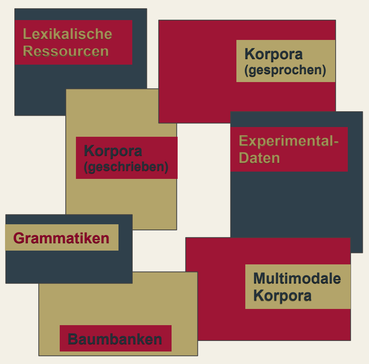
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
Seminar für Sprachwissenschaft (SfS)

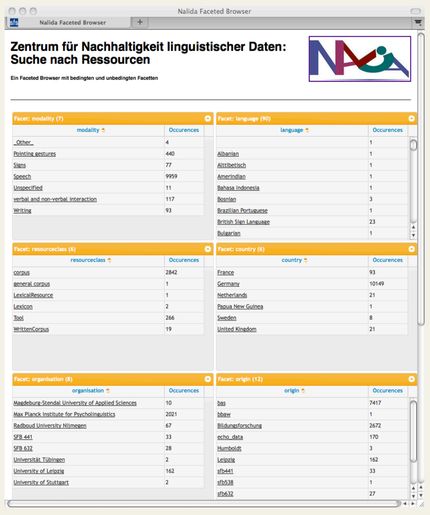
NACHHALTIGKEIT: AUCH FÜR LINGUISTISCHE DATEN!

Reinhild Barkey, Erhard Hinrichs, Christina Hoppermann, Thorsten Trippel, Claus Zinn

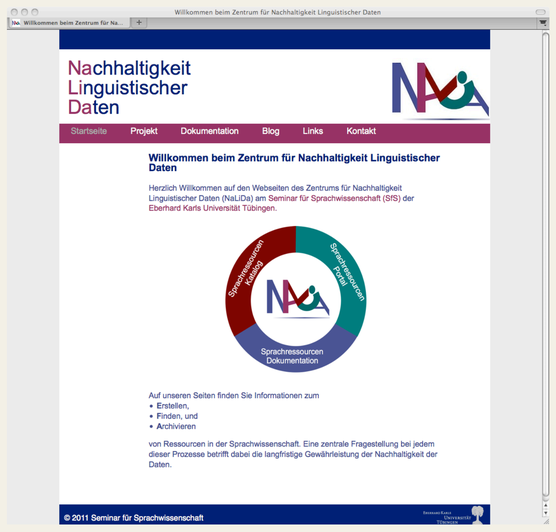


„Eine Ressource ist nachhaltig, wenn sie auch in Zukunft trotz der möglichen Veränderungen technischer Standards, Metadatenschemas oder der nicht mehr präsenten Kontaktpersonen, die eine Ressource erstellt oder verwaltet haben, noch auffindbar und verwendbar ist.“

(Aus: Glossar zu Sprachressourcen, NaLiDa-Projekt)



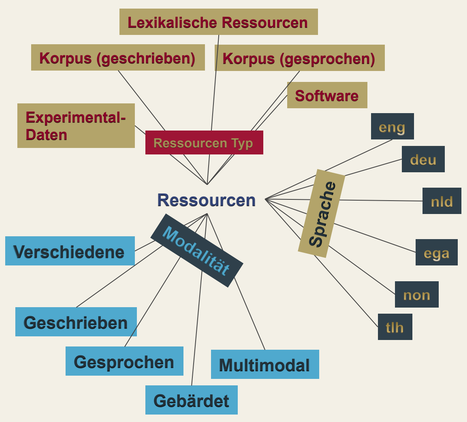
Metadaten liegen zur Archivierung und zur Suche in stark strukturierten XML-Formaten vor. Diese XML-Formate sind ebenfalls abhängig von der Klasse einer Ressource. Einige Beschreibungsebenen sind aber verhältnismäßig allgemein und werden bei vielen Sprachressourcen Verwendung finden. Trotzdem gehen auch hier die Beschreibungen über traditionelle Bibliographien hinaus. Daher eignen sich bibliographische Metadaten wie Dublin Core nur sehr eingeschränkt für Sprachressourcen.



Sprachressourcen wie Korpora, Lexika, Grammatiken, Computerprogramme oder Ergebnissammlungen werden in der linguistischen Forschung immer wichtiger. Jedoch ist ihre Erstellung häufig sehr komplex, Informationen gehen auf lange Sicht verloren oder können nicht mehr verarbeitet werden. Um dies zu vermeiden, werden sie in Repositorien archiviert und somit sowohl durchsuchbar als auch langfristig verwendbar gemacht.

