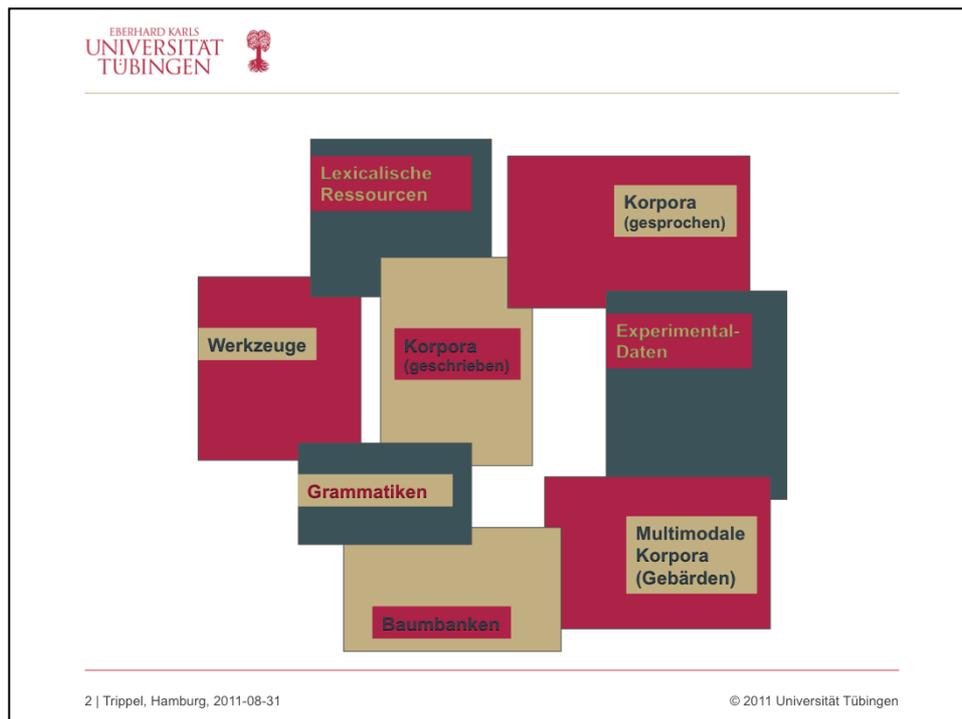




Archivierung von Forschungsdaten – langweilig, unwichtig oder was?

Thorsten Trippel
Computerlinguistik
Seminar für Sprachwissenschaft
Projekt: Zentrum für Nachhaltigkeit Linguistischer Daten



Sprachwissenschaftler machen Experimente, Aufnahmen, Annotationen, Grammatiken und verwenden diese Forschungsdaten für Unterricht, Qualifikationsarbeiten und Publikationen verwendet, teilweise über Jahre und Jahrzehnte. Wie können solche Ressourcen so lange verwendet werden? Können auch andere diese Daten verwenden. Teilweise werden Ressourcen in Drittmittelprojekten erzeugt und in anderen Projekten weiterverarbeitet und verwendet.

Aus dem Bereich der Lexika ist dies ein ganz übliches Verfahren und aus den vor-elektronischen-Zeiten wissen wir gut, wie das geht: man schreibt ein Buch und publiziert es, die Grunddaten werden auf Karteikarten gesammelt und mehr oder weniger oft durchgesehen und überarbeitet.

Heute haben wir wesentlich größere Datenmengen, die auf Computern gespeichert werden, wir erstellen sie, arbeiten damit und sind häufig ganz zufrieden damit, wie das läuft. Warum sollte man sich auch Gedanken dazu machen, ob und wie solche linguistischen Daten längerfristig verfügbar bleiben?



Archivieren? Vielleicht,
aber wir fangen doch
gerade erst an mit dem
Projekt

Niemand kann mein Daten
verwenden. Viel zu speziell

Ich habe die Daten auf 3.5"
Floppy/CD-ROM/Exabyte ...



Korpora im großen Maßstab werden noch nicht ganz so lange eingesetzt, aber diese Sammlungen von Texten mit zusätzlichen Informationen, z.B. zu Wortart oder Bedeutung oder Übersetzung, kennen wir, manchmal wissen wir auch schon, wieviel mühe es macht, so etwas über längere Zeiträume zu verwenden. Und so etwas, wie ein fertiges Korpus gibt es nicht, man hat immer noch Möglichkeiten, es zu vergrößern oder weitere Beschreibungsebenen hinzuzufügen. Und wenn man genügend Analysen damit gemacht hat, dann ist das Material so reich und so speziell für die eigenen Bedürfnisse bearbeitet, dass man sich kaum vorstellen möchte, dass es auf er Welt noch irgendeine andere Person gibt, die ähnlich detaillierte Informationen auch nur begrenzt interessant findet.

