Metadaten Crosswalk

Umsetzung exemplarischer Crosswalks in der in AP 1.2 entwickelten Metadaten-Registry (M 3.4.1)

Version 1.0
Arbeitspaket 3.4
Verantwortlicher Partner GWDG, STI, BBAW, IEG

DARIAH-DE

Aufbau von Forschungsinfrastrukturen für die e-Humanities

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1110A bis M, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

Projekt: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Förderkennzeichen: 01UG1110A bis M
Laufzeit: März 2011 bis Februar 2014

Dokumentstatus: finale Fassung
Verfügbarkeit: DARIAH-DE-intern
Autoren:

Tibor Kálmán, GWDG
Daniel Kurzawe, GWDG
Harald Lordick, STI
Beata Mache, STI
Stefan Schmunk, IEG
Niels-Oliver Walkowski, BBAW

Revisionsverlauf:

Datum	Autor	Kommentare
02.05.2012	Daniel Kurzawe (et al.)	Konsolidierung der Inhalte in das Dokument
03.05.2012	Stefan Schmunk (et al.)	Überarbeitung u. Revision
04.05.2012	Harald Lordick, Beata Mache	Konsolidierung u. Revision
07.05.2012	Daniel Kurzawe, Stefan Schmunk, Niels-Oliver Walkowski	Schlussüberarbeitung
14.05.2012	Daniel Kurzawe, Stefan Schmunk (et al.)	Schlussredaktion

Inhaltsverzeichnis:

1. Einleitung

Aufbauend auf der *Schema Registry* und der daran anküpfenden *Crosswalk Registry,* beides im Meilenstein *Schema Registry M 1.2.1* beschrieben[1], wird in diesem Dokument eine exemplarische Verwendung dargestellt und verschiedene Perspektiven auf die Organisation von Forschungsdaten und deren Verwendung in der geistes- und kulturwissenschaftlichen Forschung[2] aufgezeigt.

Ziel ist es, einen *Crosswalk* - also das Überführen von Daten zwischen unterschiedlichen Formaten - zu ermöglichen, um auf diese Weise verschiedene Datensammlungen für eine maschinelle Weiterverarbeitung vorzubereiten, beispielsweise für eine kontextübergreifende Suche. Dafür sind sowohl verschiedene technische als auch organisatorische Schritte notwendig: Zu Beginn müssen die Quell- und Zielformate jeweils beschrieben werden. Darauf aufbauend werden einzelne Elemente miteinander verknüpft. Der Prozess des Verknüpfens verschiedener Metadatenschemata wird auch als *Mapping* bezeichnet. Durch die Verbindung zweier Formate entstehen Ontologien. Die so aufgebauten Ontologien können je Anwendungsfall und je Semantik durchaus unterschiedlich sein. Das Mapping und somit der Crosswalk sind aus diesem Grund oftmals anwendungsspezifisch.

Dieser Bericht soll einerseits die Weiterentwicklung und die weitere Konzeption der Schema Registry und der darin entwickelten Tools unterstützen. Andererseits sollen aus Perspektive der geisteswissenschaftlichen Forschung, und allem geisteswissenschaftlicher IT- bzw. DH-Experten - die zukünftigen Nutzerinnen und Nutzer[3] der Schema Registry - Entwicklungs-Bedürfnisse und -Ziele formuliert werden, sodass die weitere Entwicklung an geisteswissenschaftliche Forschungsprozesse rückgekoppelt wird, in disziplinärer und interdisziplinärer Perspektive. Insofern richtet sich dieser Meilenstein vorrangig an Entwickler und geisteswissenschaftliche IT-Experten, die als Multiplikatoren in geisteswissenschaftlichen Forschungsprojekten und -prozessen beratend tätig sind.

Die *Schema Registry* ist ein innerhalb von DARIAH-DE geschaffenes Tool, um sowohl Formate und Mappings zu erstellen, als auch Ontologien abzubilden und auf Dauer, aber auch dynamisch veränderbar, für geisteswissenschaftliche Forschung vorzuhalten. Vor allem stellt es ein nachnutzbares und universell einsetzbares Werkzeug dar.

Es existieren zum heutigen Zeitpunkt eine Vielzahl unterschiedlicher Metadatenformate und -standards, die in verschiedenen Fachdisziplinen und Kontexten Verwendung finden. So werden beispielsweise Marc21 und Allegro für bibliographische Informationen verwendet und ADeX und CIDOC-CRM für die Beschreibung von archäologischen Daten. In den Philologien wird der TEI-Header für die Beschreibung von Forschungsdaten angewendet; im webbasierten Kontext finden Dublin Core und METS/MODS oftmals Verwendung. Diese Aufzählung ließe sich beliebig erweitern. Jeder der genannten Standards wird von einer oder auch mehrerer Communities unterstützt und in teils verschiedenen Versionen verwendet. An diesen Beispielen ist erkennbar, dass zum jetzigen Zeitpunkt in verschiedenen (geisteswissenschaftlichen) Fachdisziplinen, in unterschiedlichen Forschungs- und

1 Siehe: https://dev2.dariah.eu/wiki/download/attachments/2295237/M1.2.1+Schema+Registry-2.pdf?version=1&modificationDate=1334923337614

2 Im folgenden werden die Geistes- und Kulturwissenschaften unter dem Begriff Geisteswissenschaften subsumiert.

3 Im weiteren Verlauf wird die Form für Nutzerinnen und Nutzer gewählt.

Anwendungsszenarien und sonstigen Kontexten eine große Anzahl von Metadatenstandards angewendet werden. Zugleich ist ein weiteres Faktum feststellbar. In geisteswissenschaftlichen Forschungsprojekten werden Forschungsdaten häufig mit gezielt hierzu entwickelten Metadatenschemata erfasst und projektspezifischen Klassifikationen beschrieben, die in der Regel auf einem individuellen Erkenntnis- und Forschungsinteresse eines einzelnen Forschers oder einer Forschergruppe basieren.[4]

Aufgrund der skizzierten Diversität gibt es faktisch keine generischen Standards, welche in jedem Szenario ihre Anwendung finden könnten. Und selbst bei der Verwendung von generischeren Standards bleiben die erwähnten semantischen Probleme. Eine Entwicklung, die aufgrund unterschiedlicher geisteswissenschaftlicher Forschungs- und Daten-Erfassungsszenarien auch nicht wünschenswert wäre. Die Schema Registry ist deshalb ein universell einsetzbares und individuell programmierbares Tool, das Geisteswissenschaftlern und Entwicklern, aber auch Bibliothekaren und Datenspezialisten ermöglicht, Daten unterschiedlichster Formate zu übersetzen, um so eine interdisziplinäre Interoperabilität und eine transdisziplinäre Nachnutzung zu ermöglichen.

Der aktuelle Entwicklungsstand der Schema Registry findet sich unter der folgenden Adresse: http://demo2.dariah.eu:8080/federator/

Die Verwaltung und Entwicklung wird mittels Jira koordiniert. Siehe hierzu: https://dev.dariah.eu/jira/browse/SCHEREG

2. Möglichkeiten der Crosswalk-Erstellung

2.1. Perspektive

Wie bereits beschrieben, werden Forschungsdaten in geisteswissenschaftlichen Forschungsprojekten oftmals nicht mit standardisierten Metadatenschemata erfasst, sondern in den meisten Fällen mit Projekt- bzw. Forschungsfragenspezifischen Metadaten beschrieben. Dies liegt an einer Vielzahl von Gründen. Einerseits sind viele Geisteswissenschaftler mit der Anwendung aktueller Metadatenstandards nicht vertraut, da diese weder im Rahmen des Studiums, der Promotion noch in gesonderten Weiterbildungsveranstaltungen - sieht man von dem breiten Angebot an Workshops zu X-Technologien ab - gezielt vermittelt werden. Andererseits wird der Erfassung und der Beschreibung von Forschungsdaten in den meisten geisteswissenschaftlichen Disziplinen bislang noch keine besondere Bedeutung zugemessen, da das primäre Interesse der Forscher auf der Bearbeitung ihres Forschungsbereichs liegt. Zudem wurden für diese Arbeitsschritte in Forschungs- und Projektanträgen oftmals keine finanziellen und personellen Ressourcen veranschlagt bzw. beantragt, wobei sich allerdings durch eine strukturelle Änderung der Förderpolitik in Deutschland derzeit ein grundlegender Wandel abzeichnet. Deshalb erfolgt eine Anreicherung von geisteswissenschaftlichen Forschungsdaten mit standardisierten Metadatenschemata, eine Grundvoraussetzung für eine Interoperabilität und disziplinäre und erst recht interdisziplinäre Nachnutzung, in den wenigsten Fällen. Aus diesen Gründen sind eine Vielzahl von geisteswissenschaftlichen Forschungsdaten-Collections nicht standardisiert erfasst - auf eine Diskussion über die eigentlichen Datenformate kann an dieser Stelle nicht eingegangen werden, aber es ist zu vermuten, dass dort eine vergleichbare heterogene Lage anzutreffen ist - und deshalb muss

4 Siehe: http://xkcd.com/927/

diesem Umstand bei der Entwicklung der DARIAH Schema-Registry Rechnung getragen werden.
Andererseits sind geisteswissenschaftliche Forschungsgegenstände generell Objekte deren Semantik schwierig standardisierbar sind. Die Semantik - oder auch die symbolische Ordnung - in die diese Objekte eingebunden sind, ist hier nicht nur Mittel zum Zweck, um an die "eigentlichen" Forschungsgegenstände zu kommen, sie ist immer auch Bestandteil dessen, worüber Geisteswissenschaftler forschen. Heterogenität ist deshalb nicht einfach nur Folge mangelnden Bewusstseins von Geisteswissenschaftlern, sondern auch Folge originärer geisteswissenschaftlicher Forschung.
Die Schema-Registry und die dazugehörigen Tools, insbesondere die Mapping- und Crosswalk-Werkzeuge müssen die Möglichkeit bieten, zwischen beliebigen Formaten - so z.B. zwischen Nicht-Standards und Standards - zu mappen. Zudem ist davon auszugehen, dass auf Attribute-Ebene nicht in allen Fällen der gleiche semantische Inhalt vorzufinden ist. Dies soll an dieser Stelle anhand von Datumsinformationen exemplarisch verdeutlicht werden. So können Datumsangaben zwar standardisiert erfasst werden - z.B. DD.MM.YYYY oder YYYY.MM.DD usw. - aber in den entsprechenden Feldern wird zumeist nicht erfasst, welches Kalendersystem, also welcher Messbezug, diesen Angaben zugrunde liegt. So ist es beispielsweise möglich, dass diese Informationen auf dem Julianischen-, Gregorianischen-, Russischen- oder Französischen Revolutions-Kalender als Messgröße basieren, was bei einem Mapping zwar zu keinen informationstechnologischen Fehlern führt, aber aus geisteswissenschaftlicher Perspektive z.B. Suchergebnisse fehlerhaft ausgegeben würden. Vergleichbares ist bei nicht standardisierten Klassifikationen auf Attributebene denkbar, vor allem wenn diese in unterschiedlichen disziplinären Fachkontexten vergeben wurden. Es besteht also ein großer Bedarf, dass die DARIAH Schema-Registry diesen Sachverhalten Rechnung trägt und die technische Möglichkeit eröffnet, Crosswalks zwischen Nicht-Standards, deren Verwendung wohl als Regel und nicht als Ausnahme in geisteswissenschaftlichen Forschungsprojekten anzusehen ist, und standardisierten Metadatenschemata erstellen zu können. Zusammenfassen lässt sich als zentrale Forderung aus geisteswissenschaftlicher Perspektive folgendes: Die Schema-Registry muss die Möglichkeit bieten, beliebige Formate, ob Standard oder nicht, zu mappen.

