AP5: Ressourcen und Dienste

Task Force für Metadaten im VLO

—

Tischvorlage für die Weiterarbeit

Abgefasst von: *Susanne Haaf* (BBAW)
unter Mitarbeit von: *Axel Herold* (BBAW), *Peter Fankhauser* (IDS)

Beraten und verabschiedet von der VLO-Taskforce

Mitwirkende:

- *Thorsten Trippel*: **Koordination** (Seminar für Sprachwissenschaft, Universität Tübingen)
- *Thomas Eckart* (Abt. Automatische Sprachverarbeitung, Universität Leipzig)
- *Axel Herold, Susanne Haaf* (Berlin-Brandenburgische Akademie der Wissenschaften Berlin)
- *Kerstin Eckart, Jens Stegmann* (Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart)
- *Peter Fankhauser* (Institut für Deutsche Sprache Mannheim)
- *Florian Schiel* (Bayerisches Archiv für Sprachsignale, LMU München)
- *Dieter van Uytvanck* (Max Planck Institut für Psycholinguistik Nijmegen)
- *Hanna Hedeland* (Hamburger Zentrum für Sprachkorpora, Universität Hamburg)
- *Jörg Knappen* (Universität des Saarlandes, Englische Sprach- und Übersetzungswissenschaft, Saarbrücken)

Version: 07. Januar 2014

Inhalt

Einleitung

Die vorliegende Tischvorlage stellt das Ergebnis der VLO-Taskforce dar. Die VLO-Taskforce besteht aus jeweils einem bis zwei VertreterInnen aller Zentren und findet sich in regelmäßigen Videomeetings zusammen, um den Zustand des VLO und der darin dokumentierten Ressourcen zu bewerten sowie Überarbeitungen und Vereinheitlichungen zu erreichen.

Im Folgenden werden die von der Taskforce behandelten Themen erläutert sowie der gegenwärtige Stand der Arbeiten und die weitere Planung daran beschrieben.

Allgemeines zum weiteren Vorgehen

Da das VLO voraussichtlich ab Februar 2014 zusätzliche Entwicklerkapazitäten bekommen wird, werden Entscheidungen, die für die Weiterentwicklung einzelner Aspekte des VLO vonnöten sind, bis Ende Januar 2014 durch die Taskforce zu treffen sein. Fragen, die den Zuständigkeitsbereich der SCCTC-Taskforces auf CLARIN-ERIC-Ebene betreffen (z.B. die Weiterentwicklung von CMDI und ISOcat) werden in Absprache mit den jeweiligen Taskforces behandelt.

Nachdem sich die Taskforce bislang in toto per Videomeeting und E-Mail-Verteiler austauschte, sollen nun vorübergehend für einzelne Themen Teilgremien gebildet werden, die zunächst in kleinem Kreis Vorschläge für das weitere Vorgehen erarbeiten. Diese Vorschläge werden dann in der Taskforce besprochen und umgesetzt. Diese Teilgremien sind:

- **Gruppe Dokumentation:** Thorsten Trippel, Peter Fankhauser, Susanne Haaf
- **Gruppe Relationen:** Florian Schiel, Peter Fankhauser, Jens Stegmann
- **Gruppe Vokabulare:** Thomas Eckart, Hanna Hedeland, Jörg Knappen
- **Gruppe ISOcat:** Axel Herold, Dieter van Uytvanck, Kerstin Eckart

Die Ergebnisse der Gruppen werden Ende Januar 2014 in der Taskforce beraten und werden sodann der Weiterentwicklung des VLO zugrundegelegt.

Auffindbarkeit von Ressourcen im VLO

Problem Statement

Ein Problem stellte die Befüllung der Facetten aufgrund der dafür ermittelten Metadaten aus den CMDI-Metadatensätzen der Zentren dar. Grundsätzlich sollte die Befüllung aufgrund der eindeutigen Zuordnung von ISOcat Data Categories (DCs) zu CMDI-Komponenten innerhalb der jeweiligen CMDI-Profile erfolgen. Verschiedene ISOcat DCs sind jedoch nicht klar voneinander abgrenzbar, und der Skopus von Komponenten innerhalb der verschiedenen CMDI-Profile kann ebenfalls variieren. Daher kam es zu Fehlzuordnungen von Metadaten zu einzelnen Facetten.