Außerdem: Die Daten werden auf Computern abgelegt, wenn man die jemand anderem geben will, kann man die Dateien einfach weitergeben. Und geht der Computer kaputt: Wohl dem der ein Backup gemacht hat. An diesem Beispiel mit den 3.5 Zoll Disketten oder Exabytes sehen Sie aber schon ein Problem: wer hat denn sowas heute schon noch? Und USB-Speichermedien sind nicht für ihre langfristige Zuverlässigkeit bekannt.



Wenn das so kompliziert ist mit dem Archivieren, warum sollte man sich dann überhaupt darum kümmern?

Für uns, für diejenigen, die sich intensiv damit auseinandersetzen, wie das passiert, ist die Frage natürlich einfach: Es ist interessant, anspruchsvoll, es ist nicht klar, wie man dafür sorgen kann, dass Video-Dateien in 10 Jahren noch lesbar – und zugänglich – sind. Kurz: Es ist wissenschaftlich spannend und faszinierend, wenn man seine Forschungsergebnisse nach Jahren ansehen kann, auf den zugrundeliegenden Daten neue Verfahren ausprobieren kann. Oder auch zum Vergleich ähnliche Daten nochmals erfassen kann, um Längsschnittstudien zu machen, die wirklich die unterschiedlichen Grunddaten erfassen. Z.B. Dialektale Veränderungen, Änderungen am Gesteninventar oder beim Ausdruck mittels Gebärden. Oder auch nur zu überprüfen, ob Ergebnisse, die man für Erklärungsbedürftig hält, unter neuen Ansätzen und Theorien nicht erklärbar sind.

Das sagt ein begeisterter Ressourcensammler. Aber es gibt auch ganz profane Gründe:

Wissenschaftsrat zu Forschungsprimärdaten

„Der Wissenschaftsrat empfiehlt den Trägereinrichtungen die umfassende und langfristige Archivierung qualitätsgesicherter und für die jeweilige wissenschaftliche Gemeinschaft langfristig relevanter Daten“

Empfehlung zu Forschungsinfrastrukturen in den Geistes und Sozialwissenschaften, Wissenschaftsrat, Berlin, 28.01.2011, S.57f

Der Wissenschaftsrat, ein zentrales Gremium zur Beratung der Bundesregierung, hat sich mit Forschungsprimärdaten und deren Zugänglichmachung beschäftigt. Die Ergebnisse als Empfehlung wurden Ende Januar veröffentlicht. Und die wollen das.

Freie Verfügbarkeit von Forschungsprimärdaten

„Forschungsprimärdaten bilden einen wertvollen Fundus an Informationen, die mit hohem finanziellem Aufwand erhoben werden. Je nach Fachgebiet und Methode sind sie replizierbar oder basieren auf nicht wiederholbaren Beobachtungen oder Messungen. **In jedem Fall sollten die erhobenen Daten nach Abschluss der Forschungen öffentlich zugänglich und frei verfügbar sein.**“

Aus: Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, DFG: Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss für Informationsmanagement, S. 2, Januar 2009

Und die sagen auch warum: Wenn die Öffentlichkeit bezahlt hat, will sie auch darauf zugreifen können. Daher sollen die Daten öffentlich zugänglich und verfügbar sein

Wissenschaftsrat zur Plagiatsverhinderung

„Der Wissenschaftsrat begreift die umfassende öffentliche Sammlung und Bereitstellung von Forschungsprimärdaten auch als ein probates Mittel der Qualitätssicherung in der wissenschaftlichen Praxis, welches hilft, wissenschaftlichen Betrug und Plagiate leichter zu identifizieren, da die Herkunft von Forschungsdaten aus Repositorien in jedem Falle offen gelegt und die „Autoren“ der Daten zitiert werden müssen.“

Empfehlung zu Forschungsinfrastrukturen in den Geistes und Sozialwissenschaften, Wissenschaftsrat, Berlin, 28.01.2011, S.59

Und zu einem Zeitpunkt, als das Thema nur von der Fachöffentlichkeit wahrgenommen wurde, sah der Wissenschaftsrat bereits eine Anti-Plagiatsfunktion: "Der Wissenschaftsrat begreift die umfassende öffentliche Sammlung und Bereitstellung von Forschungsprimärdaten auch als ein probates Mittel der Qualitätssicherung in der wissenschaftlichen Praxis, welches hilft, wissenschaftlichen Betrug und Plagiate leichter zu identifizieren, da die Herkunft von Forschungsdaten aus Repositorien in jedem Falle offen gelegt und die „Autoren“ der Daten zitiert werden müssen."