Gerade die Entwicklung und Realisierung solch eines Tools ist für die Nachnutzung von Forschungsdaten von besonderer wenn nicht sogar von integraler Bedeutung. Nur dadurch kann gewährleistet werden, dass Forschungsdaten, die in verschiedenen Forschungskontexten erhoben bzw. von Bibliotheken und Archiven erfasst wurden, nachhaltig und interoperabel nachgenutzt werden können. Der Mehrwert für die Geisteswissenschaften liegt auf der Hand. Einerseits können Collections und Datenbestände, die bislang aufgrund ihrer nicht standardisierten Erfassung durch Metadaten, nur mit erheblichem technischen Know-How durchsucht und damit auch genutzt werden, nun durch die DARIAH-DE Entwicklungen der Schema-Registry verknüpft, maschinell lesbar und somit überhaupt für die Forschung nutzbar gemacht werden. Erst hierdurch wird eine transdisziplinäre Interoperabilität hergestellt. Andererseits verwenden verschiedene geisteswissenschaftliche Disziplinen unterschiedliche Standards für die Erfassung ihrer Forschungsdaten, obwohl sie semantisch und/oder technisch die gleichen Inhalte beschreiben. Die zentrale Aufgabe der DARIAH Schema-Registry wird es also sein, unterschiedlich beschriebene Collections miteinander zu vernetzen, Crosswalks zwischen diesen herzustellen, um so einen Mehrwert zu generieren.

3. Informationsverlust bei Crosswalks - Allgemeine Vorüberlegungen

Die Art der Informationsverluste bei Crosswalks - Mapping von Metadatenformaten - lässt sich im Allgemeinen in mehrere Kategorien unterteilen. Um dies zu verdeutlichen, soll zunächst beispielhaft von einem abstrakten Crosswalk zweier beliebiger Schemata - dem Quellformat A und dem Zielformat B - ausgegangen werden. Hierbei sind folgende Varianten erkennbar:

- (Variante 1) Das Quellformat Format A besitzt *kein korrespondierendes* Element in Format B - Fehlende Felder im Zielformat

Diese Variante ist die offensichtlichste. So ist etwa im Quellformat das Element <URL> vorhanden, welches im Zielformat B nicht vorliegt. Die Folge ist, dass Informationen während eines Mappings-Prozesses verloren gehen und deshalb im Zielformat nicht mehr repräsentiert werden können. Natürlich ist auch genau der umgekehrte Weg möglich:

- (Variante 2) Die Quelle bietet kein korrespondierendes Element zum Zielformat B - Fehlende Felder im Quellformat

Hierbei tritt ein anderer Effekt auf: Informationen werden verallgemeinert, etwa wenn eine *1 zu n* oder *0 zu n* Relation vorliegt. Diese Relationen lassen sich als Prädikate der Form: *element_bezeichnung(Inhalt)* interpretieren. Wobei der Bezeichner den Namen des Feldes und das Argument den Feldinhalt bezeichnet. Da jedes Feld aus dem Quell- und Zielformat singulär ausgezeichnet ist, können die Prädikate nur erstufiger Natur sein.

- (Variante 3) Mehrere Felder aus Format A verweisen auf Format B

Während bei den Varianten 1 und 2 offensichtlich bestehende Informationen verloren gehen und somit eine definitive qualitative Abwertung erfolgt, verlagert sich das Problem in der dritten Variante auf die semantische Ebene. Werden etwa die Felder <Tag> <Monat> <Jahr> in ein Feld <Datum> überführt, bleiben zwar alle Informationen erhalten, aber nur dann, wenn die Semantik von <Datum> geklärt ist. Deshalb ist es in diesem Fall zwingend notwendig, dass strenge Konventionen im Bezug zur Verwendung des gewählten Metadatenformats existieren. Dies widerspricht jedoch dem Gedanken, ein möglichst breit gefächertes Format zu wählen, welches eine uneingeschränkte Verwendung ermöglicht und dadurch für eine undurchschaubare Diversität an Formaten.

Neben diesen drei häufig anzutreffenden Mappingproblemen, die zu Informationsverlust führen können, können noch folgende, die hier exemplarisch genannt sind, auftreten:

1. Einige Metadatenformate geben die Möglichkeit, dass mehrere Informationen durch ein Trennzeichen (Komma, Semikolon usw.) separiert in einem Feld abgelegt werden können. Besteht diese Möglichkeit im Zielformat nicht, so muss dies beim Prozess des Mappings berücksichtigt werden.

2. Bei verschiedenen Metadatenformaten werden zwischen Kann- und Muss-Feldern (obligatorische Felder) unterschieden; also zwischen Feldern bei denen vom Nutzer beim

Anlegen des Metadatensatzes ein Inhalt hinterlegt werden muss oder nicht. Ein Mapping kann nun dazu führen, dass bestimmte im Zielformat normalerweise als "Muss-Felder" vorgesehenen Felder nicht mit Inhalten belegt sind.

3. Mapping zwischen Nicht-Metadatenstandard und Metadatenstandard - Eine Variante, die noch ausführlich in den folgenden Kapiteln behandelt wird.

4. Eine weitere Ebene des Informationsverlustes betrifft die generelle Interpretierbarkeit von Standards und Regeln in ihrer Anwendung. Die zuvor angesprochenen Informationsverlusttypen gehen von einer idealen regelgeleiteten Adaption eines Standards aus. Nun sind Regeln aber nie eindeutig und unterliegen immer einem nachgestellten Gebrauch in einer konkreten Situation. Während die zuvor vorgestellten Informationsverlusttypen im Mapping transparent gemacht werden können ist die hier beschriebene Ursache für Informationsverlust durch das Mapping selbst nicht in den Griff zu bekommen und fordert zu einem sensiblen Umgang mit einer neuen, mittels des Crosswalks durchgeführten, Datenaggregation auf.

4. Funktionalität und Eigenschaften des Software-Prototypen

4.1. Funktionalität und Graphical User Interface (GUI)

Der Prototyp der Schema / Crosswalk Registry basiert auf OpenII Harmony[5] und ist zur Zeit unter der Adresse

http://demo2.dariah.eu:8080/federator/

als Webservice zu erreichen. Eine Eingangsseite „DARIAH Federator“ bietet die Auswahl zwischen den Sprachen „Deutsch“ und „Englisch“, ein Menü „Schema Registry | Crosswalk Registry“ sowie zwei zum Menü funktionsgleiche Links auf diese Seiten.

4.1.1. Schema und Crosswalk Registry

Die Schema Registry erlaubt das Registrieren eines neuen Schemas durch Hochladen einer Struktur-Datei, das Bearbeiten der Metainformationen eines bestehenden Schemas sowie das Löschen eines Schemas. Unterstützt wird hierbei das XML-Schema-Format (= XML Schema Definition, XSD).

5 Open source information integration (II) tool suite, http://openii.sourceforge.net/

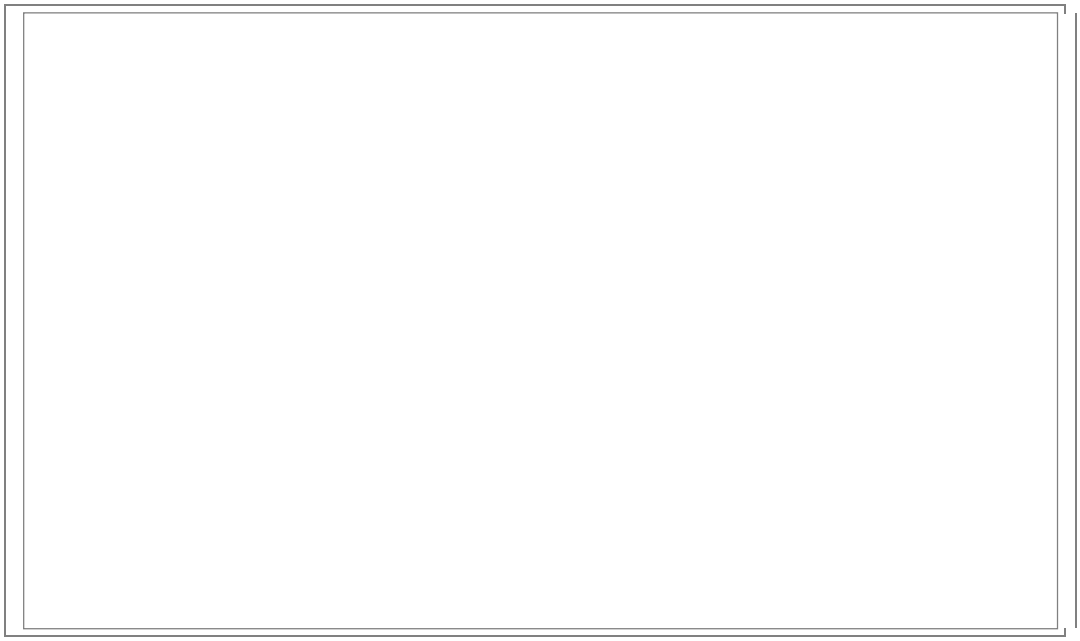

Abbildung 1: Screenshot Schema Registry

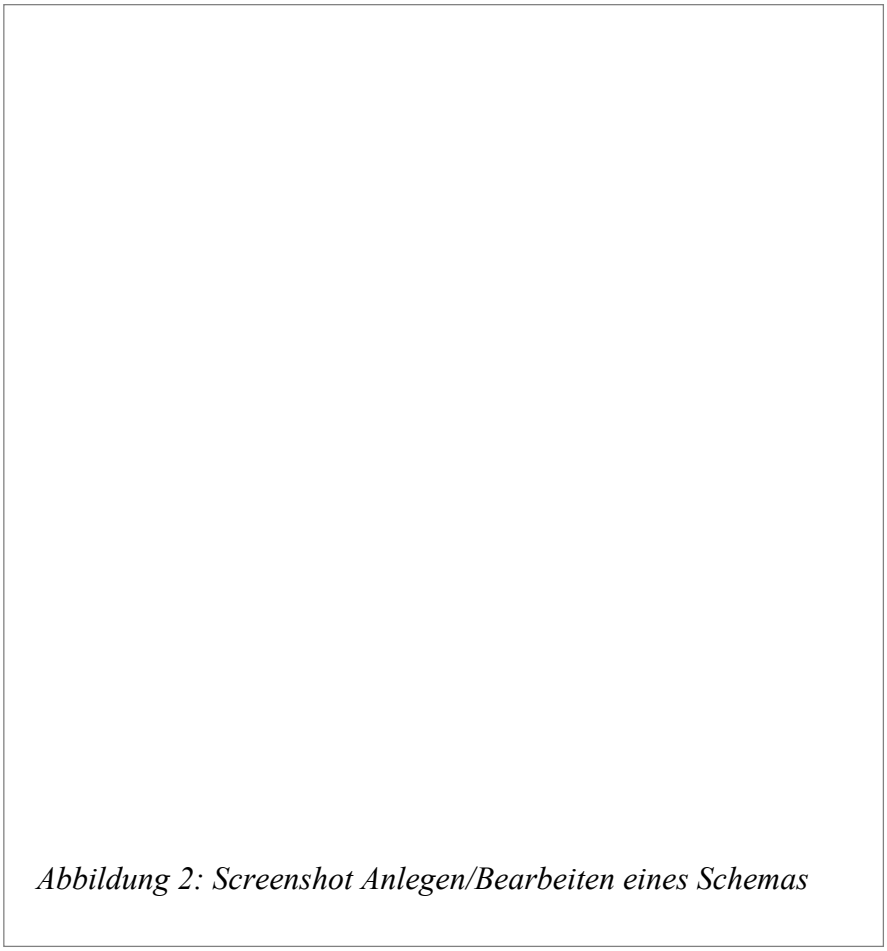

Abbildung 2: Screenshot Anlegen/Bearbeiten eines Schemas

Die Crosswalk Registry erlaubt dem User, durch Kombination von „Source Schema“ und „Target Schema“, einen Crosswalk anzulegen, diesen zu bearbeiten oder zu löschen. Neben

der Verwaltung der Schemata von Crosswalks lässt sich hier auch das Mapping (Menge der Relationen) zwischen Ausgangs- und Zielschema durchführen (siehe Screenshot „Crosswalk für Metadaten von Online-Retrodigitalisaten und ihr Mapping nach Dublin Core Simple“). Das geschieht auf einem dreigeteilten Bildschirm: links und rechts die beiden Schemata, dazwischen das Mapping. Hinzu kommen noch Regler für „Evidence“, „Depth“ und „Filters“. Per Drag-and-Drop lassen sich Links zwischen korrespondierenden Datenfeldern anlegen, die als Verbindungslinien visualisiert werden. Durch anklicken eines Links oder eines Datenfeldes mit der linken Maustaste öffnet sich ein Fenster mit der im Schema hinterlegten Beschreibung des Datenfeldes bzw. der Datenfelder. Ein Rechtsklick auf die den Link zwischen zwei Datenfeldern visualisierende Linie öffnet den Dialog „Annotations“ zur Bearbeitung des Links. Hier lässt sich eine Anmerkung eingeben, das „Confidence-Maß“ der Relation eingeben, die Verknüpfung löschen sowie eine algorithmische Funktion (Konvertierungsroutine) hinterlegen (Beispiel: Output=UPPER(DICOM_IMAGE)).