Stand

1. Die Facetten und ihre vorgesehene DC-Befüllung wurden durch die Entwickler dokumentiert und den Zentren diese Dokumentation zur Verfügung gestellt. Aufgrund der Dokumentation wurden die Metadaten-Profile der einzelnen Zentren angepasst bzw. die Metadatensätze kuriert, um die Homogenität der Einträge in den Facetten zu gewährleisten.

2. Wo eine eindeutige Zuordnung von im VLO zugrunde gelegten ISOcat DCs zu bestimmten Metadatensätzen bzw. CMDI-Komponenten nicht möglich war, erarbeiteten die Zentren jeweils Listen von XPaths, über die Metadateninhalte für bestimmte Facetten eindeutig angesteuert werden können. Diese XPath-Sammlungen umfassen sowohl Blacklists, d.h. Listen von XPaths, die zu fehlerhaften Treffern führen (Metadatenelemente, die *nicht* in der entsprechenden Facette erscheinen sollen), sowie Whitelists, d.h. Listen von XPaths, die die korrekten Metadatenelemente adressieren, soweit sie nicht bereits über die DC-Zuordnung gefunden werden. Die Verwendung dieser bereitgestellten XPaths für die Befüllung der Facetten wurde für das VLO implementiert.

Perspektive

Zukünftig soll die Auswertung der Metadaten so erfolgen, dass XPaths nicht mehr nötig sind, sondern die Zuordnung außschließlich über die ISOcat DCs erfolgen kann. Hierfür ist es nötig, die ISOcat DCs hinsichtlich ihres Skopus und ihrer Verwendung in den einzelnen CMDI-Profilen zu prüfen und ggf. semantisch zu konkretisieren oder die Zuordnungen in den jeweiligen Profilen anzupassen. Für die Ressourcen der Zentren – auch der Zentren der europäischen Partnerprojekte – soll diese kuratorische Aufgabe von den National Metadata Quality bzw. ISOcat Koordinatoren organisiert werden. Für CLARIN-D obliegt die Umsetzung der VLO-Taskforce.

Die Erarbeitung eines Vorschlags für die Spezifikation von ISOcat-Kategorien aufgrund der CMDI-Profile der CLARIN-D-Zentren wird von der *Gruppe ISOcat* übernommen. Eine Vorlage für die weitere Beratung in der VLO-Taskforce wird bis zum 20. Januar 2014 erarbeitet. Die Implementierung der Ergebnisse erfolgt bis Ende Februar 2014.

Metadatenkuration

Problem Statement

Jeder Metadaten-Provider kennt die eigenen Ressourcen und kann deren Korrektheit, Vollständigkeit und Zugänglichkeit über das VLO beurteilen. Dennoch ist für die Frage der Sichtbarkeit und des Zugangs zu Ressourcen häufig auch die Nutzerperspektive aufschlussreich.

Stand

Um die Erreichbarkeit und Nutzbarkeit der bereitgestellten Metadaten zu prüfen, wurden innerhalb der Zentren Teams gebildet, die gegenseitig die Metadatenkontrolle übernahmen:

- BBAW – UdS
- BAS – HZSK
- Uni Tübingen – Uni Stuttgart

Die gegenseitige Metadatenkontrolle und daraufhin erforderliche eventuelle Korrekturen wurden erfolgreich durchgeführt und abgeschlossen.