Unterschiedliche Typen von Ressourcen erfordern unterschiedliche Herangehensweisen. Dabei ist die Typisierung aber keineswegs eindeutig, unterschiedliche Klassifikationskriterien können verwendet werden.

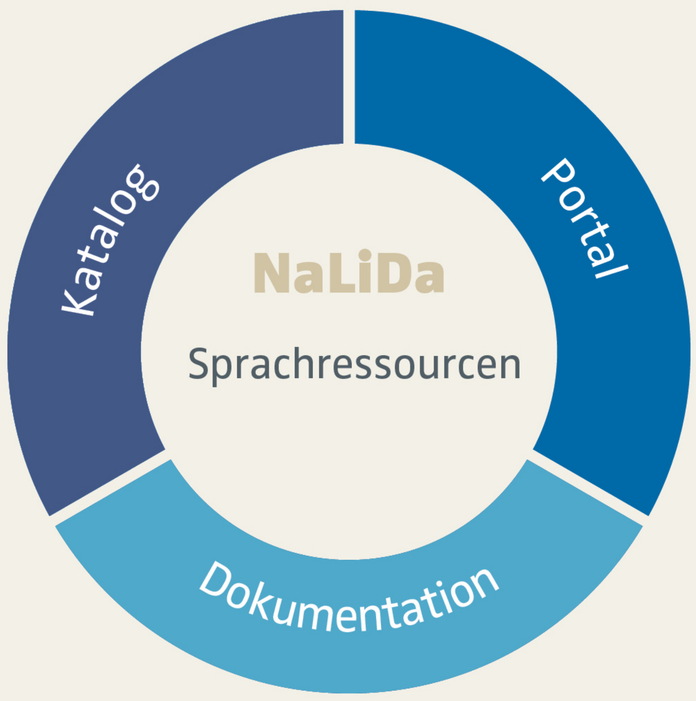
Unabhängig von den Ressourcentypen muss die Infrastruktur aber in der Lage sein, mit solchen Daten umzugehen und sie langfristig zu speichern. Dazu gibt es Nachhaltigkeitsplattformen, die die technische Abwicklung der Datenhaltung, den Zugang für Benutzer und die Verbindung zu Suchmaschinen übernehmen. Der Benutzer muss idealerweise nur noch seine Daten in der Plattform abspeichern.



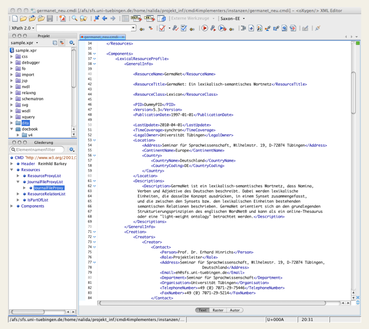
„Forschungsinfrastrukturen leisten in den Geistes- und Sozialwissenschaften einen wichtigen Beitrag zum Erkenntnisgewinn über gesellschaftliche Problemlagen und zur Erschließung unseres kulturellen Erbes.“

(Aus: *Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften stärken,* Pressemitteilung des Wissenschaftsrats, Nr. 3, Berlin, 31.01.2011)

Die Klassifikation von Ressourcen spiegelt sich in Suchfunktionen wieder, die auf Strukturen aufbauen, die in Ressourcenbeschreibungen – Metadaten – verzeichnet sind. Auf diese Weise kann man eine Suche nach Ressourcen z.B. nach Modalität, Sprache oder Genre einschränken. In Abhängigkeit von den gewählten Einschränkungen können auch neue Strukturen erscheinen. So sind Annotationsschemata für viele Ressourcentypen nicht relevant, jedoch aber für Korpora zur weiteren Einschränkung des Suchraumes.