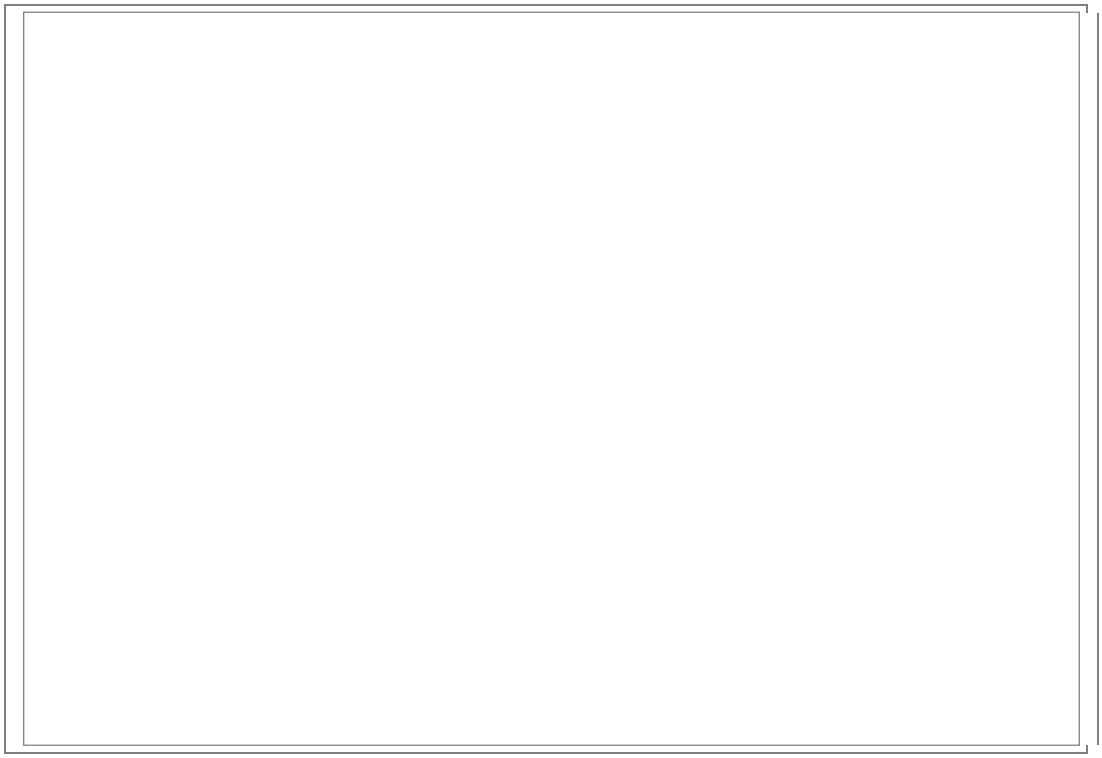

Abbildung 3: Screenshot Crosswalk Registry

4.1.2. Graphical User Interface

Das sachgerechte und übersichtliche Graphical User Interface des Prototypen wird der erforderlichen Funktionalität gerecht. Facetten des Interfaces finden sich in den abgebildeten Screenshots des vorliegenden Dokuments und in der Erläuterung der Funktionalität des Prototypen. Weitere Hinweise bringt der Abschnitt *Feature Requests.* Das GUI ist (noch) nicht selbsterklärend. Bisher nicht vorhanden ist eine Online-Dokumentation oder andere Formen der Erklärung von Zweck, Funktion und Bedienung, etwa durch eine Hilfeseite, durch Hilfe-Buttons oder Hover Infos.

4.1.3. Schnittstelle zum Wissenschaftler

Das Tool Schema / Crosswalk Registry ist insofern anspruchsvoll, als es nur gültige (valide) Schema-Dateien (XSD-Dateien) verarbeiten kann und diese voraussetzt. Ohne technische Vorkenntnisse im Umgang mit XML Schemata ist die manuelle Erstellung einer solchen Schema-Datei nicht möglich. Leichter fällt es, diverse XML-Editoren auf der Grundlage vorhandener XML-Daten zur Erstellung von diesen Daten entsprechenden Schema-Dateien einzusetzen. In der Mehrzahl der Fälle ist jedoch das Hochladen solcher Dateien in die Registry nicht gelungen. (In den FAQ zu OpenII wird diese Problematik auch erwähnt.) In der Evaluationsphase hat sich deshalb dieser der Registry vorausgehende Prozess als schwierig herausgestellt. Will man also nicht profunde XML-Schema-Kenntnisse erwarten, wäre die Bereitstellung oder zumindest Empfehlung eines Schema-Editors sinnvoll, um einen der Schema / Crosswalk Registry notwendig vorgelagerten kontrollierten und Erfolg versprechenden Workflow zu ermöglichen.

Darüber hinaus hat sich gerade bei der Auseinandersetzung mit der Schema / Crosswalk Registry herauskristallisiert, dass ein grundlegenderes Tool für das Modellieren von Daten (Entwerfen und Skizzieren von Datenstrukturen) – sozusagen ein Modelling-Tool – für den Geisteswissenschaftler hilfreich sein könnte, insbesondere auch als Einstieg und Anreiz für die Auseinandersetzung mit Schemata und Standards. Das sollte zweckentsprechend auch kollaboratives Arbeiten ermöglichen (Mapping als intellektuelle und kreative Herausforderung, als Diskussionsprozess). Gefragt wäre, so betrachtet, ein Mindmapping-Tool, konzipiert und optimiert für die Datenschemata-Erstellung. Ein pragmatischer Einstieg in eine solche Funktionalität könnte sein, wenn in der zweigeteilten Crosswalk-Registry in der Auseinandersetzung mit einem Standardschema auf der einen Seite eine freie Strukturierung von Grund auf neu („from scratch“) auf der gegenüberliegenden Seite möglich wäre. Ein Schema-Editor bzw. Datenmodellierungstool könnte auch ein Schritt zur Herstellung von Interoperabilität sein, wenn das Ausgangsmaterial in Form von CSV-, Access Format, oder etwa als Spreadsheets etc. vorliegt, das dann zwar eine technisch verarbeitbare Struktur hat, aber gegebenenfalls, insbesondere bei vorhandenen älteren Datensammlungen, durch kein ausreichend inhaltlich beschriebenes Datenmodell (wenn etwa zu Datenfeldern keine Feldbeschreibungen existieren) begleitet wird.

4.1.4. Ausbau und Erweiterung: Feature Request

- Beschreibung und Erklärung aller Funktionen (Anleitung, Handbuch, Video-Anleitung)
- Detailliertere Annotationsmöglichkeiten von Mappings
- Visualisierung der Qualität von Relationen (problematisch / unproblematisch …)
- Zusammenfassendes Linking (Wenn mehrere Quellfelder im Inhalt eines einzigen Zielfelds gespeichert werden sollen, siehe Crosswalk „Buchbildarchiv“)
- Versionsverwaltung: Speichern unterschiedlicher Versionen des Mappings
- Feedback bei Aktionen (Hochladen, Speichern, ...) bzw. Fehlermeldungen wie etwa OpenII sie zurückgibt („Failed to import schemas. PARSE_FAILURE: An element declaration with the given name, title, already exists in this scope.“)
- Hover Infos an den klickbaren Funktionsbereichen
- Mehr Felder zur Beschreibung von Schemata und Crosswalks (Provenance-Informationen wie Institution, Projekt, Schema URL

- Vererbung der ursprünglichen Meta-Information im Workflow. Beispiel: „DC:DATE war ursprünglich *bba:entstehungsJahrReproduktion*“. Das wäre für eine inhaltsreiche Suchfunktionalität wünschenswert
- Prinzipiell angelegte Sprachumschaltung durchgängig realisieren
- Downloadmöglichkeit hinterlegter Schemata und Mappings

4.1.5. Plattformen und Technik

Die Bereitstellung als Java-Applet zielt sinnvollerweise auf Plattform-Neutralität, hat allerdings in der Testphase des Prototypen nicht zu einer Plattformunabhängigkeit hinsichtlich des verwendeten Betriebssystems oder Browser führen können (Probleme mit Linux, mit Firefox, mit Internet-Explorer, mit nicht vorhanden, passenden oder deaktivierten Java-Plugins). Während der Testphase des Milestones verbreitete der Computerhersteller Apple Updates, die das Java-Browser-Plugin aus Sicherheitsgründen deaktivierten („Flashback-Trojaner“). Diese Erfahrung könnte eventuell für die Diskussion des Nachhaltigkeits-Konzepts von DARIAH-DE relevant sein.

Zurzeit offenen Punkte:

- Lauffähigkeit unter Linux (etwa Ubuntu 11.10) sowie diversen Betriebssystem- / Browserkombinationen
- Hochladen von selbst erstellten Schemata zumeist nicht möglich (das Schema erscheint kurz, verschwindet wieder, keine Fehlermeldung)
- Speichern von Mappings nicht möglich
- Fehlermeldung „nur zwischen 3 und 50 Zeichen erlaubt“ trotz entsprechenden Feldinhalts
- Zeichenkodierungsprobleme im Feld „Description“ der Schema Registry
- Crosswalk: Funktion der Depth- und Evidence-Regler
- Crosswalk: Mapping Function / nach Eingabe von Built-in Function plus Parameter [Schemafeld] geht die Built-in Function wieder verloren
- Crosswalk: Bearbeiten der „Description“ eines Datenfeldes ist möglich, geht aber verloren, sobald das Feld nicht mehr im Fokus ist

5. Crosswalk I / Mapping Metadaten-Standard / Metadaten-Standard

5.1. Ausgangssituation

Anhand des Personendaten-Repositoriums (PDR) der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) - das PDR ist einer der DARIAH-Demonstratoren - wurde die Schema-Registry evaluiert und getestet. Das Personendaten-Repositorium betont den quellenbasierten Hintergrund geisteswissenschaftlicher Forschung im Allgemeinen und biographischer Erschließungen im Besonderen. Es ist möglich für jede biographische Information zu einer Person entsprechende Quellen (Daten) für diese Information anzugeben. Während das PDR das Metadata Object Description Schema (MODS) zur Auszeichung von Referenzdaten verwendet, stellt z.B. der auf OPUS basierende eDoc-Server der BBAW dieses Format nicht zur Verfügung. Hier wird auf das sehr viel verbreitetere Dublin Core (DC)

zurückgegriffen. Wünschenswert wäre es für ein Projekt trotzdem alle Publikationen, die im Rahmen dieses Projekt erschienen sind, als mögliche Referenzen innerhalb des PDR zur Verfügung zu haben, ohne diese erneut manuell eingeben zu müssen. Das Beispiel ist insofern auch interessant, als dass es die Dimension der Schnittstelle ins Szenario mit einbezieht. Hierdurch wird deutlich, dass Daten in einem Schema häufig Bestandteil eines Wrapper-Schemas der API sein können.

