Handout zum Projekt „Resolving- und Lookup-Dienst für bibliothekarische Identifier in culturegraph.org“

Stand 21.11.2011

Inhalt

1 Formale Eckdaten 2
2 Ausgangslage 2
3 Projektziele und Umfang 2
3.1 Bezug zur Erstkatalogisierungs-ID (EKI) 3
3.2 Die Plattform *culturegraph.org* 3
4 Vorgehensweise und Planung 4
5 Aktueller Stand 5
5.1 Stand der technischen Entwicklung 5
5.1.1 Architektur 5
5.1.2 Verarbeitungskette 6
5.1.3 Implementierung und Performanz 7
5.2 Stand der Datenlieferungen und Import 8
5.3 Ergebnisse der Datenanalyse 8
5.3.1 Herausforderungen 8
5.3.2 Analyseergebnis 8
5.4 Stand der Prozessierung 8
5.5 Stand der Datenveröffentlichung 9
5.5.1 Ontologiedefinition 9
5.5.2 HTML-Oberfläche 9
6 Nächste Schritte 10
7 Fragen / Diskussionspunkte 11

1 Formale Eckdaten

Name: Resolving- und Lookup-Dienst für bibliothekarische Identifier in culturegraph.org

Laufzeit: 16.01.2011-28.02.2012 (ursprünglich geplantes Ende 31.12.2011)

Projektmittel: Eigenleistungen

Projektmitarbeiter:

- Daniel Schäfer (DNB) Projektleitung, Entwicklung (seit Mitte November in Elternzeit)
- Katja Mecklinger (DNB) – Stellvertretende Projektleitung, ÖA
- Markus Geipel (DNB) Leiter Architektur und Entwicklung
- Adrian Pohl (hbz) – ÖA, Ontologie
- Pascal Christoph (hbz) – Architektur
- Julia Hauser (DNB) - Ontologie
- Lars Svenson (DNB) - Ontologie
- Jürgen Kett (DNB) – Projektsteuerung, ÖA

2 Ausgangslage

Die Existenz vieler verschiedener bibliographischer Datenbanken von Verlagen, Bibliotheken und Bibliotheksverbünden führt dazu, dass für jede bibliographische Ressource eine Vielzahl von Beschreibungen und Identifikatoren existieren. Diese Vielfalt birgt eine Menge Probleme, deren Lösung im Rahmen der Migration bibliographischer Daten in das Semantic Web angegangen werden kann.

Betrachtet man diese Problematik nun im Hinblick auf aktuelle Linked-Data-Angebote, stellt sich der Sachverhalt folgendermaßen dar:
Die Publikation von Linked Data (LD) ist von Grund auf dezentral organisiert. Dies führt - nicht nur im Zusammenhang bibliographischer Daten - zu einer stetigen Vermehrung von Identifikatoren und Beschreibungen für ein und dieselbe Ressource. Dieser Problematik soll durch die Entwicklung eines kooperativen Identifikationssystems begegnet werden. Die Nutzung gemeinsamer, globaler Identifikatoren bietet optimale Voraussetzungen für die Verknüpfung webbasierter Informationsquellen von Gedächtnisinstitutionen. Eine solche Praxis garantiert letztlich eine stärkere Sichtbarkeit von Gedächtnisinstitutionen und ihren Beständen im World Wide Web.

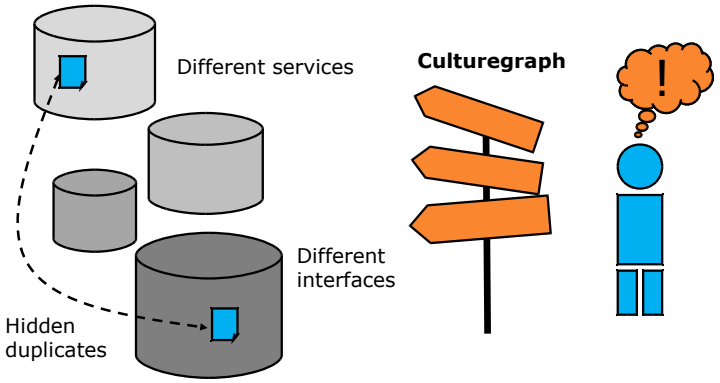
Die AG Kooperative Verbundanwendungen (AG KVA) hat den Projektpartnern DNB und hbz den Auftrag erteilt, eine Lösung für die Integration bibliographischer Daten aus verschiedenen Quellen zu entwickeln, insbesondere im Hinblick auf die zunehmende dezentrale Publikation von Linked Data durch verschiedene Institutionen.

