriffskomponenten wurden im KP bereits vertieft in den Abschnitten 2.4, 5.1 sowie 6 beschrieben.

Die aktuelle swissbib Plattform bietet diverse maschinelle Endpunkte an[20], welche von externen Diensten eingesetzt werden:

- Das Angebot des KVK[21] in Karlsruhe, der die verschiedenen Möglichkeiten der swissbib SRU[22] Schnittstelle für seine Zwecke einsetzt.
- Das Schweizer Kartenportal.
- Spezialisierte Institutionen wie die Hotelfachschule in Lausanne.
- Hybrid-Bookshelf der Firma Picibird in Berlin[23] welche die freie Suchmaschinenschnittstelle und die zugreifbaren Verfügbarkeitsinformationen von swissbib für ihr Angebot einsetzt.
- Einbindung der Informationsressourcen von swissbib in dem Service bibtip[24].

20 http:// sru.swissbib.ch, http://oai.swissbib.ch:20103/oai, http://data.swissbib.ch, http://search.swissbib.ch/solr/sb-biblio/select?q=*%3A*&wt=xml&indent=true (siehe Abbildung 3)

21 https://kvk.bibliothek.kit.edu

22 http://www.loc.gov/standards/sru/

23 http://www.hybridbookshelf.de

24 www.bibtip.com, http://aleph.unisg.ch/F?func=direct&local_base=PH&doc_number=000833116&fbt=603658703

- Seit Beginn 2017 ist die SRU Schnittstelle von swissbib in der Central Knowledge Base von ExLibris integriert. Die von swissbib bereitgestellten Daten stehen damit unter anderem allen Benutzern von *Alma* (ExLibris) weltweit zur Verfügung[25].
- Die Firma EBSCO[26] nutzt die SRU Schnittstelle für ein Dienstleistungsangebot an die Bibliothek der Hotelfachschule in Lausanne.

Solche Dienstleistungen auf Basis der vorhandenen Metadaten sind möglich, da diese Informationen relativ schnell, in der benötigten Form und offen bereitgestellt werden können.

5.2. Suchmaschine

Die Suchmaschine ist die Grundlage jedes Systems zur Informationsrückgewinnung. Das swissbib-Team hat langjährige Erfahrung mit unterschiedlichen Systemen in diesem Bereich.

- Kommerzielle Suchmaschine FAST, die swissbib bis ca. 2012 im Einsatz hatte. Eine der Stärken von FAST war das Angebot einer Vielzahl von sprachspezifischen Wörterbüchern zur Unterstützung mehrsprachiger Suchen und Terme in Indizes. Der Kauf der norwegischen Firma FAST durch Microsoft sowie die grossen Qualitätssprünge der freien Suchmaschinen Lucene und SOLR[27] veranlassten swissbib zum Umstieg auf diese Lösungen. Dies war durch die offene Schichtenarchitektur recht einfach zu bewerkstelligen.
- Für die Suche in den klassischen, auf dem MARC Format basierenden, Bibliotheksdaten wird seit mehr als 4 Jahren SOLR und Lucene eingesetzt. Die Schnittstelle ist weltweit zugänglich. (vgl. Abschnitt 3.2.1)
- Für vernetzte Daten wird Elasticsearch eingesetzt. Dieser basiert wie SOLR auf der Lucene-Bibliothek und bietet eine modernere Query-DSL[28]. Zudem unterstützt ElasticSearch Nested-Objects, was Vorteile bei der Behandlung von anonymen Quellen in RDF-Strukturen bietet.

Die Java-Bibliothek Lucene ist heute de facto Standard bei Suchmaschinen. Sie wird auch in kommerziellen Systemen wie Primo von ExLibris eingesetzt. Ein grosses Potential liegt in der Offenheit der Systeme und den damit gegebenen Möglichkeiten, Informationsspezialisten der GLAM-Institutionen an der Gestaltung der Informationsrückgewinnung aktiv und direkt auf Systemebene mitwirken zu lassen. Sechs Jahre produktiver Betrieb swissbib haben gezeigt, dass solche neuen Ansätze der Zusammenarbeit zwischen Personen in unterschiedlichen Rollen dringend erforderlich sind. Sie tragen zur Qualitätsverbesserung und stärkeren Fokussierung auf die Benutzerbedürfnisse bei.