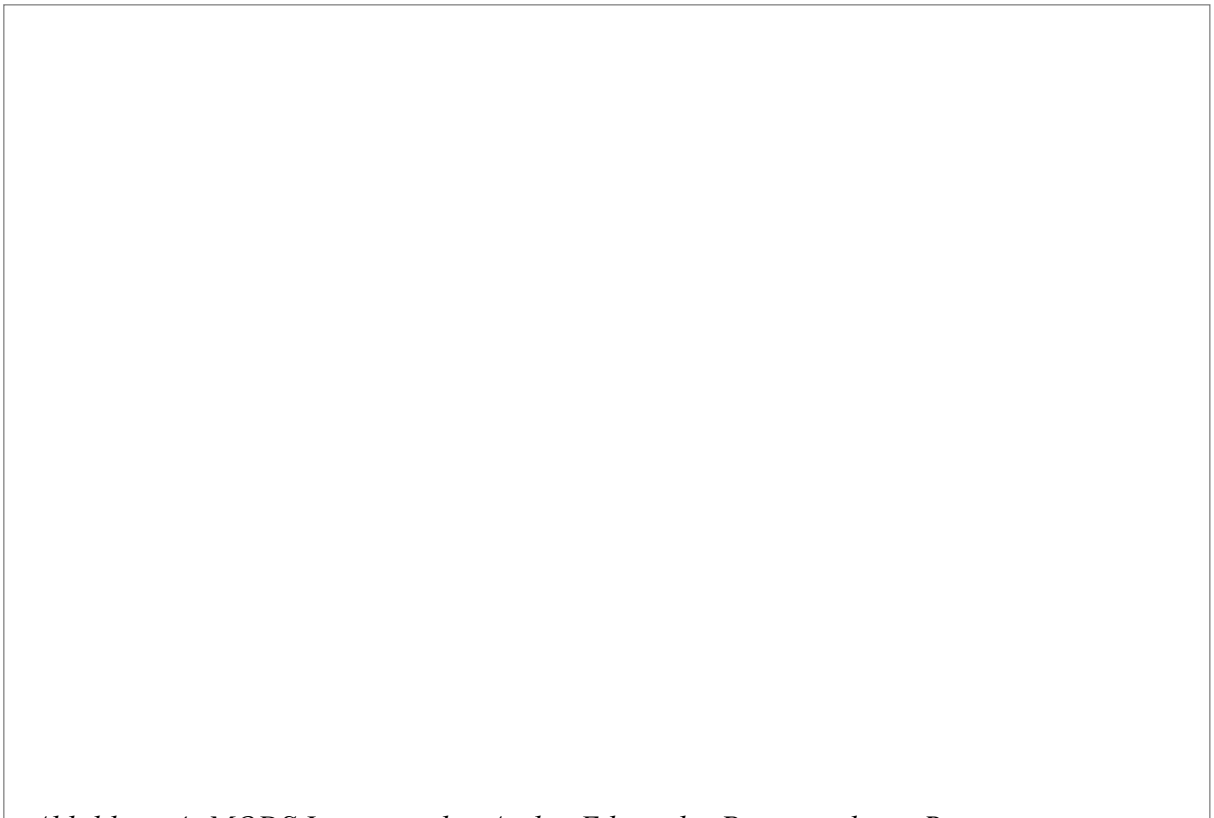

Abbildung 4: MODS Import in den Archiv Editor des Personendaten-Repositoriums

5.2. Szenario

OPUS besitzt eine OAI-PMH Schnittstelle. Über diese lassen sich Publikationen einer bestimmten Sammlung, z.B. eines Projektes auslesen und in Dublin Core ausgeben. Der Archiv Editor des PDR besitzt wiederum eine Importfunktion um Referenzobjekte in Dublin Core zu importieren. Der Worflow sieht daher folgende Schritte vor:

1. Harvesten der Quelldaten und Analyse von Quell- und Zielmodell
2. Registrieren der MODS und OAI-DC Schemata in der Schema Registry sowie des Crosswalks in der Crosswalk Registry
3. Konfiguration des Crosswalks zwischen den beiden Schemata in der Crosswalk Registry
4. Aufbereitung des Datenmaterials
5. Durchführen des Crosswalks
6. Import der Ergebnisdaten in MODS in das PDR

5.3. Analyse von Quell- und Zielmodell

Das vom PDR verwendete MODS ist eine sehr minimalistische Adaption des MODS Schemas. Für verschiedene Referenztypen gibt es XML-Templates auf der Basis von MODS auf die zurückgegriffen werden kann. Diese Templates sind nur Vorschläge, die im Projektkontext von Administratoren angepasst werden können, weshalb sie nicht in einem eigenen Schema erfasst werden. Dies hat zur Folge, dass das gesamte MODS Schema in das Mapping importiert werden muss, ohne dass es jedoch sinnvoll erscheint tatsächlich jedes Element zu mappen.

Abbildung 5: Reduzierte MODS Implementierung im Personendaten-Repositorium

OAI-DC ist ein Container für *simple dublin core* und importiert die entsprechenden Elemente in den eigenen Namespace. Darüber hinaus beschreibt OAI-DC eine spezifische Serialisierung für das Dublin Core Application Profile, die einer flachen XML-Struktur entspricht.

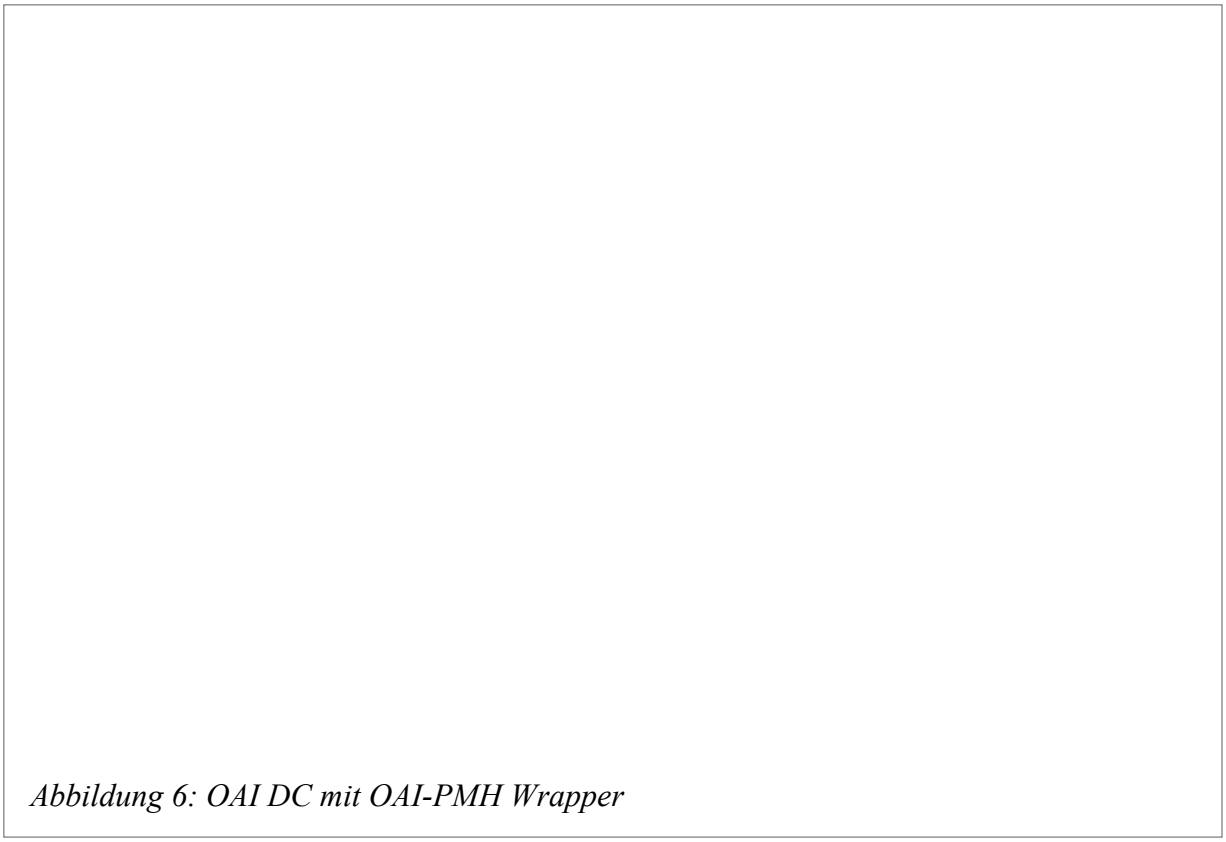

Abbildung 6: OAI DC mit OAI-PMH Wrapper

Das XML, welches die OAI-PMH Schnittstelle ausliefert ist kein reines OAI-DC XML. Vielmehr ist es eingebettet in ein XML, welches dem OAI-PMH Schema folgt. Dies fordert eine strategische Entscheidung in der Durchführung des Crosswalks. Entweder man erstellt ein Mapping von OAI-PMH-XML nach MODS oder man muss die XML-Ausgabe noch aufbereiten. Diese Aufbereitung kann z. B. so aussehen, dass man eine neue XML-Datei mit dem XML-Fragment ab dem <oai_dc:dc /> Knoten erzeugt. Da der Nutzen eines DC nach MODS Mappings größer zu sein scheint, als ein Mapping von OAI-PMH nach MODS (jenseits der sich ergebenden Problematik, dass in einem OAI-PMH-XML Dokument auch andere Schemata außer OAI-DC importierbar sind und das Mapping dadurch unendlich komplex oder unbrauchbar werden würde), entscheidet sich der vorliegende Beispielcrosswalk für diese Variante.

5.4. Registrieren der MODS und OAI-DC Schemata in der Schema Registry sowie des Crosswalks in der Crosswalk Registry

Der Import der Schemata erfolgt über die Oberfläche der Schema Registry. Einmal registrierte Schemata stehen in der Crosswalk Registry zur Verfügung. Unklar erscheint hier der Gebrauch des Feldes Autor mit dem entweder die registrierende Person oder der Ersteller des Schemas gemeint sein kann. Weitergehende Provenance Metadaten lassen sich nicht mit dem Schema assoziieren. Das Schema muss lokal auf der Festplatte vorliegen und kann so über den Dateibrowser angesteuert werden.

Registrieren des Mappings

Das zu erstellende Mapping wird ähnlich der Schemata registriert. Ausgangs- und Zielschema können nun aus einem Drop-down Menü ausgewählt werden. Erst nach der Registrierung erfolgt die Konfiguration des Mappings über das Beziehungs-Icon. Auch hier reduzieren sich die zur Verfügung stehenden Metadaten auf den Autor des Mappings. Weitere kontextualisierende Metadaten können nicht angelegt werden.

5.5. Konfiguration beider Schemata in der Crosswalk Registry, Aufbereitung des Datenmaterials und Durchführen des Crosswalks

Eine erste Schwierigkeit im vorliegenden Beispiel ergibt sich durch den unterschiedlichen Umgang mit Wurzelelementen in den beiden Schemata. Während MODS explizit zwischen einem MODS Dokument mit einem MODS-File und einem MODS Dokument mit mehreren MODS-Files unterscheidet, ist das Wurzelelement von OAI-DC gegenüber diesem Umstand agnostisch. MODS bietet die Wahl zwischen zwei verschiedenen Instanziierungen des Schemas. Deswegen müsste eines dieser zwei potenziellen Wurzelelemente von MODS auf das Wurzelelement von OAI-DC gemappt werden. Dies geht aber nicht, da OpenIIHarmony kein Mappen auf Wurzelelemente zulässt. Die Folge ist eine unsaubere Arbeit die die Flexibilität des Mappings reduziert. Im Prinzip müssten zwei Mappings für beide möglichen Instanziierungen von MODS angelegt werden.