3 Projektziele und Umfang

Die gemeinsamen Bestrebungen zielen zunächst darauf ab, zentrale URIs (im folgenden: CG- (d.h. Culturegraph-) URIs zu prägen, mit denen andere Identifier verknüpft werden. Ähnlich dem Dienst sameas.org sollen auf dieser Basis Identifier für identische bzw. verwandte Ressourcen gebündelt werden. Datensätze, die letztlich dieselbe oder eine sehr ähnliche Sache beschreiben (in der Regel dieselbe Manifestation oder zumindest eine eng verwandte Manifestation) werden in Gruppen gebündelt (Clustering) und erhalten eine gemeinsame CG-URI, die diese Gruppe identifiziert.

Der Dienst soll damit als wichtige zentrale LD-Drehscheibe etabliert werden und für die „Linked Library Cloud" eine ähnliche Funktion und Bedeutung bekommen, wie die DBPedia für die gesamte LD-Cloud: Er soll helfen, die bibliografischen Datensets möglichst eng zu verbinden und gleichzeitig den Zugang zu bibliografischen Daten der AGV und des Bibliothekswesens im Allgemeinen zu erleichtern. Die CG-URI bilden einen

zentralen Schlüssel zu den Datensets der Verbundpartner. Der LOD-Community bliebe damit erspart, jedes einzelne bibliografische Datenset einzusammeln, um diese Zusammenhänge selbst zu ermitteln. Das in der AGV organisierte Bibliothekswesen würde so als homogene, eng vernetzte Community im Web sichtbar.



Zentrales, für die Allgemeinheit sichtbares Ergebnis ist ein Webdienst auf der Plattform culturegraph.org. Dieser Dienst bietet eine Suchmaske um in den berechneten Titelgruppen zu suchen (Lookup) sowie einen Resolvingdienst: Über die Eingabe der CG-URI (oder eines alternativen eindeutigen Identifiers (Alias-URI) wird die damit assoziierte Titelgruppe zurückgegeben. Bestandteil der Darstellung einer Gruppe ist neben dem Verweis auf die einzelnen Gruppenmitglieder (also die Originalsätze), die identifizierenden Merkmale der Gruppe und eine Verweis auf das Matchingverfahren, das zur Bildung der Gruppe geführt hat.

Neben dieser HTML-GUI wird auch ein LD-Service angeboten. Das heißt, alle Titelgruppen, deren identifizierenden Merkmale und Verknüpfungen werden nach den Linked-Data-Prinzipien auch in einer RDF/XML-Repräsentation zur Verfügung gestellt[1].

3.1 Bezug zur Erstkatalogisierungs-ID (EKI)

Das Projekt verfolgt mit der Etablierung von verbundübergreifenden Identifiern für Titeldaten ein ähnliches Ziel, wie das EKI-Verfahren. Allerdings sind die Ansätze sehr unterschiedlich: Der Resolving- und Lookupdienst richtet sich zunächst primär an existierende Daten, die nun nachträglich miteinander in Bezug gebracht werden und gemeinsame Identifer zugewiesen bekommen. Hierzu wird anders als beim Ringtausch des EKI-Verfahrens ein zentraler Datenpool benötigt. Ein Folgeziel ist die Integration des Dienstes in das laufende EKI-Verfahren: Bevor ein neuer Datensatz angelegt wird, sollte die CG-Plattform auf gegebenenfalls bereits existierende Datensätze durchsucht werden. Im Idealfall wird jede neue EKI automatisch auch in CG registriert.

3.2 Die Plattform *culturegraph.org*

Das laufende Projekt ist nicht mit der Plattform culturegraph.org identisch. Vielmehr ist es ein Infrastrukturprojekt, dessen Ergebnisse auf der Plattform culturgraph.org veröffentlicht werden.

culturegraph.org ist eine Plattform für Dienste und Projekte rund um die Themen Datenvernetzung, Persistent Identifier (PI) und Linked Open Data für kulturelle Entitäten. Wichtige Kernmerkmale der Plattform sind:

- Linked-Data-Prinzipien
- Persistenz und Transparenz: verlässliche und persistente Referenzierbarkeit durch PIs, Änderungshistorie und Provenienzangabe.
- Offenheit: offene Nutzungslizenz, Open-Source-Entwicklung

Der Dienst wird zurzeit im Rahmen dieses initialen Unterprojektes aufgebaut. Die Plattform wird in Zukunft allerdings auch die Heimat weiterer Projekte und Dienste sein.

[1] Vgl. http://linkeddatabook.com/editions/1.0/

Diese Dienste werden sich dann vermehrt auch an die Kulturdomäne im Allgemeinen richten und auch Archive und Museen einbeziehen. Wir sind bestrebt, die Plattform auch im Umfeld der DDB anzubieten.