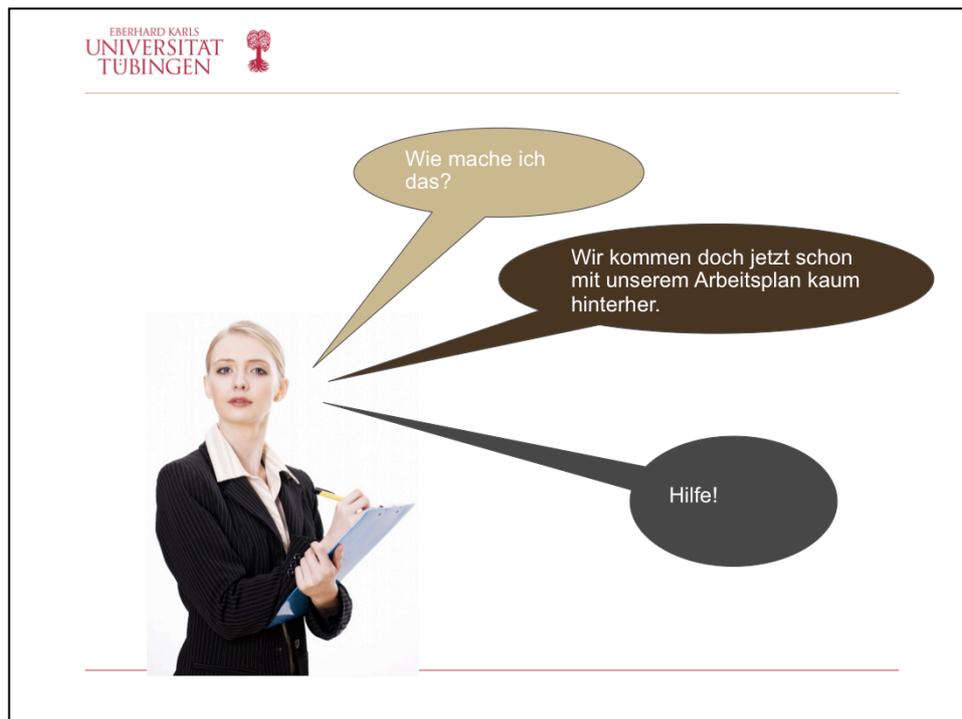
Mittels der Archivierung von Primärdaten sollen also auch Betrug und gefälschten Studien vorgebeugt werden.

Konsequenzen aus der Verfügbarkeit

„Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für **zehn Jahre** aufbewahrt werden.“

Aus:
Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, DFG: Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss für Informationsmanagement, S. 2, Januar 2009, Hervorhebung hinzugefügt.

Die Deutsche Forschungsgemeinschaft geht sogar von Datenhaltungen für 10 Jahre aus, und gibt damit den Wissenschaftlern den sprichwörtlichen schwarzen Peter zur Archivierung und Zugänglichmachung.



Der einzelne Wissenschaftler wird gerne damit alleine gelassen. Es wird einfach davon ausgegangen, es wird vorausgesetzt, dass er das tut, ohne dass es Anleitungen oder qualifizierte Hilfestellungen gäbe. Für Bücher gibt es normalerweise Bibliotheken mit Bibliothekaren, die große Bestände langfristig verwalten, die haben die Erfahrung, die katalogisieren, die wissen, wie das geht. In der Zwischenzeit gibt es aber eine Reihe von Projekten, die auch von den Geldgebern für Forschungsprojekte getragen werden, die sich darum bemühen, die Forschung im Bereich der Archivierung voranzutreiben und Infrastrukturen und Expertisen für die Archivierung aufzubauen

Zentrale Archivierung und Datenmanagement

- Datenmanagement von:
 - Forschungsprimärdaten
 - Metadaten
 - Verknüpfung von Forschungsprimärdaten und Publikationen
- Kooperation mit Rechenzentren, Bibliotheken, Archiven
 - Sicherung durch fortlaufende Migration auf aktuelle Hardware
 - Ausfallrisiken minimieren
 - Gewährleistung von Datenmanagement über das Projektende hinaus

Als zentrale Anlaufpunkte etablieren sich Archive, die häufig mehr oder weniger eng mit etablierten Hochschuleinrichtungen wie Rechenzentren und Bibliotheken zusammenarbeiten. Dabei gibt es sowohl lokale Einrichtungen, etwa hier in Hamburg das Hamburger Zentrum für Sprachkorpora; hier am Institut für Gebärdensprache das Archiv, das meines Wissens das beste und größte Archiv für Gebärdensprachendaten in Deutschland ist; es gibt aber auch große Verbundprojekte und Einrichtungen, die im Bereich der Standardisierung und Archivierung zusammenarbeiten. Allen gemein ist: es werden Forschungsprimärdaten erfasst, mit beschreibenden Informationen. Es werden, genau wie bei Büchern und Bibliotheksnummern, Identifikatoren vergeben, so dass auf die Forschungsprimärdaten verwiesen werden kann.