Abbildung 7: MODS Schema mit der Option für zwei unterschiedliche Implementierungen, das in OpenIIHarmony zu Schwierigkeiten führt

Eine weitere Folge der Komplexität des MODS Schemas zeigt sich darin, dass OpenIIHarmony das Schema nur bis zur Ebene der zwei möglichen Wurzelelemente liest, aber nicht darüber hinaus. Konkret wird der selbstdefinierte @type *modsType* nicht interpretiert. Zur Durchführung des Beispielcrosswalks wurde daher das MODS Schema so angepasst, dass es ohne einen selbstdefinierten Element-@type und einer Element-group auskommt.
MODS Elemente werden häufig mehrfach innerhalb anderer Elemente referenziert, d. h. ein Element kann in verschiedenen Kontexten auftauchen. Wenn in OpenIIHarmony dieses Element innerhalb eines Elternelementes gemapped wird, so wird es implizit auch für jedes andere Elternelement in dem es auftaucht durchgeführt. Dieses Mapping ist nicht wieder lösbar. Dies kann zu Problemen führen, da der Kontext eines Elements bedeutungskonstituierend ist. Beim vorliegenden Mapping wirkt sich dieses Problem pragmatisch aus. Ein MODS Dokument kann mittels <relatedItem /> ein anderes Objekt angeben, welches in einer Beziehung zu diesem Element steht. Dieses Objekt kann bei MODS mittels einer kompletten MODS Beschreibung in das Dokument eingefügt werden. Beim Mapping von oai_dc:title nach mods:title wird dadurch automatisch auch auf das Sekundärobjekt referenziert, eine Information, die der Benutzer aber nicht in seinem Ergebnis XML vorfinden möchte und die nicht zum Primärobjekt gehört.
Prinzipiell ist es möglich für jede Beziehung eine Wahrscheinlichkeit zu spezifizieren mit der ein Knoten im Ausgangsschema einem Knoten im Zielschema entspricht. Viele der Mappingprobleme stellen sich aber nicht als ein Wahrscheinlichkeitsproblem dar, bzw. ist eine entsprechende Information ist wenig aussagekräftig. Bei MODS gibt es z. B. eine nicht ganz leichte Grenzziehung zwischen <genre /> und <typeOfResource />. Ob sich beides auf OAI-DC Subject mappen lässt hängt von der MODS Implementierung und dem Verständnis von OAI-DC Subject ab und ist entweder gegeben oder nicht. Die Information, dass dieses Mapping vielleicht Gültigkeit hat, bildet nicht die Situation ab und abstrahiert zu stark von

den tatsächlichen Daten. Dieser Punkt verdeutlicht die Notwendigkeit von dynamisch anpassbaren Mappings während der Anwendung auf konkrete Datensammlungen.
Sowohl Schema Registry als auch Crosswalk Registry kamen bei der Registrierung nicht mit Sonderzeichen klar.

5.6. Ergebnis

Das Zusammenführen von unterschiedlichen Datenbeständen in einen neuen Datenbestand kann abhängig von den Daten und der Forschungsfrage wie auch der -methodik interessant und hilfreich sein. Im durchgeführten Crosswalk besaßen die auf MODS gemappten Daten, die mit den vorhandenen im PDR aggregiert wurden eine sekundäre Priorität. Es ging um den Import von Quellendaten für den Beleg von Personendaten, der Primärressource. Hierbei fallen semantische Feinheiten nicht in gleicher Weise ins Gewicht, wie im Falle eines Mappings von Forschungsdaten die im Fokus der Forschung stehen. Trotzdem sind bei der Konfiguration des Mappings einige Probleme aufgetaucht, die für das letzteres Szenario von Bedeutung sind.
Generell lassen sich die Probleme beim Durchführen des Crosswalks in technische und Modellierungsprobleme einteilen. Das gravierendste technische Problem besteht in einer anscheinend mangelnden Unterstützung der vollen XML Schema Expressivität durch die OpenIIHarmony Software. Ebenso sollte das Tool so flexibel sein unterschiedliche Implementierungskonzepte von XML Schema unterstützen zu können. Die automatisierte Verknüpfung eines Elements mit einem anderen Element, dass in unterschiedlichen Kontexten innerhalb seines Schemas auftaucht, erzeugt semantische Probleme, die unnötig sind.
Inhaltlich steht das <genre /> / <typeOfResource /> für die Problematik, dass für ein Mapping der Bezugspunkt nicht das abstrakte Schema sein kann, sondern die Adaption des Schemas in einem definierten Kontext sein muss. Dieser Kontext kann z. B. eine Domain sein. Da dieses Problem jedoch hier nicht aufhört ist die Anpassbarkeit eines Mappings im konkreten Anwendungsfall eine wichtige Voraussetzung für einen sinnvollen Crosswalk. Darüber hinaus ist es für den richtigen Umgang mit Mappings empfehlenswert sowohl Schemadateien als auch und gerade angelegte Mappings mit umfangreichen *Provenance* Metadaten versehen zu können. Hierzu sollten zumindest Projektkontext, Institution und registrierende Person gehören. Getrennt werden sollte auch zwischen registrierender Person und Person, Einrichtung etc., die das Schema erstellt hat.
Eine Empfehlung für die Verbesserung des Workflows könnte die Möglichkeit eines URL-Imports von Schemata sein. Da bereits die XML-Techniken in einem webnativen Kontext eingebunden sind und es best practice ist, ein Schema unter einer URL zur Verfügung zu stellen, wäre dies ein naheliegendes Feature. Im vorliegenden Beispielworkflow mussten die Schemadateien heruntergeladen werden, damit sie dann in der Schema-Registry wieder hochgeladen werden konnten. Es würde die alltägliche Arbeit mit der Schema-Registry ebenfalls erleichtern diese Schema URL in der Auflistung der Schemata zur Verfügung zu stellen.
Kein Werkzeug kann sämtliche Probleme antizipieren, die in Situationen auftreten für die es erstellt wurde. Die Evaluation der Schema-Registry machte deutlich, dass davon auszugehen ist, dass in den meisten Anwendungsfällen händische Arbeit von Nöten ist, bevor ein Crosswalk gewinnbringend durchgeführt werden kann. Ebenso konnte am konkreten Beispiel gezeigt werden, was zuvor bereits abstrakt beschrieben wurde: kein Crosswalk ist ohne Informationsverlust möglich. Die Tragweite beider Punkte fordert vor dem Einsatz des vorliegenden Werkzeuges eine genaue Abschätzung aller Problemfelder im Kontext der jeweiligen Forschungssituation: lohnt sich der Aufwand? Ist der Informationsverlust

essenziell oder peripher? Derlei Überlegungen durchzuführen sollte eine dringende Empfehlung an potenzielle Nutzer sein. Sie setzt außerdem eine Kenntnis der Mechanismen der angewendeten Techniken voraus. In diesem Bereich kann DARIAH-DE auf der Seite von Empfehlungen und Schulungskonzepten einen wichtigen Beitrag leisten

6. Crosswalk II / Mapping Nicht-Standard / Standard

6.1. Strukturentwurf für ein Buchbildarchiv und sein Mapping nach DC Simple

Die Datenbank mit Bildnachweisen aus Druckwerken ist als Informationssystem für Geisteswissenschaftler konzipiert. Das Schema beinhaltet die Möglichkeit, Personen- und Ortsinformationen mit präzisierenden Identifiern zu versehen. Der Aufwand bei der Datenerfassung ist durch den Strukturentwurf in pragmatischen Grenzen gehalten. Die ohnehin zur (Wieder-)Beschaffung notwendigen bibliographischen Informationen dienen zugleich als wesentliche Kontextinformationen der Abbildung. Die ursprünglichen Attribute pnd, tgn und typ sind hier als Elemente abgebildet. Einfachste Beispieldatensätze als Dublin-Core-Metadaten finden sich in einer virtuellen Maschine der DARIAH-Infrastruktur.[6] Das Mapping als Handskizze ist farbcodiert von unproblematisch (grün) nach problematisch (rot). Die Pfeilspitzen deuten an, dass die Relation nicht umkehrbar ist, und hier und da findet sich eine Relation annotiert.

Bei den wenigsten Links zeigt sich dieses Mapping eindeutig. Die Verknüpfungen sind meist nicht umkehrbar, und führen oft zu einer unerwünschten Mehrdeutigkeit. Beispiel: „entstehungsjahrOriginal“, „entstehungsjahrReproduktion“, „jahr(enthaltenInDruckwerk)“ allesamt nach DC:DATE zu linken, ist nicht völlig ohne Sinn, aber doch von jeweils ganz verschiedener Bedeutung. Einen prägnanten Informationsverlust bringt auch das Mapping aller weiteren Felder von „enthaltenInDruckwerk“ zu einem zusammengefassten DC:RELATION-Element, weil auch hier die ursprüngliche Metainformation verloren geht. Zudem stellt sich die Frage nach der Reihenfolge der Feldverarbeitung sowie der notwendigen Interpunktion der zusammengefassten Feldinhalte und der Definition von Regeln hierfür. Im Dublin Core Schema sollte jeder Datensatz ein DC:TYPE = „Image“ enthalten. Im Source Schema existiert jedoch kein korrespondierendes Feld, die Metainformation „Abbildung“ ist hier nur implizit enthalten, so dass ein Link an dieser Stelle nicht möglich ist. Gleiches gilt für den impliziten Medientyp „gedruckt“ ~ DC:FORMAT. Für manche Felder und Attribute, „kommentar“ oder „typ“ etwa, findet sich keine Entsprechung in DC.

Für den eigentlichen Zweck des Buchbildarchivs sind die enthaltenen Datenfelder von unterschiedlicher Relevanz, und gerade diese Relevanzunterschiede gehen beim Mapping

6 DARIAH-jOAI Testserver des Steinheim-Instituts, eingerichtet als Virtuelle Maschine vom Jülich Supercomputing Center. Weblinks (Anmeldung erforderlich):
a) zum Webinterface: https://dariah.zam.kfa-juelich.de/;
b) über einen OAI-PMH-Request: https://dariah.zam.kfa-juelich.de/oai/provider?verb=ListRecords&metadataPrefix=oai_dc&set=bba:kohut-bd1.

nach DC verloren. Das zeigt sich unter anderem an den im Ausgangsschema an verschiedenen Stellen vorgesehenen Identifiern für Personen und Orte, deren eigentlicher Sinn kontextabhängig ist, was bei unterschiedsloser Kodierung als DC:IDENTIFIER verloren geht.

Die intensive Auseinandersetzung mit verschiedenen Crosswalks lässt vermuten, dass das konkret durchgeführte Mapping auch von der mehr oder weniger konkreten Vorstellung beeinflusst ist, wofür es im weiteren Workflow gebraucht wird. Das Mapping würde also von potenziellen Fragestellungen nicht unabhängig und so auch von dieser Seite her nicht eindeutig durchführbar sein. Wenn etwa von einem differenzierteren zu einem weniger differenzierten Schema „gemappt“ wird, und sich damit die Filterungsmöglichkeit der Daten reduziert, kann für einen Suchprozess im weiteren Workflow das Weglassen von Daten aus Relevanzgründen sinnvoll sein (etwa die Jahresangaben im Beispiel „Buchbildarchiv“). Andere Anwendungsszenarien werden diesen Datenverlust andererseits nicht hinnehmen wollen. Das Crosswalk Tool unterstützt den eher technischen Prozess des Mappings, gibt naturgemäß keine Hilfestellung beim Verstehen und Anwenden des Schemas (Beispiel: enthaltenInDruckwerk ~ DC:PUBLISHER oder DC:RELATION ?).