Perspektive

Es sollen Verfahren zur automatischen Prüfung der Qualität von Metadaten entwickelt werden. Diese sollen etwa auf den für die Form von Metadaten festgelegten Best Practices, auf Regeln für die Einheitlichkeit erfasster Daten sowie auf dem Verhältnis der angegebenen Daten zu dem jeweils verwendeten Metadatenprofil beruhen. Ziel ist es, für einzelne Metadateninstanzen oder Mengen von Instanzen eines CMDI-Profils aussagekräftige Analysen zu generieren, die auf mögliche Probleme der Metadatenmodellierung aufmerksam machen. Von einem solchen Verfahren profitieren auch die europäischen Partnerprojekte direkt.

An der Implementierung wird sich die VLO-Taskforce aktiv beteiligen. Die Planungen hierfür werden im Anschluss an die Fertigstellung der Dokumentation zu den Best Practices für die Metadatenerfassung in CLARIN erfolgen (i.e. ab April 2014).

Auswahl der Facetten der VLO-Startseite

Problem Statement

Das VLO soll mittels einen Facettenbrowser einen facettierten Zugriff auf die einzelnen verzeichneten Ressourcen ermöglichen. Die Facetten stellen dabei bestimmte inhaltliche Kategorien dar, nach welchen die Metadaten jeweils gefiltert werden sollten. In der Taskforce wurde die bislang festgelegte Auswahl der Facetten zur Diskussion gestellt: Die Frage war, inwieweit die bisher festgelegten Facetten a) die im VLO bereitgestellten Ressourcen ausreichend repräsentieren und b) möglichen Szenarien für den Ersteinstieg durch die Nutzer gerecht werden.

Stand

1. Die Zentren prüften die Facettenauswahl für die eigenen Metadaten und stellten jeweils Vorschläge für sinnvoll erscheinende Facetten bereit. Gemeinsam legten die Zentren aufgrund dieser Vorschläge eine Auswahl von **Facetten** fest, **die perspektivisch für den Sucheinstieg auf der VLO-Startseite verfügbar sein sollen**:

- Resource Class (text, lexical resource, video data, audio data, …; Grundlage bildet das für DC 3806 vorgeschlagene Vokabular, das aber noch erweitert werden muss)
- Modality (speech, writing, facial-expressions, …)
- Format (TXT, JPG, TEI-XML, … – das technische Repräsentationsformat der Ressource)
- Language(s) of the resource (Sprache(n), die als Primärsprachen der Ressource fungieren; nicht die Sprachen von Sprechern/Autoren usw.)

- Organisation (Institution, die die Ressource aktuell betreut und bereitstellt; Institution, an die man sich wenden sollte, um Fragen in Bezug auf die Ressource zu klären oder um Zugriff zu erhalten)
- Country (Land, in dem die Ressource entstanden ist; *nicht* das Land, in dem die Ressource aktuell vorgehalten wird)
- Project (Name des Projekts, in dem die Ressource erstellt wurde)
- Collection (Sammlung von Ressourcen, der eine Ressource angehört)
- Time Coverage (Zeitraum, den die Primärdaten repräsentieren; Aufnahmedatum; *nicht* Zeitraum der Ressourcenerstellung, die ja später erfolgt sein kann)

2. **Gegenüber der ursprünglichen Auswahl neu** sind dabei die folgenden Facetten:
- Resource Class
- Resource Type
- Modality (speech, writing, facial-expressions, …)
- Project
- Time Coverage

Zunächst konnten die Facetten *Resource Class* und *Format* auf der VLO-Startseite realisiert werden. Die Zentren konnten ihre Metadaten jeweils in der Form kurieren, dass die Menge der ursprünglich in *Resourcetype* vereinigten Werte aufgeteilt werden konnte auf die Facetten *Format* und *Resource Class*.

3. Der direkte Einstieg zu den Ressourcen über *Name* und *Description* von der Startseite aus bleibt erhalten. Die Menge der Ressourcen, die je Seite sichtbar sind, wurde von 10 auf 30 erhöht.

Perspektive

Aus der unter 2. genannten Auswahl konnten folgende Facetten **bislang nicht realisiert** werden:
- Modality (speech, writing, facial-expressions, …)
- Project
- Time Coverage

Die Umsetzung dieser Facetten inkl. der entsprechenden Kuration der Metadaten durch die Zentren zur Befüllung der Facetten wird bis Ende Januar 2014 erfolgen.