Suchmaschinen werden in swissbib zusätzlich zur üblichen Suche für folgende Anwendungsszenarien verwendet:

25 Dies ist uns erst nachträglich bekannt gegeben worden. Die betreffende kurze Mitteilung von ExLibris: «As requested on your http://www.swissbib.org/wiki/index.php?title=SRU page, we are here to notify you that we have created a search configuration for SwissBIB based on the SRU details provided on your site.We are planning to release it in our January 2017 Central Knowledgebase release. This search configuration will be also available to our Alma users. »

26 Vgl. https://www.ebsco.com

27 Vgl. http://lucene.apache.org/solr

28 https://de.wikipedia.org/wiki/Dom%C3%A4nenspezifische_Sprache

- ElasticSearch bietet eine Lösung zur Sammlung, Analyse und Visualisierung von Metriken an[29]. Diese Lösung werden wir sowohl für Statistik- als auch Analysezwecke einsetzen.
- SOLR kann auch bei Big-Data eingesetzt. Hier soll insbesondere die Indexierung auf Basis eines verteilten HDFS Systems (s. Kap. 3.1.2) evaluiert werden.
- SOLR bietet eine freie Lösung zur Unterstützung der Traversierung von Graphen an. Im Umfeld von verlinkten Daten bietet dies Möglichkeiten, die verstärkt genutzt werden sollten.

5.3. Services für vernetzte Daten

Die Erstellung von neuen Services für vernetzte Daten im Bibliotheksumfeld auf Basis von swissbib wurde im Projekt *linked.swissbib.ch* realisiert. *linked.swissbib.ch* besteht aus drei Teilen:

1. Infrastruktur (Siehe 3.1.3)
2. Webandwendung mit Benutzeroberfläche. Dieser Teil wurde durch die HTW Chur vorrangig realisiert. Das Konzept wird unter der Adresse http://linked.swissbib.ch beschrieben. Mehr Details zum Konzept finden sich in einem Blogbeitrag[30].
3. Maschine-Maschine-Schnittstelle zur Nachbenutzung der verlinkten Strukturen. Die API basiert auf dem Protokoll mit dem Namen «Hydra – Hypermedia Driven Web API» [31] Diese Form eines Clients ist Teil einer aktuellen Entwicklung mit der Bezeichnung «Linked Data Fragments» und ist eine Möglichkeit, die Schwächen der Sparql- Server zu umgehen. Im November 2016 wurde zu diesem Thema ein «library science talk» an der Zentralbibliothek Zürich gegeben[32]. Sollten sich die produktiven Einsatzmöglichkeiten der Sparql- oder Graphenserver in der nächsten Zeit verbessern, können diese in swissbib innerhalb des Servicelayer integriert werden. Von der Gesis in Köln wurden Algorithmen zur Verlinkung mit den Datasets von DBPedia und VIAF entwickelt.

Weitere Details zu den Artefakten des linked.swissbib.ch Projekts können einem noch nicht veröffentlichten Artikel entnommen werden. Ein Entwurf kann auf Anfrage zugesandt werden. Ein abrufbarer Vortrag wurde an der swib 2016 gehalten[33].

Unsere nächsten Schritte im Bereich linked data: Die neuen Services noch stärker innerhalb der swissbib Plattform nutzen und weiterentwickeln. Das Potential der Services besser zu kommunizieren und damit für andere nutzbar zu machen.