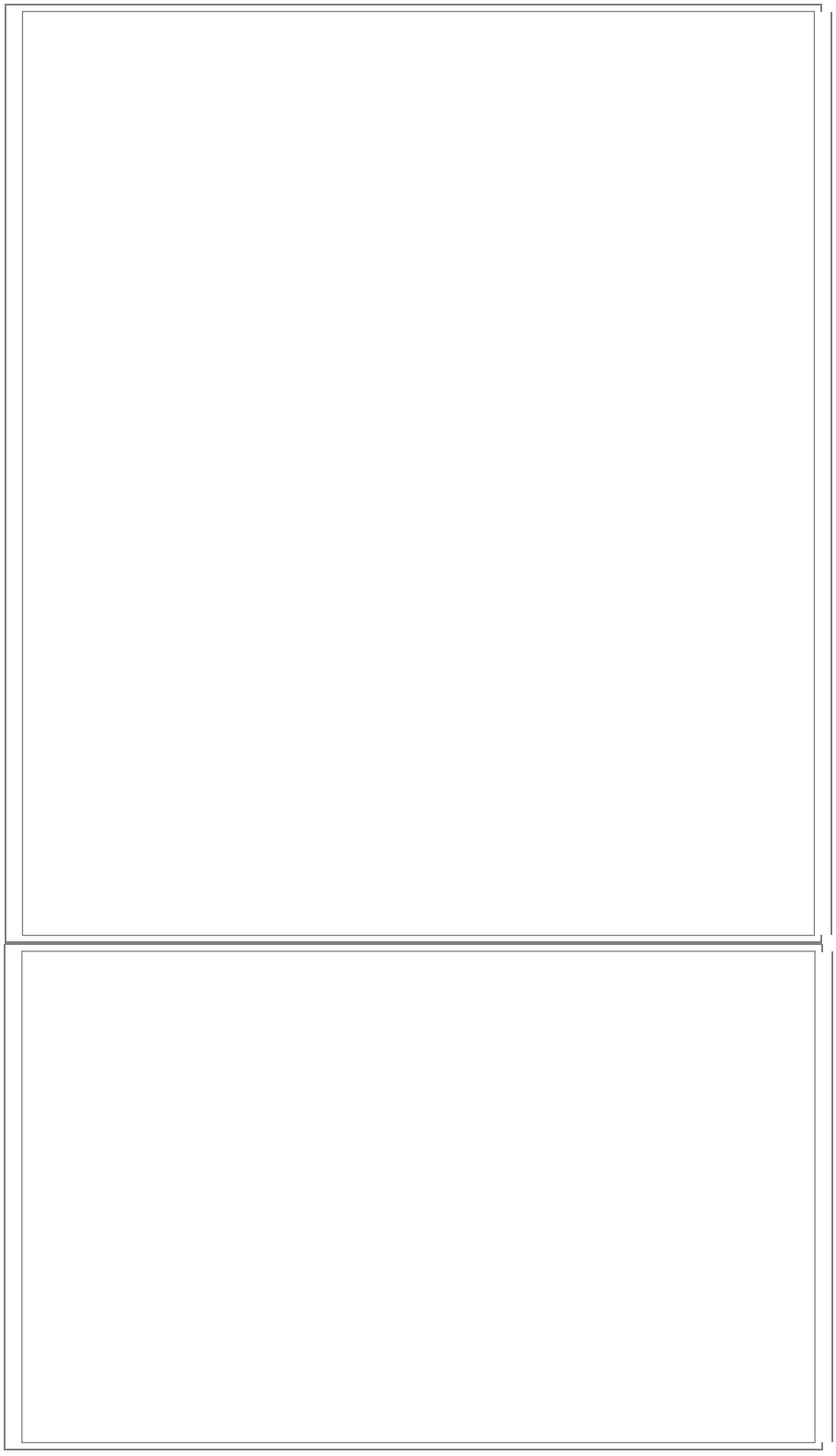

Abbildung 8: Screenshot: Buchbildarchiv | DC Simple (OpenII) sowie Vorentwurf (Handskizze)

6.2. Allegro-C N / Dublin Core Simple

Gerade für die Geisteswissenschaften stellt die Möglichkeit der Digitalisierung verstreuter und oft nur noch vereinzelt vorhandener Quellen einen großen Gewinn dar. Zahlreiche Digitalisierungsprojekte sind von verschiedenen Institutionen durchgeführt worden. Die Beschreibung der so entstandenen Dateien ist mit unterschiedlichen Mitteln und in differenter Ausführlichkeit geschehen. Oft sind hierfür die in Instituten schon vorhandenen bibliographischen Datenbankstrukturen genutzt worden.
Als Beispiel werden hier die Daten aus dem im Steinheim-Institut durchgeführten Projekt Deutsch-jüdische Publizistik des 19. Jahrhunderts benutzt. Im Rahmen dieses Projektes sind Broschüren, Zeitschriftaufsätze, Privatdrucke, aber auch Fragmente aus umfangreicheren Werken und eine komplette Zeitschrift digitalisiert, in Volltext übertragen, ediert und als Neu-Ausgabe online veröffentlicht worden. Die Metadaten sind in einer Allegro-C-Datenbank (N-Format) erfasst, die eine detaillierte und sehr flexible Beschreibung der Objekte ermöglicht. Das führt dazu, dass der Umfang der Metadaten unterschiedlich und objektspezifisch ist. Allegro erlaubt einen XML-Export, der ausgesuchte Elemente aufnimmt. Bei diesem Schritt entstehen allerdings erste Datenverluste, z.B., wenn ein Feld zweifach belegt ist (was ein Software-Problem von Allegro-C ist).
Für einen einfachen Datensatz (selbständiges Objekt ohne hierarchische Struktur) lässt sich, importiert mit EditiX, ein Schema erstellen und als XSD-Datei exportieren. Diese wurde in die Schema Registry hochgeladen und ein Crosswalk Allegro N / DC simple angelegt.

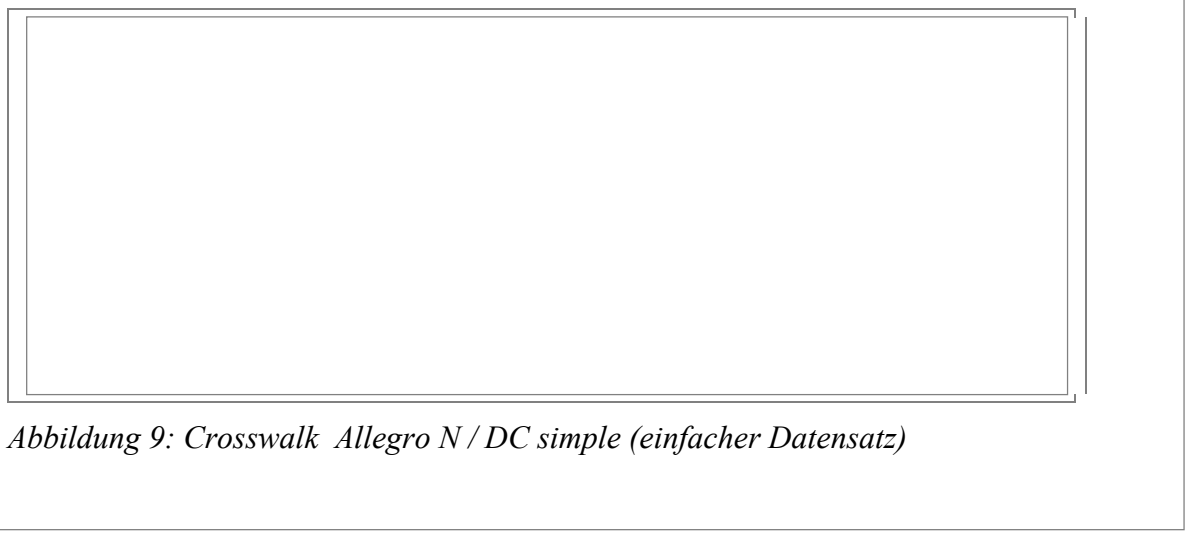

Abbildung 9: Crosswalk Allegro N / DC simple (einfacher Datensatz)

Bei dem eher einfachen Datensatz zeigte sich das Problem, dass die zahlreichen Identifier des Originals mit jeweils eindeutiger Metainformation (URL, URN, ID aus der Datenbank, Signatur) unterschiedslos zu DC:IDENTIFIER werden, so dass sie ihre Interpretierbarkeit verlieren.
Bei Datensätzen mit Bezügen zu über- und untergeordneten Datensätzen (verknüpfte Untersätze in der Ursprungsdatenbank) wird diese Relation nicht wiedergegeben. Es entsteht ein Datenverlust: Nicht darzustellen sind nämlich die Relation zwischen einzelnen Objekten, wie z.B. zwischen einzelnen Ausgaben einer Zeitschrift, die in der ursprünglichen Struktur über eine LinkId (a22060912+020 - 20. Heft der Zeitschrift) festgelegt waren. Auch die Trennung der Angaben zur Erstausgabe und der online-Ausgabe (Herausgeber, Datum, Umfang) ist nicht möglich.

Für hierarchisch strukturierte Daten wurde ein Schema mit EditiX erstellt. In diesen Daten sind in Unterfeldern die Relationen zwischen den neu herausgegebenen Abschnitten und dem ursprünglichen Gesamtwerk dargestellt. Beispiel: Es wurde ein Fragment eines Kapitels aus einem mehrbändigen Werk, dessen Bände nicht gleichzeitig erschienen waren, neu herausgegeben. Diese Informationen sind in der Ursprungdatenbank kodiert. Im Crosswalk erscheint jedoch nur das Attribut uf (= Unterfeld), ohne weitere Differenzierung. So entstehen weitere Datenverluste, die so gravierend sind, dass die Objekte nicht mehr identifiziert werden können (siehe unten, technischer Informationsverlust).

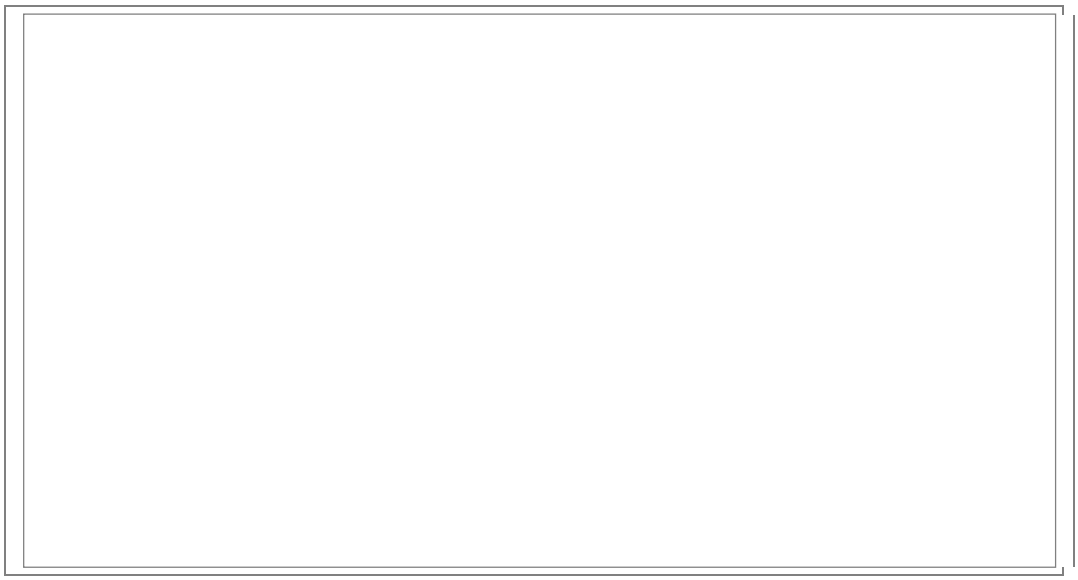

Abbildung 10: Inhaltliche Zuordnung

Bei umfangreichen Datenstrukturen zeigt sich, dass relevante Datenverluste entstehen, etwa bei mehrfacher Zuordnung zu einem Element (Identifier, Source). Ursprünglich vorhandene logische Strukturen gehen verloren.