Bislang wird culturegraph.org nicht durch Drittmittel finanziert. Um die Plattform dennoch stetig auszubauen und Drittmittel einzuwerben, suchen wir daher Kooperationspartner und interessante Projektideen. Die Idee ist es, eine Heimat für Forschungs- und Entwicklungsprojekte rund um das Thema Datenvernetzung zu schaffen, deren Ergebnisse dann rasch in produktive Dienstleistungen überführt werden können. Wir sind der Überzeugung, dass es mit vereinten Kräften auf einer offenen Plattform, mit modernen Technologien gelingen kann, das Bibliothekswesen entscheidende Schritte voran zu bringen. Es ist nicht sinnvoll, dieses Feld den Softwareherstellern und Branchenriesen wie OCLC alleine zu überlassen, da dies uns wesentlicher Gestaltungsmöglichkeiten beraubt.

4 Vorgehensweise und Planung

Alle Titelsätze der AGV werden mittels Matchingalgorithmen auf einer zentralen Infrastruktur (Datalab & Hub) prozessiert und die Prozessierungsergebnisse zu Verfügung gestellt. Die Distribution geschieht auf zwei Ebenen: Die CG-URI und die zugehörigen CG-Cluster werden als Linked Open Data unter einer offenen Lizenz für die allgemeine Öffentlichkeit publiziert. Der Zugriff auf die Original-Daten und spezielle Analyseergebnisse (Fehler, Statistiken, etc.) ist nur für die Verbundpartner vorgesehen.

Dem zentralen Ansatz liegen die Thesen zugrunde, dass Matchingalgorithmen in Abhängigkeit vom vorhandenen Datenbestand stetig fortentwickelt werden müssen und, dass die Daten der Partner sich im Detail in wichtigen identifizierenden Eigenschaften unterscheiden. Die Heterogenität der Daten und geeignete Matchingeigenschaften / Regeln, lassen sich anhand eines zentralen Datenpools effizienter ermitteln und neue Verfahren können direkt ausprobiert werden. Damit das Datenlabor ein solch exploratives Vorgehen unterstützt, muss die zentrale Infrastruktur in der Lage sein, den Gesamtbestand innerhalb weniger Minuten/Stunden zu verarbeiten.

Grob lässt sich das Projekt in zwei Abschnitte unterteilen. Die erste Hälfte des Projektes von Januar bis Ende Juni 2011 beschäftigte sich vor allem mit dem Aufbau der Basis-Infrastruktur für den Resolving- und Lookup-Service und der Einarbeitung in die Thematik und in neue Technologien (wie zum Beispiel NoSQL-Datenbanken und Map-Reduce-Verfahren).

Die zweite Projekthälfte hat den Import der AGV-Daten, deren Analyse, die Entwicklung und Verfeinerung von Matchingalgorithmen und die Veröffentlichung der Ergebnisse (also der Titelgruppen) in der LOD-Cloud zum Schwerpunkt. Das Projekt endet mit der Produktivnahme des Datahubs. Insbesondere das Ende des Projektes muss durch eine intensive Öffentlichkeitsarbeit begleitet werden, damit der Dienst auch seinen Weg ins Bewusstsein der Anwender findet. Diese Aktivität reicht über das Projektende hinaus.

Die folgenden Hauptarbeitspakete wurden identifiziert:

1. **Projektleitung**
2. **Technische Entwicklung:** Konzeption und Aufbau der zentralen Infrastruktur des Datalab&Hub
3. **Datenlieferung und Import**: Auf einer Sitzung der AG KVA am 16.06.2011 einigte man sich auf die Lieferung aller Daten zu Titeln, die seit 1945 publiziert wurden
4. **Datenanalyse:** Analyse der importierten Daten / Besprechung der Ergebnisse mit der AG KVA
5. **Prozessierung (Matching):** Entwicklung von Matchingalgorithmen, Erstellen von Matchingberichten und Statistiken, Bewertung der Matchingergebnisse durch die AG KVA

6. **Veröffentlichung der Daten:** Repräsentation der Ergebnisse in einem geeigneten Datenmodell (Ontologie), Datenpräsentation in HTML, Publikation der Titelgruppen in der LOD-Cloud
7. **Öffentlichkeitarbeit**

5 Aktueller Stand

5.1 Stand der technischen Entwicklung

Es wurde eine neue Infrastruktur (Datalab & Datahub) aufgebaut, um den Import und die Analysen auf großen Datenbeständen effizient durchführen zu können und deren Ergebnisse zu präsentieren. Diese basiert auf den Technologien Apache hadoop/hbase/lucene.