Die folgenden Facetten der ursprünglichen Auswahl sollen zunächst anhand der im VLO vorliegenden Metadaten **auf ihren Nutzen für den Ersteinstieg der Recherhe hin übergeprüft** werden:
- Continent
- Genre
- Subject
- Dataprovider
- Nationalproject

- Tag (aka Keywords)

Diejenigen Facetten, die sich weiterhin als notwendig erweisen, bleiben bestehen; die übrigen Facetten werden von der VLO-Startseite entfernt.

Perspektivisch ist außerdem zu prüfen, ob eine nach Nutzer individualisierte und/oder nach Recherchetiefe variierende Facettenauswahl sinnvoll und umsetzbar wären.

Der Prozess der Prüfung, ggf. Metadatenkuration und Festlegung der endgültigen Facettenauswahl wird bis Februar 2014 abgeschlossen sein. Die Implementierung der endgültigen Facettenauswahl erfolgt bis Ende März 2014.

Kategorienauswahl auf der Zielseite der Ressourcen

Problem Statement

Die Kategorienauswahl auf der Zielseite der jeweiligen Ressourcen beläuft sich derzeit auf folgende Einträge:

- collection
- continent
- country
- dataProvider
- description
- genre
- id
- languages
- metadataSource
- name
- nationalProject
- organisation
- projectName
- subject
- year

Diese Kategorienauswahl erwies sich für eine Reihe von Ressourcentypen als nicht passend bzw. nicht ausreichend.

Stand

Die Überarbeitung der Kategorien der Zielseiten wurde aufgenommen. Dabei erarbeiteten die Zentren unter Berücksichtigung der eigenen Ressourcen Vorschläge, welche die folgenden Kategorien umfassten:

- LiveCycleStatus (DC 3818)
- Title (e.g. Abhandlung über den Ursprung der Sprache)
- Author (e.g. Johann Gottfried von Herder)
- Creation date of source (z.B. 1751)
- Publication date of resource (z.B. 2013-09-05T13:40:57Z)

- Contact
- Rights Holder (dc-rightsHolder o.Ä.)
- Licence (dc-rights, recommended Licence-Komponente etc.)
- NationalProject
- Creator
- Project
- Version

Perspektive

Die Kategorienauswahl der Zielseiten wird von der Taskforce neu überlegt. Hierzu werden die Zentren Vorschläge erarbeiten, auf Grundlage derer eine Kategorienauswahl festgelegt wird. Dieser Prozess wird bis Ende Dezember 2014 abgeschlossen sein. Die Umsetzung erfolgt bis Ende März 2014.

Einheitliche Vokabulare für die Befüllung von Facetten

Problem Statement

Bislang gab es kaum formale Festlegungen von Wertebereichen für die Befüllung einzelner CMDI-Komponenten mit Metadaten. Dies führt dazu, dass die in den Facetten zugänglichen Werte zum Teil sehr heterogen sind, sich z.B. hinsichtlich der verwendeten Sprache, der Groß- oder Kleinschreibung, Getrennt- oder Zusammenschreibung etc. unterscheiden.

Für einzelne Facetten erscheint die Festlegung kontrollierter Vokabulare, die ggf. erweiterbar wären, möglich (z.B. für die Facette Ressource Class, Language etc.), um gleichartige Inhalte mit gleichartigen Ausdrücken und somit geschlossen zugänglich zu machen (z.B. keine Trennung von Sprachressourcen in *German, Deutsch, ger* etc., sondern ein einheitlicher Wert *German*). Die Problematik wirkt sich nicht allein auf die Darstellung der Werte innerhalb der Facetten aus, sondern auch auf die Suche in sämtlichen Metadatensätzen über das Suchfeld des VLO.