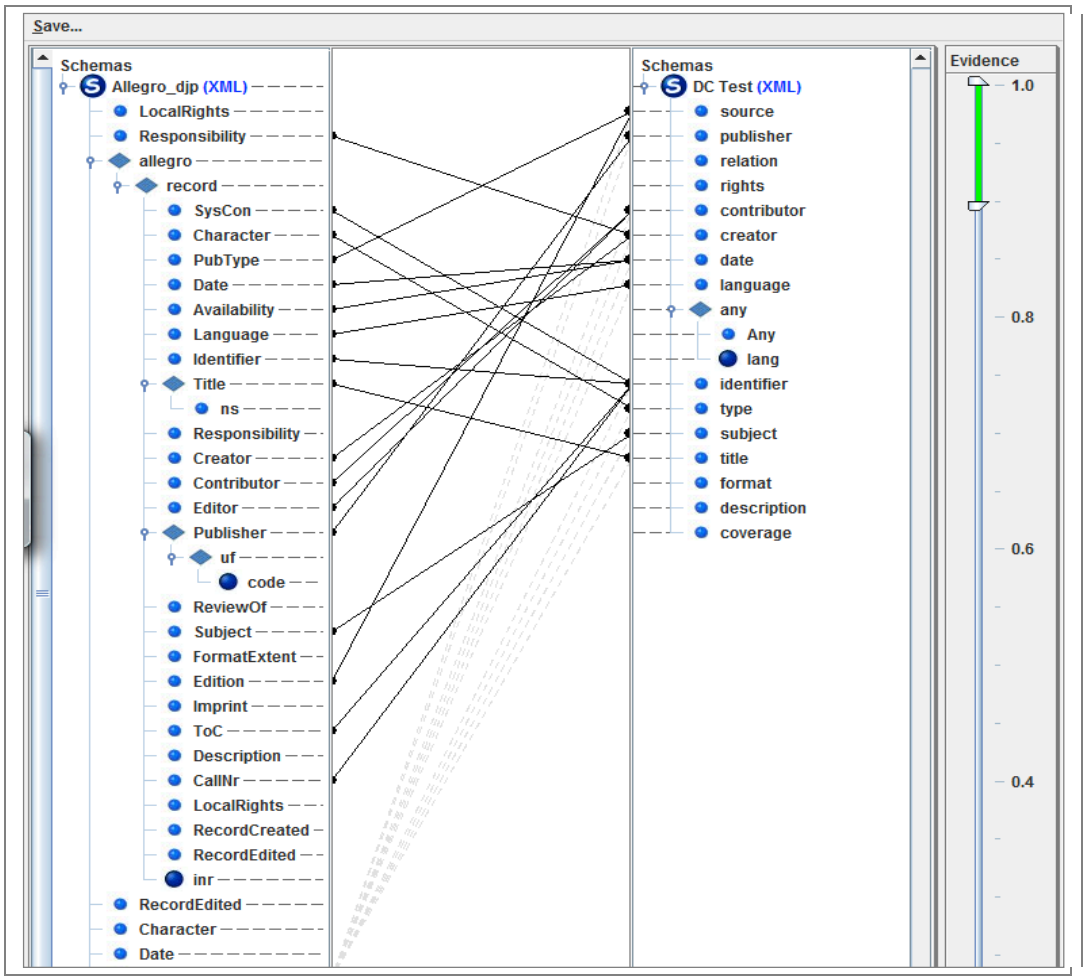

Abbildung 2: Crosswalk Allegro N / DC simple (hierarchischer Datensatz)

Nach der Durchführung des Crosswalks mit Allegro-Daten scheint folgende Empfehlung nahezuliegen: differenziertere Metadaten sollten nicht nach DC Simple übertragen werden, sondern in einen entsprechend umfangreicheren Standard, wenn der weitere Workflow dies erlaubt. Für nichtstandardisierte Datenstrukturen scheint jedenfalls eine Unterstützung durch geeignete Softwarelösungen für das Erstellen eines Schemas unverzichtbar.

6.3. Problematisierung Daten- und Qualitätsverlust

Inhaltliche Divergenzen

Das Mapping von nichtstandardisierten Daten nach Dublin Core Simple bringt erkennbar Datenverluste mit sich, das zeigen die exemplarisch durchgeführten Crosswalks. Zusammenhänge zwischen Ursprungsdatenfeldern lassen sich nicht transportieren, einzelne Feldinhalte lassen sich schlechter oder gar nicht mehr interpretieren und es entsteht ein erheblicher Verlust an Filterungsmöglichkeiten. Das kann beispielsweise wiederum zu unerwünscht vielen Treffern bei einer nachgeschalteten Suche führen. Inwieweit dies problematisch ist, hängt auch von der Art der Weiterverwendung der Daten ab. Zum Umgang

mit dieser Frage könnte auch gehören, für gleiche Schema-Kombinationen unterschiedliche Mappings für unterschiedliche Zwecke anzubieten.

6.3.1. Technischer Informationsverlust

Ein Standard-XML-Export aus Allegro-C kann zu XML-Dateien führen, die ihre inhaltlichen Strukturinformationen in Attributen speichern. Ein daraus exportiertes Schema enthält diese Informationen nicht mehr (siehe Abbildung 3, Beispiel I) und ist daher für einen Crosswalk und insbesondere den weiteren Workflow in Richtung generische Suche wertlos. Gleiches gilt wohl für MARC21. Hier bedarf es Konvertierungsroutinen oder aber manuell nachgebildeter Schemata oder eben alternativ, dass im gesamten Workflow, also vor Laden in die Schema Registry, Strukturinformationen in Attributen unterstützt werden. Die abgebildeten einfachen Beispiele betreffen bibliograpgische Daten, andere Daten lassen sich aber ebenso (unglücklich für den Zweck eines Crosswalks) formulieren. Im Sinne von *best practice* wäre Beispiel II dem Beispiel I vorzuziehen, wobei die Unterfelder <uf/> hier noch nicht korrigiert sind.

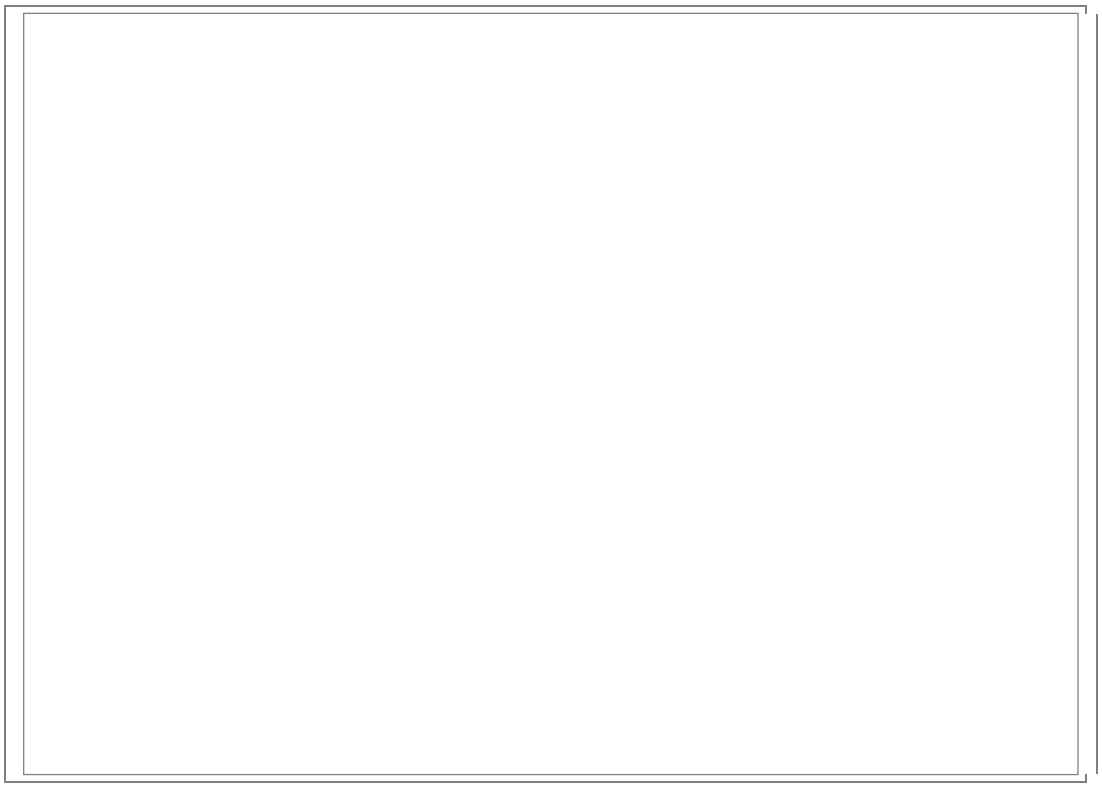

Abbildung 3: Strukturinformationen in Attributen (Beispiel I)

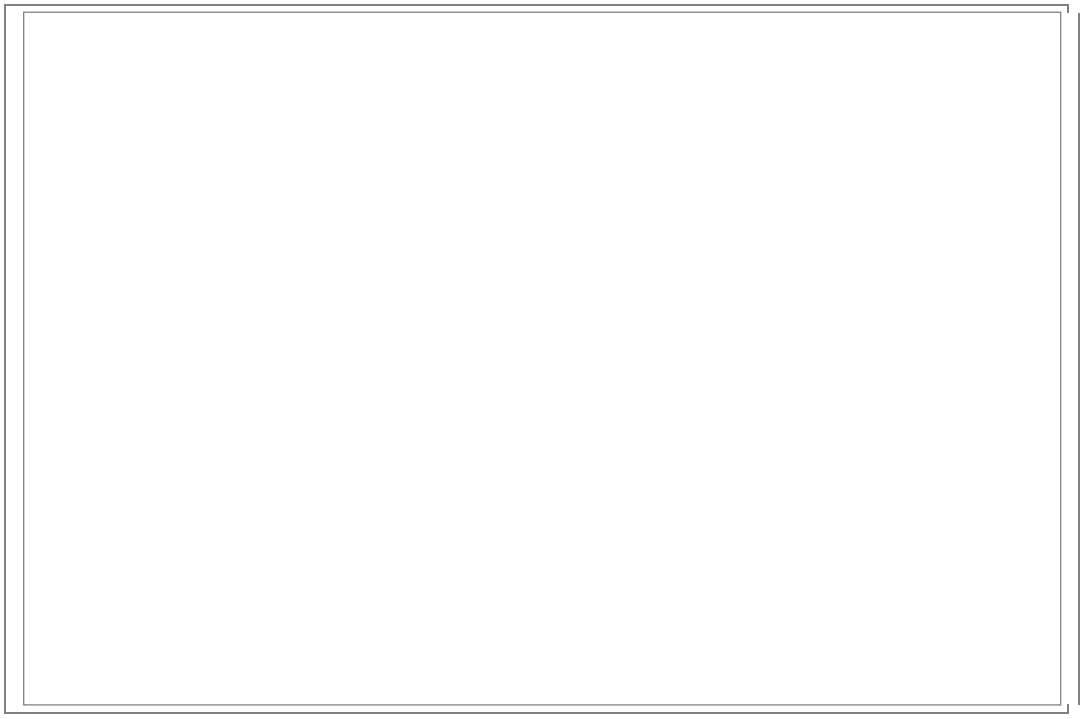

Abbildung 4: Strukturinformationen als Elementnamen (Beispiel II)