Und weil bestimmte Probleme mit der Archivierung grundlegender Art sind, sorgt die Kooperation mit zentralen Infrastruktureinrichtungen wie Bibliotheken und Rechenzentren für ein Maximum an Sicherheit, sowohl was die Datenhaltung und Backups, angeht, als auch die Verfügbarkeit.

Forschungsprimärdaten

- Experimentaldaten
 - Erhebungen
 - Fragebögen
 - Versuche
- „Quellen“
- Annotationen
- Rohanalysen
- Unpublizierte Projektberichte
- ...

Worüber sprechen wir? Also über verschiedene Daten, aber auch über Aufnahmen und andere Quellen mit ihren Annotationen und Rohanalysen. Natürlich ist es immer eine Frage, was Rohdaten sind und was bereits analysiert ist, aber lassen Sie mich das heute ein wenig lax angehen.

Metadaten: Dokumentation von Primärdaten

- Alle Informationen für:
 - (wieder) finden von Daten
 - (gefundene) Daten verstehen
 - Autorenschaft
 - Relevanz für die eigene Forschung
 - Format und Erstellungsdatum
- In einem spezifischen Format
 - Übersichtlich
 - Standardisiert
 - Durchsuchbar
- Maschinell verarbeitbar

Zur Beschreibung dieser Primärdaten werden Metadaten verwendet. Metadaten können daher als die Informationen angesehen werden, die man benötigt, um Daten, also z.B. eine Datei auf einem Computer, wieder zu finden oder um zu verstehen, um was es sich bei einer Datei, die man gefunden hat, wirklich handelt.

Dazu müssen diese Informationen in einem spezifischen Format vorliegen, das übersichtlich, standardisiert und durchsuchbar ist, so dass sowohl Suchmaschinen in einer maschinellen Verarbeitung darauf zugreifen können, als auch Menschen einen Eindruck davon erhalten, was diese Daten denn beinhalten. Beispiele für Metadaten sind etwa die bibliographischen Angaben in Publikationen, in Bibliotheken oder auch die Titelei.



Standardisierung im Infrastrukturbereich

- Kommunikation der Lösungen miteinander
 - Austausch von Metadatenkatalogen
 - Zugriff über Webservices aufeinander
- Gemeinsame Metadatenstrukturen
 - Nach Typ der Ressource
 - Zentrale Definitionen von Datenkategorien
- Datenformate
 - Wo möglich: Datentypspezifische Standardformate
 - Offen für Reimplementierung
- Lösungen
 - Authentifizierung und Autorisierung
 - Datenschutz
 - Zentrale Infrastruktureinrichtungen

OAI-PMH

Komponentenmodell für
Metadaten: CMDI

ISO 12620

Valides XML, TEI, ...
(Datentyp abhängig)

Shibboleth

Verfahrensverzeichnis

Um eine Inselbildung zu vermeiden, wird großer Wert auf Interoperabilität der Datenzentren und deren verschiedenen Lösungen gelegt. Um z.B. Metadaten auszutauschen, werden Verfahren und Protokolle verwendet, die sich im Bibliothekswesen bewährt haben. Das Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH) stellt so ein Protokoll dar, es erlaubt Metadaten auch in verschiedenen Formaten und unterschiedlichem Detailreichtum auszutauschen.

Um über die strukturierten Metadaten hinreichend gut suchen zu können, ist es wiederum notwendig, dass es gemeinsame Strukturen gibt. In der Entwicklung hat sich allerdings herausgestellt, dass die Beschreibungen von Büchern und anderen traditionellen Medien sich sehr stark davon unterscheidet, was man z.B. für die linguistische Feldforschung braucht, wo neben Audio- und Videoaufnahmen, Texten, Fragebögen, geographische Informationen, etc. beschrieben werden. Linguistische Software wiederum hat ebenso stark davon abweichende Beschreibungsebenen. Um möglichst viele Metadaten ähnlich zu gestalten, werden Beschreibungsebenen zu Blöcken zusammengefasst, die für verschiedene Ressourcentypen relevant sind. So kann etwa ein Autor, Ursprungsprojekt, der Ressourcename als Einheit zusammengefasst für viele Ressourcentypen verwendet werden, diese gibt es meist auch für traditionelle Medien in Bibliotheken, wo solche Datentypen