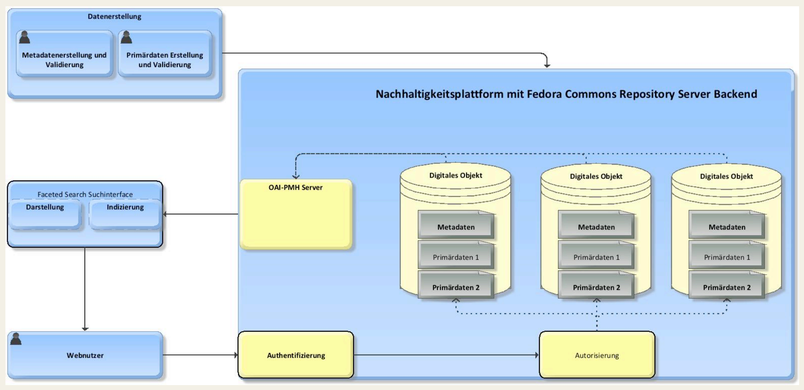
http://www.sfs.uni-tuebingen.de/nalida



Technische Ansichten auf Metadaten sind nur für Computerprogramme (und Programmierer) interessant. Im Rahmen eines Sprachressourcenportals, das sich an Sprachwissenschaftler wendet, die sich für Sprachressourcen oder deren nachhaltige Bereitstellung interessieren, geht es darum, einen Überblick über existierende Ressourcen zu schaffen. Wenn man die Strukturen der Metadaten ausnutzt, kann man verschiedene Perspektiven und Aspekte der Beschreibung bündeln. Eine Möglichkeit zur Visualisierung bieten z.B. Karteireiter.

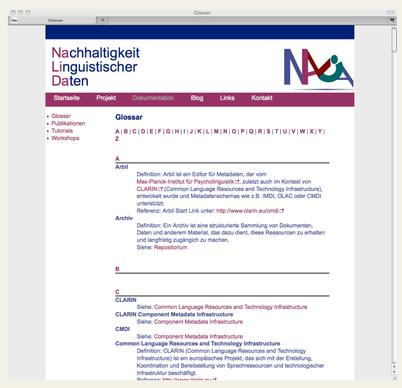
„Der Wissenschaftsrat begreift die umfassende öffentliche Sammlung und Bereitstellung von Forschungsprimärdaten auch als ein probates Mittel der Qualitätssicherung in der wissenschaftlichen Praxis [...].“

(Aus: Empfehlung zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften, Wissenschaftsrat, Berlin, 28.01.2011, S.59)



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

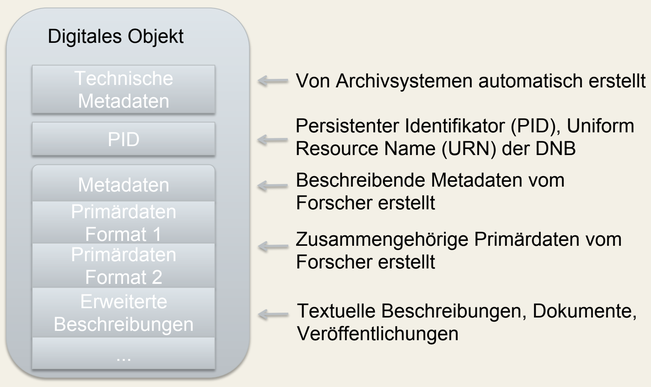
(Aus: Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, DFG: Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss für Informationsmanagement, S. 2, Januar 2009)