5.4. Services für die Datenanalyse

Typische Benutzeranfragen an das swissbib-Team sind Fragen wie:

29 Elasticsearch / Logstash / Kibana (https://www.elastic.co/products)

30 http://swissbib.blogspot.ch/2016/05/swissbib-data-goes-linked-teil-3.html

31 http://www.hydra-cg.com/spec/latest/core/ - Details dazu aus Sicht swissbib folgen in einem Blogbeitrag

32 https://www.zb.uzh.ch/Medien/Ausbildung/library_science_talks_slides_verborgh-ruben_20161206.pdf

33 Vgl. http://swib.org/swib16/programme.html

- «Der Bestand unserer Bibliothek wird auf swissbib.ch abgebildet. Gibt es über Ihren Katalog eine Möglichkeit herauszufinden, welche oder wie viele Titel sind ausschliesslich in unserer Bibliothek vorhanden sind?»
- Bereitstellung von Informationen für das Umsignierungsprojekt einer Bibliothek auf Basis des kompletten swissbib Datenbestandes. Die anfragende Bibliothek war vor allem an Ergebnissen von bereits abgeschlossenen Prozessen anderer Bibliotheken interessiert, um damit den eigenen Arbeitsaufwand reduzieren zu können
- Forschende möchten eine computerunterstützte Sacherschliessung auf Basis der in *jusbib*[34] verwendeten Rechtsklassifikation analysieren. Die Verwendung dieser speziellen Rechtsklassifikation wird im spezialisierten jusbib-Service durch visuelle Komponenten vereinfacht[35].

Neben den informationswissenschaftlichen Themen können die Metadaten auch für andere Forschungsfelder interessant sein. Für diese Forschungsfelder können in Zukunft Schnittstellen und Services zur Verfügung gestellt werden.

swissbib verfügt über Methoden und Verfahren, mit denen die oben beschriebenen Anfragen bedient werden können. Hierzu zählen die Schnittstellen der Suchmaschinen und das Metafacture-Framework, dessen Transformationssprache leicht zu erlernen ist. Metafacture wurde zum Beispiel zur Ermittlung der Informationen für ein Umsignierungsprojekt eingesetzt.

Mehr Möglichkeiten bietet das Verfahren des Streaming-Processings (vgl. Kap. 3.1.4), mit denen auch die Datenressourcen untersucht werden können. Hierfür verwenden Forschende Werkzeuge wie Apache Zeppelin[36]. In *linked.swissbib.ch* wurden erste Schritte in diese Richtung unternommen.

Dies sind nur wenige Beispiele für neue Service-Möglichkeiten der «Data Analytics» auf Basis der Schweizer Metadatenplattform. Solche Services sind jedoch nur dann realisierbar, wenn die neue Plattform Methoden und Verfahren einer modernen Datenverarbeitung unterstützt, wie diese im Kap. 3.1 beschrieben wurden.

5.5. Zukünftige Services auf Basis der Schweizer Metadatenplattform

Die Anforderungen und Wünsche an eine Schweizer Metadatenplattform können sich schnell ändern und die Architektur muss sich daran orientieren. Die Hauptmerkmale der vorgestellten Architektur soll hier nochmals zusammengefasst werden:

- Eine entkoppelte Plattform zur Datenverarbeitung als Basis für die Dienste des Servicelayers.
- Ein Servicelayer dessen Bausteine frei kombinierbar sind und die Möglichkeiten der Datenverarbeitung nutzen kann. Es werden offene und freie Schnittstellen zur internen und externen Nutzung bereitgestellt.
- Die Datenverarbeitung besteht aus Bausteinen, die Methoden und Verfahren der Big-Data-Techniken unterstützen.

34 jusbib ist ein Derivat der swissbib-Suche, welche speziell für die Juristen entwickelt wurde, vgl. http://jus.swissbib.ch

35 Vgl. https://jus.swissbib.ch/Search/AdvancedClassification

36 Vgl. https://zeppelin.apache.org

- GLAM-Institutionen sind in der Lage, die Schweizer Metadatenplattform unabhängig und selbständig zu betreiben.