Datensicherheit

- Sensible Daten
 - Biographische Informationen
 - "Unfertige" Daten
 - Copyright-Probleme
- Lösungen
 - Authentifizierung und Autorisierung
 - Wer darf was lesen/schreiben?
 - Wie weiß man, dass es die vorgebliche Person ist?
 - Datenschutz
 - Schutz vor Hackern
 - Verfahren zum Datenschutz
 - Zentrale Infrastruktureinrichtungen

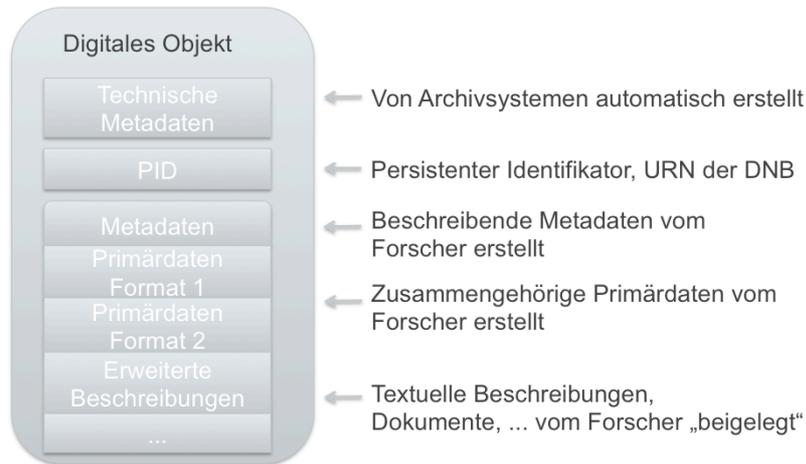
Shibboleth

Verfahrensverzeichnis

Lassen Sie mich auf die beiden unteren Boxen, eingehen, die direkt mit der Datensicherheit und der Privatsphäre zu tun haben. Da wir im Bereich der Sprache teils mit sehr sensiblen Daten umgehen, müssen wir dafür auch Vorkehrungen treffen. Daten wie Korpora oder Experimentaldaten enthalten häufig Rückschlüsse auf biographische Informationen. Auch sind Daten, über die noch nicht publiziert wurde, häufig noch vertraulich, manchmal sind auch andere Eigentumsrechte betroffen. Der Besitzer darf natürlich darauf zugreifen, aber es ist auch notwendig dafür zu sorgen, dass nur berechtigte Personen zugreifen dürfen. Technische Verfahren wie das Shibboleth-System, was von vielen Universitäten unterstützt wird, erlauben dabei Nutzer zuverlässig zu erkennen, Archivsysteme müssen dann nur noch feststellen, ob diese Person auf Daten zugreifen darf. Zentrale Infrastrukturen wie Rechenzentren haben auch ein große Arsenal zur Hand, um die Daten technisch zu schützen, z.B. gegen Diebe und Hacker, für die Zugangsbeschränkung, also wer etwas lesen darf, müssen aber noch weitere Instrumentarien verwendet werden. Wir verwenden dafür Verfahrensverzeichnisse, wie sie auch von Behörden und ähnlichen Einrichtungen verwendet werden, in denen detailliert geregelt wird, wer das Recht hat, anderen Zugangsrechte zu geben und welche Voraussetzungen der zu erfüllen hat. Für einen Einzelforscher ist dieser Overhead zu groß, für Spezialeinrichtungen mit Verbindung zu Infrastruktureinrichtungen ist dies zu befähigen.



Ablage in Verzeichnissen und Archivsystemen



Forschungsprimärdaten und beschreibende Metadaten werden zusammen zu sogenannten digitalen Objekten zusammengefasst und im Archivsystem mit Bezug zueinander abgelegt werden. Wie in einer Bibliothek, in der ein neues Buch dadurch aufgenommen wird, dass die bibliographischen Informationen in den Katalog eingefügt werden, Bibliotheksspezifische Angaben in die Titelei eingestempelt werden und das Buch einem Standort zugewiesen wird, werden auch elektronische Primärdaten in ein Archivsystem dadurch aufgenommen, dass beschreibende Metadaten und die Primärdaten zusammen aufgenommen werden, technische Metadaten und ein Identifikationsmerkmal, ein Persistenter Identifikator, angelegt wird.