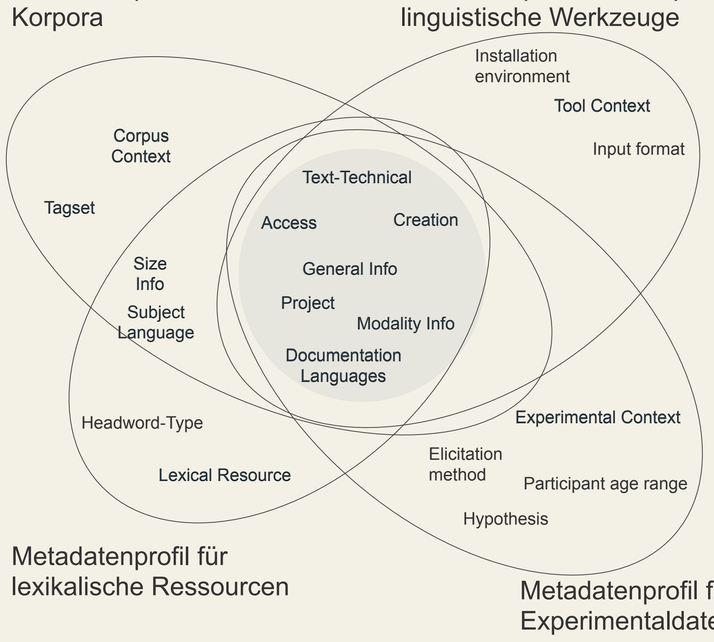


Suchfunktionen und Metadatenansichten innerhalb eines Sprachressourcenportals erlauben den Zugriff auf Forschungsprimärdaten, wie z.B. Korpora, lexikalische Ressourcen, Experimentaldaten, Dokumentationsdaten, etc. Dabei wird vom Archivsystem gewährleistet, dass Primärdaten tatsächlich nur von denjenigen gesehen werden, die dazu auch berechtigt sind. Dies ist insbesondere bei Daten entscheidend, die anderweitig veröffentlicht werden, aus ethischen Gründen nicht veröffentlicht werden dürfen oder es andere Hindernisse für eine generelle Freigabe gibt. Trotzdem können solche Ressourcen über Metadaten gefunden werden, so dass der Kontakt zu den Erstellern hergestellt werden kann.

Zentraler Bestandteil der Nachhaltigkeitsplattform sind digitale Objekte. Ein digitales Objekt fasst die Ressource samt ihrer Beschreibungen und aller zugehörigen Informationen zusammen. Auch wissenschaftliche Artikel, die die Ressource erläutern oder als Grundlage verwenden, interne Berichte und Notizen, etc. können vom Ressourcenersteller hinzugenommen werden. Zugangsbeschränkungen sorgen dafür, dass nur autorisierten Kreisen ein Zugang zu den Ressourcen gewährt wird.

Digitale Objekte können dabei immer eindeutig durch persistente Identifikatoren, kurz PIDs, erkannt und adressiert werden. Hierzu haben Bibliotheken wie die Deutsche Nationalbibliothek (DNB) Nummernkreise und Formate für Identifikatoren entwickelt.





Neben der Bereitstellung von Ressourcen und einer Suche über diese Ressourcen dient das Sprachressourcenportal auch dem Austausch von Informationen zur nachhaltigen Erstellung und Archivierung von Ressourcen. Um die unterschiedlichen Teildisziplinen bei der Kommunikation zu unterstützen, gibt es neben Dokumentations- und Schulungsmaterialien auch ein Glossar, um für die Sprachressourcen relevante Begriffe schnell zugänglich zu machen. Zur Nachhaltigkeit gehört auch, Ressourcen verstehbar zu machen und dafür zu sorgen, dass die verwendeten Definitionen dokumentiert sind.

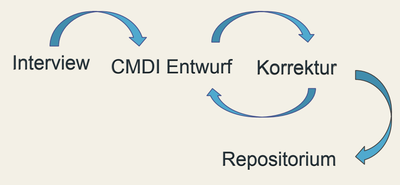
Zur langfristigen Archivierung gehört nicht nur die technische Aufbewahrung, die Ablage der Primärdaten in sogenannten Datenströmen, sondern auch die Beschreibung der Ressourcen. Unterschiedliche Datenströme können auch zu unterschiedlichen Ressourcentypen gehören.

Für jede Klasse von Ressourcen gibt es unterschiedliche Profile, um sie zu beschreiben. Gleichzeitig werden Datenkategorien – zu Komponenten zusammengefasst – auch von verschiedenen Profilen verwendet. Komponenten, die häufig eingesetzt werden, bilden dabei einen Kern, der mit Dublin Core vergleichbar ist aber für Sprachressourcen weitere Datenkategorien beinhaltet. Der Kern ist dabei weder exklusiv noch universell.

Im NaLiDa-Projekt wurde bisher in der Regel ein mehrstufiges Verfahren zur Erstellung von Metadaten verwendet. Dies dient im Wesentlichen zum Training der Datenersteller, die Metadaten selbst erfassen zu können, aber auch der Aufnahme neuer Ressourcentypen. Basierend auf ersten Gesprächen und Ressourcenbeschreibungen wird ein Entwurf für ein Metadatenschema mit einem Mustermetadatensatz entwickelt. Bei Bedarf werden auch weitere Komponenten und Profile entwickelt.

Der Metadatensatz kann dann ohne weitere technische Vorkenntnisse vom Ressourcenersteller korrekturgelesen und überprüft werden. Bei Bedarf kann es noch Anpassungen an das Metadatenformat geben.

Ist der Metadatensatz vollständig, kann er mit den Primärdaten zusammen in die Nachhaltigkeitsplattform aufgenommen werden. Ein späteres Bearbeiten ist natürlich jederzeit noch möglich.



Universität Tübingen · Philosophische Fakultät · Seminar für Sprachwissenschaft
Wilhelmstraße 19 · 72074 Tübingen
www.sfs.uni-tuebingen.de/nalida
nalida@sfs.uni-tuebingen.de