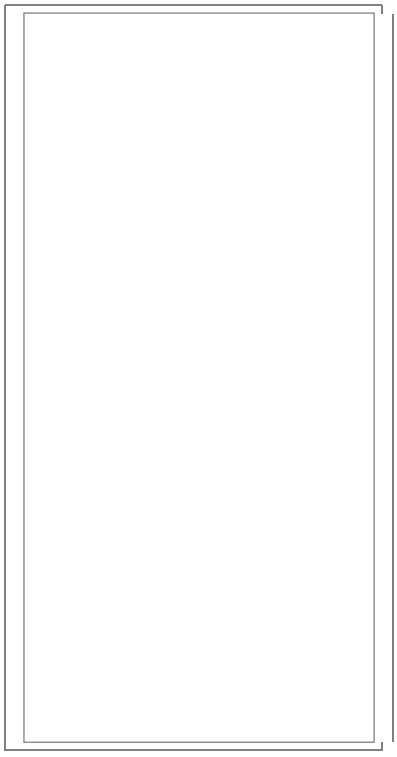

Abbildung 5: Marc21 in der Schema Registry

7. Fazit

Das im Milestone M 1.2.1 vorgestellte Konzept zur Schema Registry ist eine sinnvolle Entwicklung, das in vielerlei Hinsicht aktuelle Bedürfnisse im Umgang mit Forschungsdaten aus geisteswissenschaftlichen Forschungsprojekten aufgreift. Jedoch sind noch praktische Probleme bei der technischen Umsetzung erkennbar und machen daher an manchen Stellen eine Überarbeitung notwendig. Dies liegt sicherlich auch daran, dass es sich bei dem evaluierten Tool um einen ersten Prototypen handelt, der sich noch in der Entwicklung und nicht in einem releasefähigen Zustand befindet.

Allerdings sind aus Sicht der an der Evaluation beteiligten Fach- und Geisteswissenschaftler noch einige Eigenschaften bzw. Features notwendig; sie sollten bei der weiteren Programmierung berücksichtigt bzw. thematisiert werden. Dies betrifft nicht nur die Funktionalität und das Konzept des User-Interfaces, sondern insbesondere auch die Anforderungen, die dem Wissenschaftler gestellt werden, um sinnvoll mit der Software arbeiten zu können (z.B. der Umgang mit Schema-Beschreibungen in Form von xsd-Daten). Generell ist festzustellen, dass die Umsetzung als Java Applet aus Gründen der Usability leider zu Problemen geführt hat. Die Evaluation hat gezeigt, dass das Tool auf einer Vielzahl der derzeit gängigen Browser und Betriebssysteme nicht funktionierte. De facto war eine problemlose Verwendung nur unter *Windows* mit *Chrome* möglich. Andere Browser wurden teilweise oder gar nicht unterstützt, was sicherlich am Entwicklungsstand des Prototpyen liegt. Nichtsdestotrotz muss bei der weiteren Entwicklung sicher gestellt werden, dass das Applet zumindest unter den am häufigsten verwendeten Browsern Firefox und Internet Explorer - jeweils in den aktuellen, aber auch in älteren Versionen - lauffähig ist. Dies ist insofern von besonderer Bedeutung, als geisteswissenschaftliche Nutzer, die oftmals in universitären Netzen und Domains arbeiten, in der Regel an die jeweiligen IT-Policies der Universitäten und Forschungseinrichtungen gebunden sind und daher nicht flexibel bestimmte Systemvoraussetzungen erfüllen können. Über die notwendigen Administratoren-Rechte, um selbständig Installationen veranlassen zu können, werden nur die wenigsten verfügen. Deshalb müssen die Anforderungen an zusätzlich zu installierende Java-Packages so gering wie möglich gehalten werden, da ansonsten davon auszugehen ist, dass aufgrund der technischen Hürden kaum ein Geisteswissenschaftler die Software einsetzen kann.
Die der Evaluation zugrunde liegende Version zeigt zwar nur den aktuellen Stand der Entwicklung, doch bedarf es gerade hinsichtlich der Kompatibilität noch weiterer Nachbesserungen. In vielen Forschungsumgebungen sind die Computer fremdadministriert. Da dem Forscher nicht immer ein spezifischer Browser unter einem spezifischen Betriebssystem zur Verfügung stehen wird, wäre es sinnvoll, generische Technologien wie HTML5 oder vergleichbares für das Frontend in Erwägung zu ziehen. Die an der Evaluation beteiligten Fachwissenschaftler werden deshalb weitere Rücksprache mit dem Entwicklerteam halten.
Bei der Evaluation wurde zudem deutlich, dass eine langfristige und auf Dauer angelegte Pflege und Wartung notwendig sein wird. Wie aus anderen IT-Projekten bekannt ist, sind Bedienkonzepte und verfügbare Technologien einem ständigem Wandel unterworfen, die es sind laufende Anpassungen erfordern. So könnten etwa neue Browserversionen benötigte Features und Funktionalitäten nicht mehr unterstützen, Updates der Programmierumgebung könnten Anpassungen notwendig machen, geisteswissenschaftliche Forscher und Nutzer könnten Bedarf an neuen Funktionalitäten äußern. So ist mit einem Pflege-, Wartungs- und Anpassungsaufwand zu rechnen, der eine permanente Betreuung notwendig macht. Hierzu muss auf Ebene des deutschen DARIAH-DE-Konsortiums eine Lösung gefunden werden, da nur Angebote, die eine dauerhafte Verfügbarkeit gewährleisten, auch von der angesprochenen

Zielgruppe genutzt werden dürften. Aus diesem Grund ist ein langfristiges Entwicklungs- und Betreuungskonzept notwendig.

Wie bereits beschrieben, hat gerade der Umgang mit XSD-Dateien zu Irritationen geführt. Dies liegt vor allem daran, dass eine Editor-Komponente bzw. ein entsprechendes Feature fehlt, das bei der Erstellung und Validierung von XML Schema Definitionen unterstützt. Hier wäre mindestens eine Empfehlung oder Integration eines geeigneten Editors wünschenswert.[7] Es sollte hierbei beachtet werden, dass die Verwendung von Tools wie Eclipse oder vergleichbaren nicht zielgruppengerecht ist. Die empfohlenen bzw. implementierten Tools sollten geisteswissenschaftlichen Forschern und Nutzern einen möglichst einfachen technischen Einstieg ermöglichen, welcher nur Grundlagenwissen in XML-Techniken und Datenbanken voraussetzt.

Neben den bereits im Rahmen der Evaluation genannten Fehlerbeschreibungen und gewünschten Verbesserungen ist eine vollständige Dokumentation der Funktionalität notwendig, die auch einen direkten Einstieg ermöglicht. In dieser Dokumentation muss der Umgang mit den Komponenten, deren Zusammenspiel und die Möglichkeiten und Grenzen der Applikation beschrieben werden. Gleichermaßen müssen geisteswissenschaftliche Nutzer in der Verwendung des Tools angeleitet werden. Auch über Schulungs- bzw. Präsentationsmaterialien ist nachzudenken.

In dieser Art weiterentwickelt, so die Einschätzung der Arbeitsgruppe, wird das Tool einen Beitrag dazu leisten können, dass Geisteswissenschaftler sich mit unterschiedlichen Metadatenformaten auseinander setzen und dadurch wertvolle Unterstützung beim Design ihrer je individuellen Forschungsanwendung erhalten. Die isolierte Betrachtung von Crosswalks zu Standards hin, wie sie hier insbesondere für Dublin Core durchgeführt wurde, führt naturgemäß vor allem die dabei auftretende Reduktion von Komplexität vor Augen, die vom einzelnen Wissenschaftler als eher problematisch empfunden wird. Die großen Vorteile für die Herstellung von Interoperabilität, die man erhoffen und erwarten kann, und für die hier Grundlagenarbeit geleistet wurde, können im Rahmen des vorliegenden Meilensteins noch nicht sichtbar werden, zeichnen sich jedoch schon vielversprechend für den weiteren Projektverlauf ab.

7 Siehe Abschnitt 4.1.3.

8. Anhang: Tabellarische Auflistung der identifizierten Anforderungen an die Schema Registry

Art	Beschreibung	Vorschlag aus AP3.4	Bewertung
Plattformunabhängigkeit	Voller Funktionsumfang unter allen in der Forschung gängigen Betriebssystemen und Browsern	Ggf. zunächst Standardkonfigurationen komplett abdecken (MS mit IE/FF) und dann erweitern.	Wichtig, um von größeren Kreisen geisteswissenschaftlicher Forscher eingesetzt zu werden
Volle Unterstützung von XML Schemata	Bei den Beispielen zu Mods wurden Probleme mit der Verarbeitung von komplexen Schemata mit mehreren Wurzelelementen deutlich	Es werden genauere Fallbeispiele aus AP3 gesammelt und Probleme gemeinsam analysiert.	Wichtig, um bei komplexeren Daten eingesetzt werden zu können
Mehr Eingriff in das automatische Mapping	Die automatisch generierten Mapping-Verbindungen sollten besser editierbar sein	Eine Editierfunktion für diese Verbindungen	Moderat. Wenn man das automatische Mapping verwenden soll jedoch unabdingbar.
XSD Editor	Es fehlt an einer konkreten Editierkomponente	Empfehlungen für geeignete Editoren aussprechen und diese mit in eine Dokumentation aufnehmen	Moderat, aber gerade für Collections, die mit Nicht-Standardisierten Metadaten-Schemata erfasst bzw. angelegt wurden, unabdingbar.
Dokumentation	Es fehlt noch an einer übergreifenden Dokumentation mit passenden Beispielen	Verfassen einer Dokumentation, welche möglichst wenig technisches Vorwissen voraussetzt	Wichtig, um Fachwissenschaftlern den Zugang zu ermöglichen
Visualisierung der Qualität von Relationen	Relationen sollten gekennzeichnet werden können um etwa den Grad der Passung visuell darstellen zu können	Farbliche Kodierung der Linien könnte dieses ermöglichen	Moderat
Versionsverwaltung	Es sollte die Möglichkeit gegeben werden mehrere Versionen eines Mappings anzulegen		Moderat
Visuelle Rückmeldungen	An diversen Stellen ist eine visuelle Rückmeldung notwendig. Auch gerade bei der Identifikation von Fehlern, etwa beim Import von XSDs	Visuelle Rückmeldungen beim Speichern, Import, Export und wichtigen Arbeitsschritten	Wichtig, um ein Arbeiten zu ermöglichen
Vererbung der ursprünglichen Meta-Information im Workflow	Beispiel: „DC:DATE war ursprünglich *bba:entstehungsJahrReproduktion*“. Das wäre für eine inhaltsreiche Suchfunktionalität wünschenswert	Beispiele zur genaueren Analyse können von AP3 erbracht werden	Wichtig
Export	Downloadmöglichkeit hinterlegter Schemata und Mappings		Wichtig
Zeichenkodierungsprobleme im Feld „Description“ der Schema Registry		Beispiele zur genaueren Analyse könnten von AP3 erbracht werden	Moderat
Bearbeiten der „Description“ eines Datenfeldes ist möglich, geht aber verloren, sobald das Feld nicht mehr im Fokus ist		Beispiele zur genaueren Analyse könnten von AP3 erbracht werden	Wichtig
Beschreibung von Schemata	Mehr Felder zur Beschreibung von Schemata und Crosswalks (Provenance-Informationen wie Institution, Projekt, Schema URL, ...)	AP3 unterschützt gerne bei der Identifikation der notwendigen Felder	Moderat