Faceted Browser

- Basierend auf OAI-PMH-Feeds
- Verwendung fachbezogener Facetten
 - z.T. vom Ressourcentyp gesteuert
 - Allgemeine Facetten für Suchraumbeschränkung
- Verwendung des Komponentenmodells für Metadaten

Nalida Faceted Browser
DIP 633, 918 Tübingen

Facet: modality (7)		Facet: language (10)	
modality	Documents	language	Documents
Other	4	Arabic	1
Portuguese resources	440	Arabic	1
Sign	77	Arabic	1
Speech	999	Arabic	1
Unspecified	11	Bahasa Indonesia	1
verbal_and_non-verbal_interaction	117	Basque	3
writing	93	Bahasa Portugese	1
		Bahasa Sunda	23
		Bulgarisch	1

Facet: resources (8)		Facet: country (8)	
resources	Documents	country	Documents
corpus	2342	France	93
general_corpus	1	Germany	10149
LexicalResource	1	Netherlands	21
Lexicon	2	Spain	1
Text	214	Sweden	8
WrittenCorpus	19	United Kingdom	21

Facet: organization (8)		Facet: origin (12)	
organization	Documents	origin	Documents
Hamburg School, University of Applied Sciences	10	ban	7417
Max Planck Institute for Psycholinguistics	2021	babel	1
Radboud University Nijmegen	87	Bibliografische	2472
SFB 441	33	icclib_data	139
SFB 532	28	Humboldt	3
Universität Tübingen	2	Lexikon	142
University of Leoben	142	uflib1	33
University of Southwest	2	uflib2	1
		uflib3	27

Fertig

Diese Informationen können aber auch für die Suche nach Ressourcen verwendet werden. Hier zum Beispiel haben wir ein Suchinterface für Sprachressourcen, die auf solchen Ressourcen aufbauen. Abhängig vom Typ einer Ressource werden unterschiedliche Unterkategorien verwendet, die man als Facetten bezeichnet



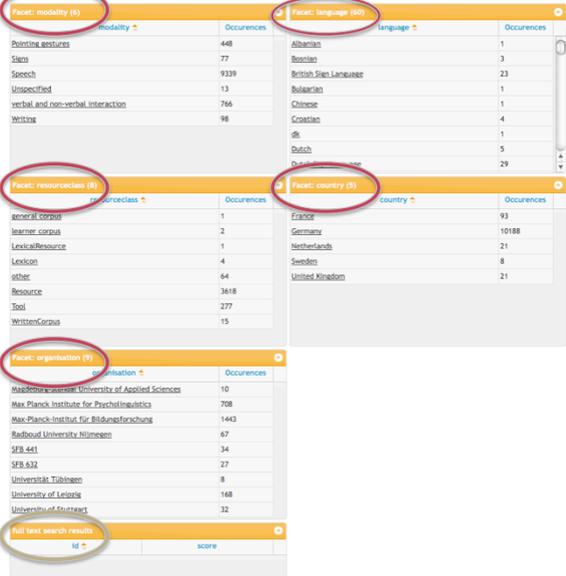
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Zentrum für Nachhaltigkeit linguistischer Daten: Suche nach Ressourcen

Ein Faceted Browser mit bedingten und unbedingten Facetten

- Unbedingte Facetten
 - Modality
 - Language
 - Resourceclass
 - Country
 - Organisation
- Special: Full text search



18 | Trailblazing, DH 2011

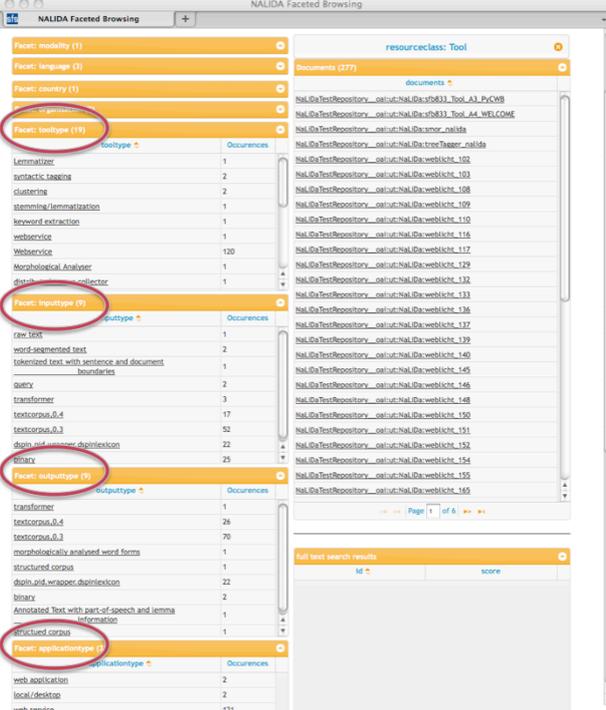
Unbedingte Facetten sind diejenigen, die unabhängig von weiteren Einschränkungen verwendet werden, z.B. Modalität, Sprache, Organisation, etc. Im Allgemeinen können alle Facetten, die in allen Ressourcen verwendet werden, so benutzt werden, in der Praxis muss man aber eine Vorauswahl treffen, um die Übersichtlichkeit zu gewährleisten.