5.1.1 Architektur

Soll der Resolving- und Lookupdienst mehr sein als eine Projektstudie, ein Prototyp, sondern ein operatives System welches die gesamten deutschen bibliographischen Daten mühelos verarbeitet, dann sind hohe technische Anforderungen zu erfüllen. Die deutschen Verbunddaten überschreiten nach unseren Schätzungen die Marke von 100 Millionen Datensätzen und liegen damit in Größenordnungen jenseits von ambitionierten Projekten wie zum Beispiel der DDB (ca. 2 Millionen Datensätze). Die wichtigste Maßgabe ist daher Skalierbarkeit. Ein weiterer wichtiger Aspekt ist die Verlässlichkeit der Datenbereitstellung für den Nutzer.

Unsere Architektur sieht daher eine Zweiteilung vor (vgl. Abb. 1): Ein Portal zur Präsentation der Ergebnisse steht einem Hadoop-Cluster gegenüber, welches als Datenlabor die eigentliche Verarbeitung übernimmt.

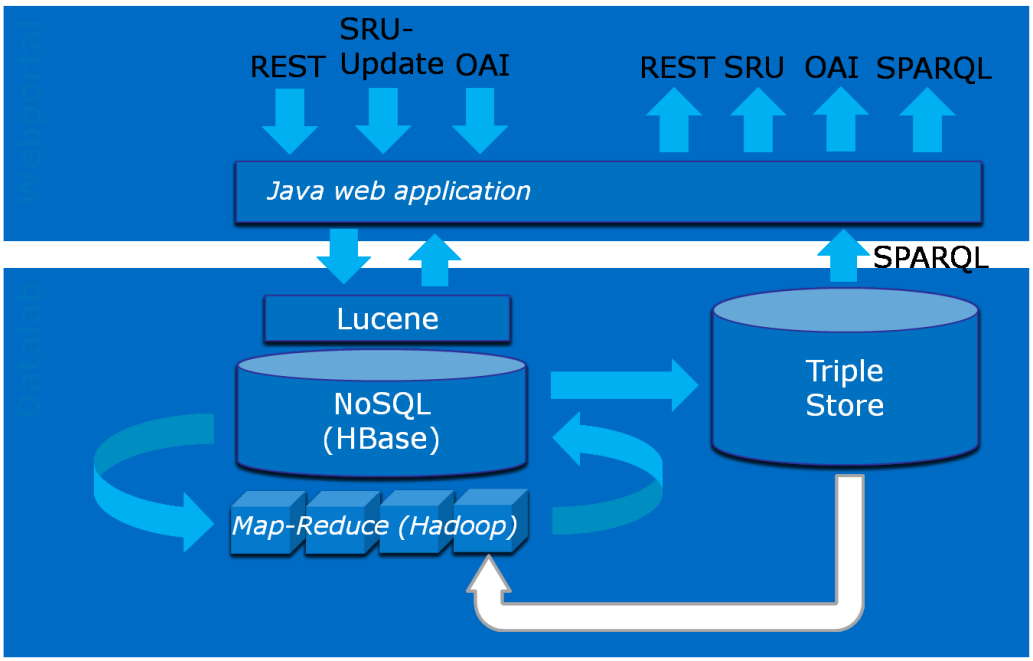


Abb. 1: Architektur Datalab&Hub

5.1.1.1 Portal

Das Portal bietet die Schnittstelle zu Benutzer und nützt ein Tomcat Cluster als Server Infrastruktur. Die Daten liegen in Form eines Lucene Indexes in einem NFS-Storage vor. Das Portal nutzt somit die erprobte und etablierte Infrastruktur der DNB. Sämtliche angebotenen Maschinenschnittstellen (REST, SRU etc.) werden auch vom Portal bereitgestellt.

5.1.1.2 Datenlabor

Abgeschirmt hinter dem Portal liegt das Herzstück des Systems: das Datenlabor. Um eine hochperformante und skalierbare Datenverarbeitung zu gewährleisten werden hier neueste Technologien wie Hadoop und HBase eingesetzt. Hadoop ermöglicht es, intensive Rechenprozesse mit großen Datenmengen auf eine beliebige Anzahl von Rechnern nach dem Map-Reduce Paradigma zu verteilen. Als Datenbanklösung wird HBase eingesetzt. HBase basiert auf Hadoop und unterliegt nicht den Skalierungsgrenzen herkömmlicher Relationaler Datenbanken. Die Infrastruktur des Datenlabors orientiert sich damit an der Infrastruktur mit der auch datenintensive Unternehmen wie Yahoo!, Google und Facebook operieren. Dem System beigestellt ist ein Triple-Store, welcher eine zusätzliche Datenhaltung in RDF bietet.