Stand

In der Taskforce wurden verschiedene Lösungen diskutiert:

1. Festschreibung von Best Practices für die Metadatenaufnahme auf Ebene der Elementinhalte für Metadaten-Provider;
2. automatische Vereinheitlichung der geharvesteten divergierenden Metadaten und Präsentation mit vereinheitlichten Schreibweisen (z.B. die Werte *German, Deutsch, ger* werden zunächst abgebildet auf einen einheitlichen String *german*, der als normalisierter Wert *German* im VLO dargestellt wird.

Beide Lösungen werden angestrebt. Einerseits sollen den Metadatenprovidern kontrollierte Vokabulare und Best Practices an die Hand gegeben werden, an denen sie sich orientieren können. Andererseits sollen die Provider weiterhin die Freiheit behalten, ihre Daten den eigenen Vorstellungen entsprechend aufzunehmen. Die daraus resultierende Inhomogenität soll VLO-seits durch die automatische Nachbearbeitung der eingehenden Daten entsprechend des

unter Punkt 2 skizzierten Weges aufgefangen werden. Für die Suche werden die eingehenden Werte grundsätzlich zu Klein- und Zusammenschreibung (bei Mehrwortausdrücken) normalisiert.

Perspektive

Für eine neue CMDI-Version (2.0; geplant 2014) ist es vorgesehen, die Einbindung externer fester Vokabulare in CMDI-Profile zu ermöglichen. Dadurch können Profile und Wertebereiche von Elementen getrennt gepflegt werden.

Für die Formulierung von Best Practices zur Metadatenerfassung wird die *Gruppe Vokabulare* der VLO Taskforce bis zum 20. Januar 2014 einen Vorschlag vorlegen. Nach Beschlussfassung durch die VLO-Taskforce obliegt den Zentren dann die Kuration der eigenen Daten gemäß den formulierten Best Practices, die bis Ende März 2014 erfolgen soll. Die Dokumentation der Best Practices wird durch die *Gruppe Dokumentation* koordiniert.

Trennung Sprachressourcen und Tools/Services

Problem Statement

Die ursprüngliche Konzeption des VLO sah die Vereinigung von Metadaten für Sprachressourcen, Tools und Services (TS) vor, die durch CLARIN zur Verfügung gestellt werden, um diese Ressourcen trotz ihrer genuinen Heterogenität über eine gemeinsame Umgebung einheitlich zugänglich zu machen. Problematisch dabei ist, dass die Schnittbereiche der Metadatenbeschreibungen von den Sprachressourcen einerseits und TS andererseits so gering sind, dass sich dies auf die Ermittlung geeigneter Facetten als Ersteinstieg auswirkt, während die Notwendigkeit einer Suche über der Vereinigungsmenge von Sprachressourcen und TS nicht unmittelbar einleuchtet. Derzeit gehen darüber hinaus die TS in der Fülle der verzeichneten Sprachressourcen unter. Ein direkter Einstieg nur zu den TS kann zwar über die Filterfunktion nach Ressource Type erreicht werden; dieser Sucheinstieg setzt jedoch ein gewisses Maß an Vorwissen beim Nutzer voraus.

Stand

In der Taskforce wurden verschiedene Lösungsszenarien für das geschilderte Problem diskutiert:

1. Auslagerung von TS aus dem VLO, z.B. in die WebLicht-Umgebung oder in CLARIN-fremde Verzeichnisse;
2. mehrere Seiten innerhalb des VLO mit unterschiedlichen Facetten für Sprachressourcen einerseits und TS andererseits;
3. weitere gemeinsame Verzeichnung von Sprachressourcen und TS im VLO, wobei die Möglichkeit einer individualisierten Facettenauswahl als Ersteinstieg möglich ist.

Voraussetzung für die unterschiedliche Behandlung von Sprachressourcen einerseits und TS

andererseits wird jedenfalls eine entsprechende Angabe in den jeweiligen CMDI-Metadaten der Provider sein.