Die Zahl in jeder Zeile gibt an, wie viele Ressourcen mit den aufgeführten Eigenschaften im Suchraum vorhanden sind.



Bedingte Facetten

- Tools
 - Tool type
 - Input type
 - Output type
 - Application type
- Unbedingte Facetten minimiert
- Andere bedingte Facetten bei anderen Ressourcentypen



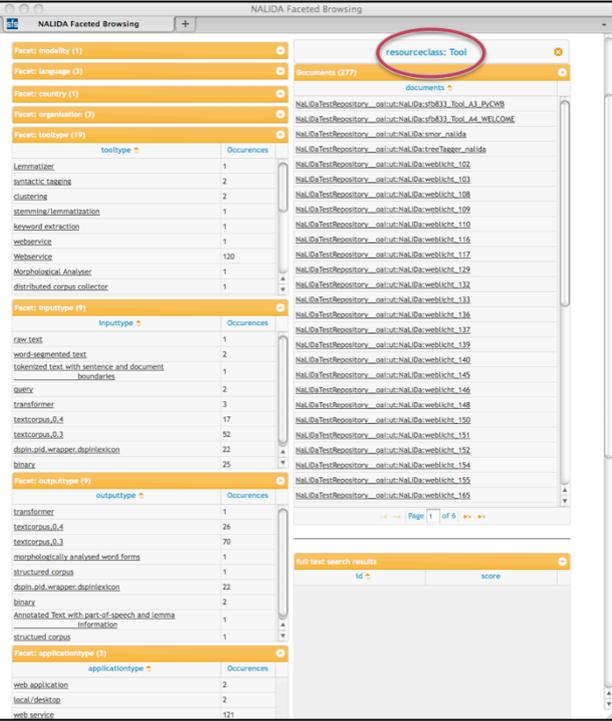
19 | Trailblazing, DH 2011

Abhängig von der Ressourcenart werden auch bedingte Facetten definiert. Für Textkorpora sind z.B. Genre-Informationen zur Auswahl relevant, für Computerprogramme, wie hier dargestellt, ist der Input und Output wichtiger. Auch hier müssen Fachleute eine Vorauswahl treffen.



- Vorausgewählte Facetten mit Werte
- Wertemengen

20 | Trailblazing, DH 2011



Note also that after the selection of facets, the facets and the selected value is still communicated to a user and the user may deselect these facet. Additionally, at least below a threshold, the documents that match to the facet values are listed. These lists can be rather long, hence a threshold should be selected based on human expertise.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Zentrum für Nachhaltigkeit linguistischer Daten: Suche nach Ressourcen

Ein Faceted Browser mit bedingten und unbedingten Facetten

Wenn man alles ausgewählt hat...

Facet: language (0)

country: Germany

modality: Writing

organisation: University of Stuttgart

resourceclass: Tool

tooltype: Lemmatizer

Documents (1)

TreeTagger - a language-independent part-of-speech tagger

21 | Trailblazing, DH 2011

Im Idealfall erhält man nach der Auswahl der für einen interessanten Facetten nur eine kleine Anzahl von Dokumenten, hier z.B. habe ich nach einem Lemmatizer aus Stuttgart für geschriebene Ressourcen gesucht, in unserem Suchraum bleibt nur der TreeTagger übrig.

Und damit hat man dann zugriff auf den vollständigen Katalogeintrag.



Katalog linguistischer Ressourcen

Document View: GermaNet

Document View: GermaNet

Document View: GermaNet

General Info | Creation | Documentation | Access | Lexicon Content | Size Info | Text-Technical

General Information

Resource Name: GermaNet
Resource Title: GermaNet: Ein lexikalisch-semantisches Wortnetz
Resource Class: Lexicon
Version: 5.3
Publication Date: 1997-01-01
Last Update: 2010-04-01
Time Coverage: synchron
Legal Owner: Universität Tübingen
Location: Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen,false,false
GermaNet ist ein lexikalisch-semantisches Wortnetz, das Nomina, Verben und Adjektive des Deutschen beschreibt. Dabei werden lexikalische Einheiten, die dasselbe Konzept ausdrücken, in einem Synset zusammengefasst, und die zwischen den Synsets bzw. den lexikalischen Einheiten bestehenden semantischen Relationen beschrieben. GermaNet orientiert sich an den grundlegenden Strukturierungsprinzipien des englischen WordNet® und kann als ein online-Thesaurus oder eine "light-weight ontology" betrachtet werden.
Description:
Contact: germanetinfo@sfs.uni-tuebingen.de

Hier nochmal ein Katalogeintrag, wie man ihn als Ergebnis erhält. Wenn man also so einen Eintrag erhält und dadurch eine Ressource erhält, was kann man damit machen? Man kann diese Ressource in Publikationen und Analysen zitieren.