5.1.2 Verarbeitungskette

Ziel ist es im Sinne von FRBR Manifestations äquivalente Datensätze zu erkennen und zu bündeln. Wie bereits bei der Architektur gilt auch für den Algorithmus das Gebot der Skalierbarkeit. Bei Daten jenseits der Millionen-Marke schränkt sich die Auswahl der realistisch anwendbaren Algorithmen ein: Lediglich solche mit einer asymptotisch oberen Schranke von NlogN für die Laufzeit in Abhängigkeit von der Eingabegröße n kommen noch in Frage. Gegeben diese Herausforderung wurde folgende Verarbeitungskette implementiert:

1. **Umwandlung in Internformat und Normierung**
   Daten werden in den Formation MAB2 und MARC21 geliefert. Um die Daten geschlossen weiterverarbeiten zu können müssen diese in ein einheitliches Internformat gewandelt und normiert werden. Normierungen beinhalten zum Beispiel das bereinigen von ISBN Nummern.
   Die Umwandlung muss gleichzeitig effizient und flexibel konfigurierbar sein. Um diese Anforderung zu erfüllen wurde eine neue of XML basierende

Datentransformationssprache entwickelt welche sich gegenüber anderen Alternativen durch leichte Verständlichkeit und hohe Verarbeitungseffizienz auszeichnet.

2. **Einspielen in Datenbank**
   Die Daten werden nun in eine Non-SQL Datenbank eingespielt welche als Eingabe für diverse weitere Berechnungen dient. Sie ist jenseits des aktuellen Projekts nachnutzbar.
3. **Erzeugung identifizierender Eigenschaften (unique properties)**
   Für jeden Datensatz werden Eigenschaften berechnet, welche ihn eindeutig identifizieren – immer unter dem Vorbehalt dass dies die Datenlage zulässt. Die Eigenschaften werden wiederum in der bereits erwähnten Datentransformationssprache formuliert. Wichtig ist hierbei dass die Definition auch nicht lineare Kombinationen zulässt. Eine Beschränkung auf Linearkombinationen wie in anderen Systemen (DDB) der Fall macht verlässliche Dedublizierung unmöglich. Dafür ein Beispiel: In einer Linearkombination würde die Aussagekraft jeder elementaren Eigenschaft eines Datensatzes unabhängig von den restlichen definiert. Wie aussagekräftig ist jedoch die ISBN? Dies lässt sich nicht pauschal beantworten. Nur in Kombination mit anderen Eigenschaften - Auflagenummer, Monographie etc. – ist die ISBN zu bewerten und wiederspricht damit der linearen Kombinierbarkeit.
4. **Gruppierung der Daten anhand dieser Eigenschaften**
   Die Daten werden nun mittels einer Ordnungsrelation über den unique properties gruppiert. Ein Datensatz kann dabei in mehreren Gruppen auftauchen. Zum Beispiel könnte sich das Buch „Die Konstruktion der gesellschaftlichen Wirklichkeit" von John R. Searle sowohl in (EKI= DNB1008350362) als auch in (ISBN-AUFLAGE=9783518296059-1) wiederfinden.
5. **Verschmelzen von äquivalenten Gruppen**
   Wie in dem vorangegangenen Beispiel existieren Datensätze welche eine Brücke schlagen zwischen verschiedenen Äquivalenzgruppen. Gemäß der Symmetrie und Transitivität der Äquivalenzrelation müssen diese Gruppen verschmolzen werden. Dies geschieht in effizienter Weise indem über die bereits existierende Ordnungsrelation über den unique properties genutzt wird um eine präferierte Gruppen ID dynamisch festzulegen. Wird ein Datensatz in mehr als eine Gruppe geschrieben – und bildet dadurch eine Brücke – wird an alle nicht präferierten Gruppen ein Verweis auf die präferierte gehängt.
6. **Einspielen von Gruppen in Datenbank**
   In diesem Schritt werden die gefundenen Äquivalenzgruppen persistiert. Gruppen mit einem Verweis auf eine präferierte Gruppe werden in diesem Schritt Migriert so dass in der Datenbank bereits das endgültige Ergebnis des Dedublizierungsprozesses steht.
7. **Plausibilitätskontrollen**
   Wie bereits ausgeführt, können die Daten Fehler verschiedener Art enthalten. Bevor persistente ID vergeben werden können müssen daher die gefundenen Äquivalenzgruppen auf Plausibilität geprüft werden. Algorithmen dafür befinden sich in der Entwicklung.
8. **Vergabe von Persistenten IDs**
9. **Indizierung der Daten und Veröffentlichung**
   Um die Daten nun zu veröffentlichen und über Schnittstellen anzubieten werden diese mit Lucene indiziert. Der Index wird anschließen in das Webportal übertragen.

5.1.3 Implementierung und Performanz

Für das Portal wird die in der DNB bereits existierende Infrastruktur nachgenutzt. Die Arbeit am Portal konzentriert sich daher auf die Softwareentwicklung.