Perspektive

Der Weg einer vollständigen Auslagerung der TS aus dem VLO, wie unter Punkt 1 beschrieben, soll nicht weiter verfolgt werden. Welcher der beiden alternativen Wege eingeschlagen werden soll, wird bis Ende Dezember 2013 von der VLO-Taskforce beraten. Die Umsetzung des jeweils gewählten Lösungsansatzes wird bis Ende April 2014 erfolgen. Der getrennten Auswertung von TS und Sprachressourcen werden die Angaben im Feld “Ressource Class” zugrunde gelegt.

Gruppierung zusammengehöriger Ressourcen

Problem Statement

Zusammenhänge zwischen einzelnen Sprachressourcen können bislang im VLO (z.B. “X ist Teil von Y”; “Y umfasst A, B und C”) nicht dargestellt werden. Der CMDI-Header bietet über die Relationen *hasPart* und *isPartOf* die Möglichkeit, Teil-Ganzes-Beziehungen zwischen Ressourcen in den Metadaten abzubilden. Diese Relationen sind jedoch nicht ausreichend, da vielfältige andere Relationen zwischen Ressourcen denkbar sind, z.B. *source*, *isVersionOf* oder allgemein *Relation*.

Stand

Die Taskforce einigte sich auf das Desiderat, verschiedene Arten von Beziehungen zwischen Sprachressourcen abzubilden und diese im VLO auswertbar zu machen. Die Umsetzung muss noch erfolgen.

Perspektive

Für die Umsetzung sind zwei Schritte erforderlich: Zum einen soll die für 2014 vorgesehene neue CMDI-Version 2.0 um weitere mögliche Relationen erweitert werden. Zum anderen soll das VLO so angepasst werden, dass angegebene Relationen abgebildet und Ressourcen aufgrund ihrer Relationen zueinander auswertbar gemacht werden.

Die Erarbeitung eines Vorschlags für den Umgang mit Relationen zwischen Dokumenten und Korpora im CMDI-Header wird durch die *Gruppe Relationen* übernommen, die der VLO-Taskforce bis zum 20. Januar 2014 einen Vorschlag für eventuelle Anpassungen des CMDI-Headers für diese Aspekte vorlegt.

Dokumentation der VLO-Facetten und Richtlinien für CMDI-Profile

Problem Statement

Aktuell sind die Facetten des VLO nicht ausreichend dokumentiert. Die Dokumentation ist auf mehrere, teilweise widersprüchliche Quellen (Handbuch, informelle Facetten-Liste, Tooltips im VLO, Tool zum Testen des VLO-Mappings und Dokumentation unter http://lux13.mpi.nl/clarin/vlo/mapping/index.html) verteilt. Darüber hinaus gibt es keine klaren Richtlinien für die Erstellung und Wiederverwendung von CMDI-Profilen mit besonderer Berücksichtigung der Facetten des VLO.

Stand

Die Taskforce einigt sich auf das Desiderat, methodische und technische Entscheidungen in Bezug auf VLO Facetten umgehend zu dokumentieren.

Perspektive

Im ersten Schritt sollen die stabilen Facetten umfassend dokumentiert werden.
Das umfasst: (1) Eine kurze informelle Beschreibung der Funktion der Facette im VLO (Zweck, Semantik), (2) eine semi-formale Beschreibung der Wertebereiche, (3) eine exemplarische Beschreibung der Facetten anhand von konkreten Beispielen im VLO für typische Sprachressourcen.
In den nächsten Schritten sollen die technischen und methodischen Entscheidungen der Taskforce zeitnah in die Dokumentation einfließen.

Die Erarbeitung eines Konzepts für die umfassende Nutzerdokumentation zu den im VLO verfügbaren Funktionen und den Richtlinien für die CLARIN-konforme Erhebung von Metadaten wird von der Gruppe Dokumentation überneommen, die der VLO-Taskforce bis zum 20. Januar 2014 einen Vorschlag für das weitere Vorgehen vorlegt. Die darüber hinaus gehende weitere koordinative und redaktionelle Betreuung der Dokumentationsarbeiten obliegt ebenfalls der Gruppe Dokumentation.