Wissenschaftsrat zur Archivierung und Zitierbarkeit

„Der Wissenschaftsrat fordert insbesondere die Forschungsförderer auf, Anreize zu schaffen, um qualitativ hochwertige Daten zu archivieren und langfristig zu sichern. Zu diesem Zweck sollten Referenz- und entsprechende Zitationsmöglichkeiten für Datensätze aufgebaut werden. *Persistent identifier* (PI) bzw. *digital object identifier* (DOI) erlauben eine eindeutige Identifizierung und Zitierbarkeit digital hinterlegter Daten selbst dann, wenn sie ihre Speicherorte (in der Regel referenziert über den *uniform resource locator*, URL) wechseln.“

Empfehlung zu Forschungsinfrastrukturen in den Geistes und Sozialwissenschaften, Wissenschaftsrat, Berlin, 28.01.2011, S.58

Der Wissenschaftsrat sagt dazu: Der Wissenschaftsrat fordert insbesondere die Forschungsförderer auf, Anreize zu schaffen, um qualitativ hochwertige Daten zu archivieren und langfristig zu sichern. Zu diesem Zweck sollten Referenz- und entsprechende Zitationsmöglichkeiten für Datensätze aufgebaut werden. *Persistent identifier* (PI) bzw. *digital object identifier* (DOI) erlauben eine eindeutige Identifizierung und Zitierbarkeit digital hinterlegter Daten selbst dann, wenn sie ihre Speicherorte (in der Regel referenziert über den *uniform resource locator*, URL) wechseln.

Da elektornische Primärdaten anders aufgebaut und anders archiviert werden, müssen sie auch anders als in Bibliotheken referenziert werden.

Referenzsysteme

- URL: Uniform Resource Locator (<http://www...>)
- URN: Uniform Resource Name (<urn:nbn:de:bsz:21-Id-000193>)
- DOI: Digital Object identifier ([doi:10.4242/BalisageVol7.Broeder01](https://doi.org/10.4242/BalisageVol7.Broeder01))
- Handle: <http://nascent.nature.com/openhandle/handle?format=json&id=10100/openhandle>

Der Vorteil von elektronischen Ressourcen ist, dass man sie eindeutig und einfach referenzieren kann. Im Gegensatz zu Büchern, deren bibliographischen Angaben man braucht, um sie im Bibliothekskatalog zu finden, wodurch man ihren Standort finden kann.

In der Welt des Internets kennen wir die URL, die leider nicht sehr persistent ist: Webserver ändern sich. Um das Problem zu lösen, gibt es verschiedene Serviceprovider, die dauerhafte identifikatoren vergeben, ändert sich dann ein Server, muss das nur bei diesem Serviceprovider geändert werden. Die Deutschen Nationalbibliothek etwa verwendet das URN-System, ursprünglich vom Buch kommend, DOI werden insbesondere von Verlagen verwendet, Handles sind ein System, bei dem der Identifikator auch gleichzeitig eine URL ist: Ein click und man wird zum Standort der richtigen Ressource weitergeleitet.

URNs, DOI und Handle kann man angeben, um auf Ressourcen dauerhaft zu verweisen, ohne Datumsangabe oder ähnliches, weil sie immer eindeutig zu genau einer Ressource gehören.



Was will NaLiDa

Archivierung von Forschungsdaten – spannend, wichtig und einfach zu bewerkstelligen...

- Motivation
 - Gefordert von Forschungsförderern
 - Neue Fragestellungen
- Datenzentren für den Overhead
- Metadaten zum Finden und gefunden werden
- Verweise über PIDs
- Informationen beim NaLiDa-Projekt
 - nalida@sfs.uni-tuebingen.de
 - <http://www.sfs.uni-tuebingen.de/nalida>

Fotos

- Man reading Image: Arvind Balaraman / FreeDigitalPhotos.net
 - Business Woman Taking Notes Image: Michal Marcol / FreeDigitalPhotos.net
-