Im Falle des Datenlabors musste hingegen Pionierarbeit geleistet werden, da die verwendeten Technologien in der klassischen Bibliotheks IT bisher nicht vorhanden waren. Es musste daher erst ein Hadoop Cluster prototypisch aufgesetzt werden. Momentan läuft dieses provisorische Cluster stabil mit sehr guter Performanz: Bei ca. 80

Millionen momentan verfügbaren Daten ergeben sich folgende Laufzeiten für das Matching:

Schritte 1. – 2.: 240 min

Schritte 3. - 6.: 45 min

Für die weiteren Schritte sind noch keine endgültigen Aussagen möglich.

Ein Produktiv-Cluster mit entsprechender Hardware wird diese Zeiten voraussichtlich auf ein Viertel drücken. Selbst wenn neue komplexere Versionen des Matching-Algorithmus diesen Performanz-Gewinn ausgleichen sollten, zeichnet sich ab, dass das Datenlabor auch auf lange Sicht wachsenden Datenmengen problemlos bewältigen wird.

5.2 Stand der Datenlieferungen und Import

Eine Komplettlieferung der Titeldaten seit 1945 ist inzwischen von fast allen Verbundpartnern. Aktuell befinden sich rund 90 Mio. Datensätze im System. Es wurden effiziente Importverfahren neu entwickelt, die die Formate MAB, MARC und Pica unterstützen. Sowohl beim Importieren als auch Analysieren der Daten wurde eine sehr hohe Performanz erreicht.

5.3 Ergebnisse der Datenanalyse

5.3.1 Herausforderungen

Grundsätzlich gilt, dass nur Daten verarbeitet werden können die erstens existieren und zweitens qualitativen Mindestanforderungen genügen.

Dies bedingt, dass gelieferte Daten statistisch analysiert werden. Auch hierfür musste eine technische Infrastruktur geschaffen werden. Gegenwärtig können sowohl die gelieferten Rohdaten in Marc21 und Mab2 analysiert werden als auch deren Umsetzung in das Culturegraph Intern-Format. Desweiteren werden bestimmte Feldinhalte auf Gültigkeit geprüft. Zum Beispiel werden fehlerhafte ISBN erkannt und protokolliert. Andere Fehler wie die inkorrekte Verwendung von Formatstandards oder semantische Fehler wie die Vergabe der selben EKI an Datensätze, die sich auf unterschiedliche Werke beziehen sind oft schwerer zu identifizieren; oft auch erst im Nachhinein. Um dennoch einen hohen Qualitätsstandard der Berechnungsergebnisse aufrecht zu erhalten, sieht der Matching-Algorithmus einen Schritt „Plausibilitätskontrolle" vor, in welchem gefundene Äquivalenzgruppen vor der Veröffentlichung noch einen Filter durchlaufen (siehe Schritt 7 in der Prozesskette).

5.3.2 Analyseergebnis

Eine statistische Analyse der Daten, die die Häufigkeit der relevanten Identifer-Informationen ausweist, kann unter der Adresse https://wiki1.hbz-nrw.de/display/SEM/Datenlieferungen+der+Verbuende eingesehen werden. Die Qualitätskontrolle der Datenabzüge vor dem eigentlichen Import ergab einige Unregelmäßigkeiten, wie z.B. falsche ISBN-Längen (in über 10.000 Fällen), fehlende lokale Identifier, falsches Encoding.

Die Erstkatalogisierungs-ID in den MARC-Lieferungen entspricht nicht bei allen Abzügen der Spezifikation. Teilweise wurden diese von den Verbünden nachgebessert bzw. mussten beim Ingest der Daten korrigiert werden. Den einzelnen Verbünden wurde ein Feedback zu den Datenlieferungen gegeben, und die entsprechende Fehlerprotokolle wurden kommuniziert.

5.4 Stand der Prozessierung

Alle rund 90 Mio. Datensätze wurden mit prototypischen Matchingregeln verarbeitet. Gematcht wurde auf EKI-Gleichheit sowie Gleichheit einer Kombination aus normierter ISBN und Erscheinungsjahr im Falle einer Monographie.

Nach diesen Regeln konnten aus den Daten 50 Mio. identifizierende Eigenschaften (siehe Beschreibung der Prozesskette) generiert werden. Ca. 18 Mio. Datensätze konnten daraufhin gebündelt werden.

Bei einer Änderung der Matchingregeln ist ein Update der Bündelung innerhalb einer Stunde möglich. Der Einsatz komplexerer Algorithmen ist prinzipiell jederzeit möglich, erfordern allerdings eine bessere Datengrundlage (s. Abs. 5.3). Wir hoffen, dass es in Kooperation mit den Partnern möglich ist, diese Schritt für Schritt zu aufzubauen.

5.5 Stand der Datenveröffentlichung

5.5.1 Ontologiedefinition

Es wurde eine Ontologie entwickelt, um Äquivalenzklassen von bibliographischen Datensätzen in RDF darzustellen. Ein wichtiger Punkt war es hierbei soweit wie möglich bestehende Ontologien nachzunutzen. Die RDF Modellierung der bibliographischen Datensätzen selbst wird mit ähnlichen Vorhaben in der DNB koordiniert.

5.5.2 HTML-Oberfläche

Ein erster Prototyp der HTML-Oberfläche wurde entwickelt und ist online unter http://www.culturegraph.org/demonstrator einsehbar. Die Oberfläche basiert noch auf der ersten (inzwischen verworfenen) technischen Infrastruktur, die auf der NoSQL-Datenbank CouchDB aufbaute. Eine Migration der Oberfläche auf die aktuelle Infrastruktur und Anpassung der Funktionalität ist gerade in Arbeit.

Funktional bietet die Oberfläche aktuell:

- eine Resolvingfunktion (Beispiel: www.culturegraph.org/resource/BSZ273011103)
- eine Äquivalenzsuche: Zu einem Datensatz werden die anderen Mitglieder der Gruppe angezeigt. (Beispiel: http://www.culturegraph.org/match/BSZ273011103)
- Eine Vergleichssicht, die mehrere Metadatensätze hinsichtlich ihrer Eigenschaften gegenüberstellt und so verdeutlich, warum Datensätze einer Gruppe zugeordnet wurden bzw. warum nicht (vgl. Abb. 2). Diese Sicht soll der späteren Analyse von Stichproben dienen.
  (Beispiel: http://www.culturegraph.org/compare/BSZ273011103/DNB985584866)

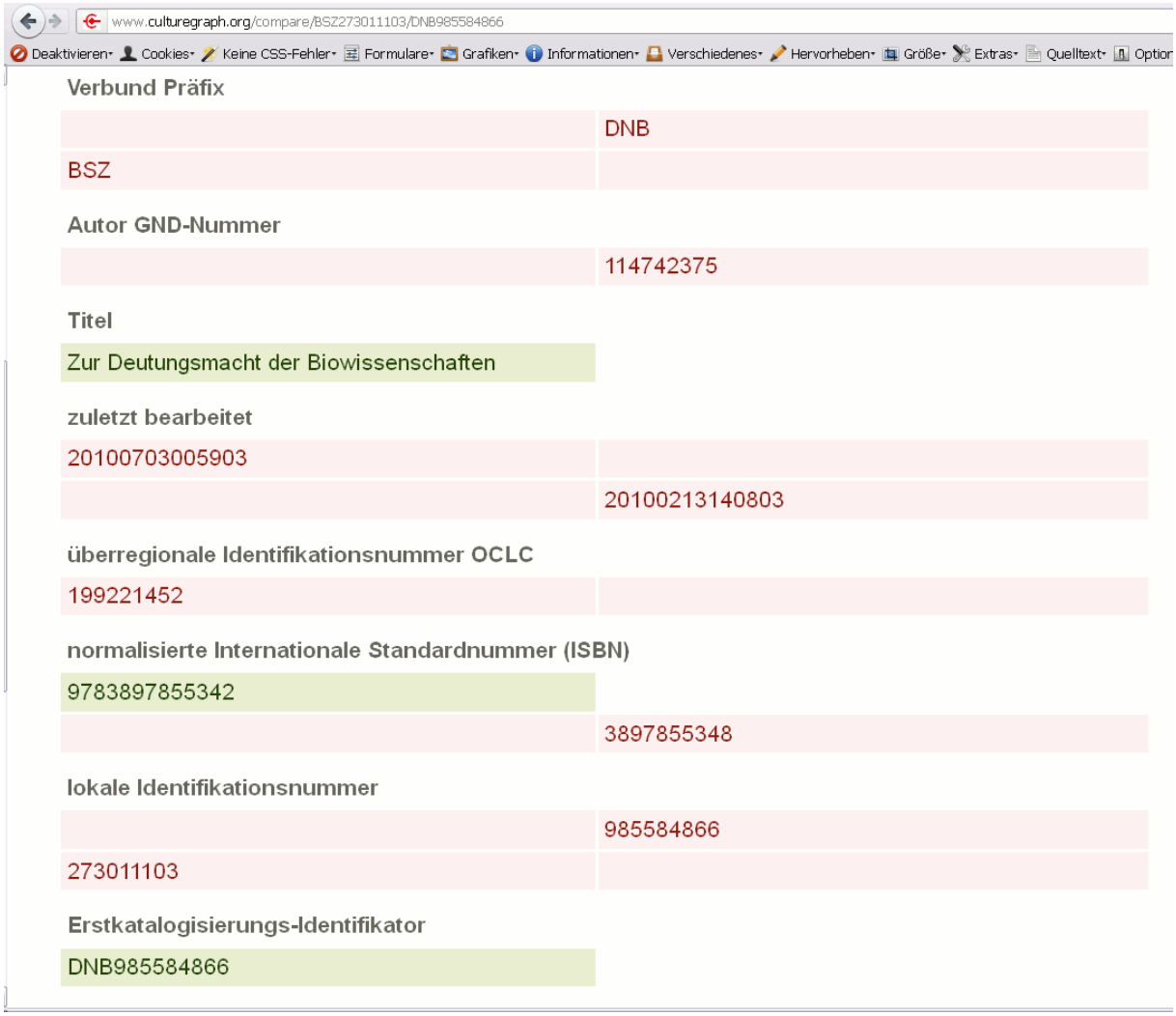


Abb. 2: Vergleichssicht des aktuellen Demonstrators (Bildausschnitt).
Übereinstimmende Eigenschaften sind grün hinterlegt, nicht übereinstimmende rot.

6 Nächste Schritte

Das Projekt soll bis 28.02.2012 abgeschlossen sein. Das entspricht einer Verlängerung der ursprünglichen Projektlaufzeit um rund 2 Monate. Gründe für dieser Verschiebung sind zum einen die Verschiebung der technischen Strategie am Ende der ersten Projekthälfte (CouchDB konnte die Anforderungen hinsichtlich Skalierbarkeit und Performance nicht erfüllen, weshalb eine Re-Implementierung auf Basis von Hadoop/hbase erfolgte) und zum anderen die hohen Aufwänden bei der Organisation der Datenlieferungen (verzögerte, teils fehlerhafte Datenlieferungen).

Wichtige noch zu erledigende Aufgaben sind:

- **Feedback und Konsequenzen aus den Analyseergebnissen (bis Ende Dezember):** Die Analyseergebnisse wurden den Verbundpartner mit der Bitte um Prüfung zugesandt. Kritisches Feedback wird nötig sein, um die Datenlieferungen bzw. die Interpretation der bisherigen Lieferungen zu verbessern. Nur so ist es möglich, die Datenbasis zu verbessern, damit sie sich in der Zukunft für komplexere Matchingalgorithmen eignet. Der Schritt ist wichtig für die Zukunft, aber nicht kritisch für den Projekterfolg, da auch auf Basis der aktuellen Lieferungen und Verabredungen (DNB-Titel sind prägend für alle Veröffentlichung vor EKI-Vergabe) bereits große Gruppen veröffentlicht werden können.

- **Matchingbericht (bis Mitte Januar):** Parallel wird für jeden Verbund ein Matchingbericht (inkl. Statistik) erstellt. Anhand dieses Berichts können sich die Partner ein Bild von der Überschneidung ihrer Daten mit den Daten anderer Partner machen und die Matching-Resultate stichprobenartig analysieren. Dieser Bericht wird automatisch erstellt und soll in Zukunft regelmäßig aktualisiert werden.
- **Fertigstellung der GUI (bis Ende Januar):** Die GUI wird aktuell auf die neue Infrastruktur migriert und im gleichen Zuge auch angepasst.
- **Fertigstellung der Infrastruktur (bis Ende Januar):** Im Bereich der Datensuche sind noch Restarbeiten zu erledigen.
- **Veröffentlichung der Daten als LOD (erster Betadienst Ende Januar, Prägung von persistenten URIs ab Ende Februar):** Die Titelgruppen werden in einer RDF-Repräsentation veröffentlicht. Das Mapping hierzu ist bereits prototypisch fertiggestellt. Allerdings wird an der Ontologiedefinition noch gearbeitet. Ein weiterer (noch völlig neuer Schritt) ist die Prägung von **persistenten** CG-URIs. Datensätze mit diesem Status werden nicht merh gelöscht und sind damit zitierfähig. Dieser Statuswechsel muss noch entwickelt werden. Er wird anfangs nur für eine kleinere Teilmenge der Daten gelten (z.B. für Titelgruppen mit EKI und Titelgruppen mit DNB-Titel vor EKI-Vergabe).
- **Erstellen eines Abschlussberichts**

7 Fragen / Diskussionspunkte

Fragen zu den Projektzielen

Anregungen an das Projekt / Kritik

Wünsche an die Plattform / weitere Projektideen

Kooperationsmöglichkeiten

Öffentlichkeitsarbeit: Wie bauen wir eine Community auf?