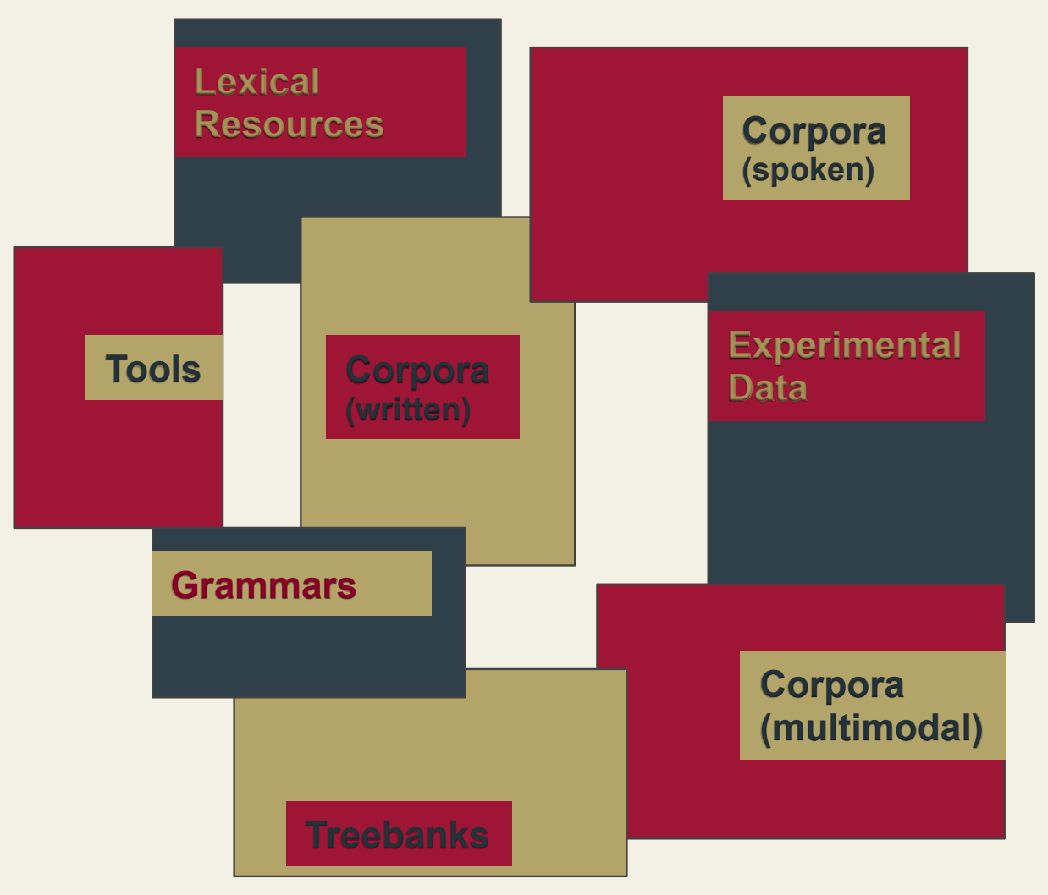
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Humanities
Department of Linguistics (SfS)
General and ompCutational Linguistics

Presented at:
Balisage 2011
The Markup Conference
August 2011, Montréal, Canada

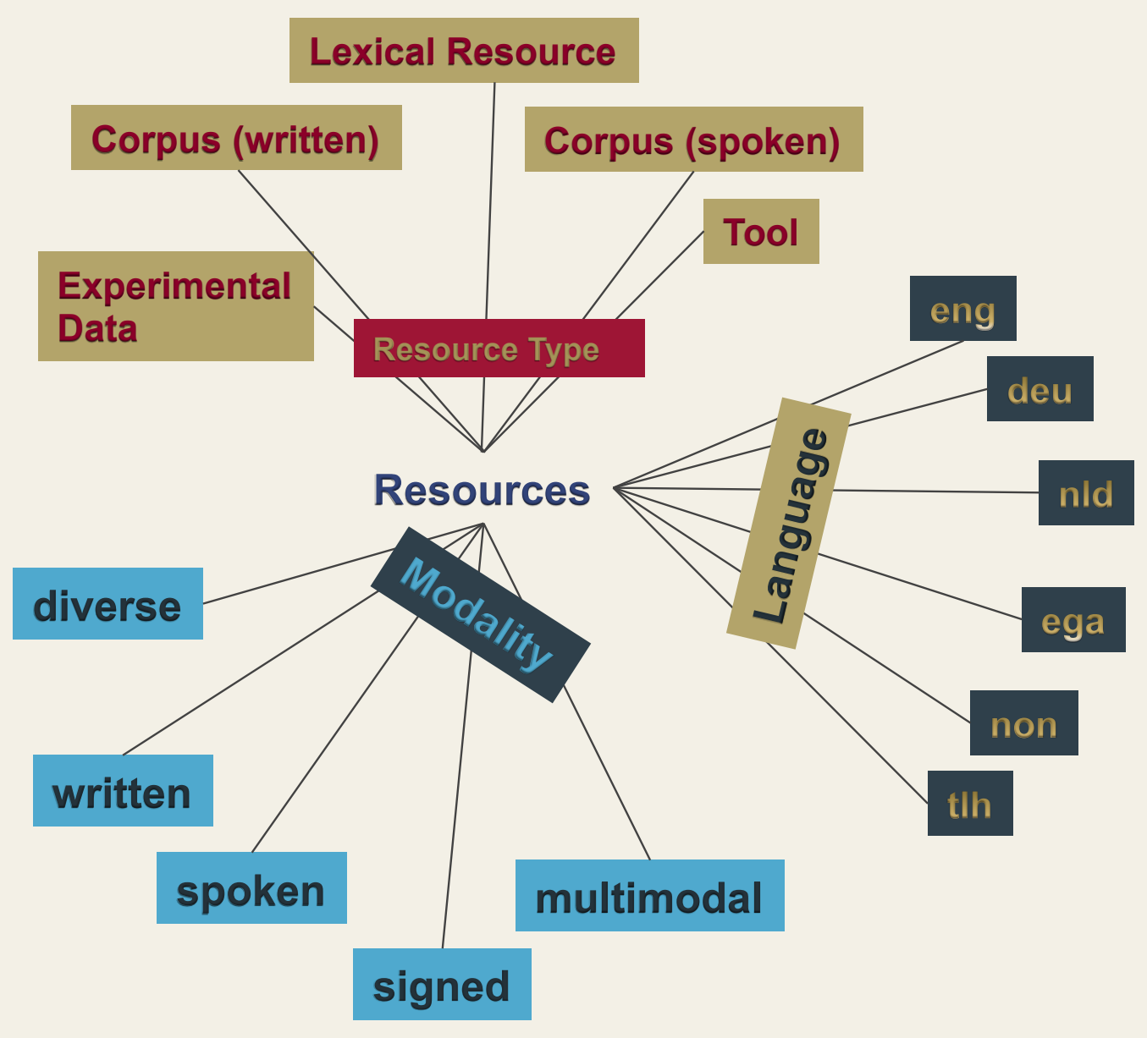
SUSTAINABILITY OF LINGUISTIC RESOURCES

Erhard Hinrichs, Christina Hoppermann, Thorsten Trippel, Claus Zinn



> *"A language resource is sustainable if it is findable and usable independent of the development of technical standards, metadadata schemes and contact persons creating or maintaining a resource."*
>
> *„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."*
>
> (From: Glossary of language resources, NaLiDa-Projekt)

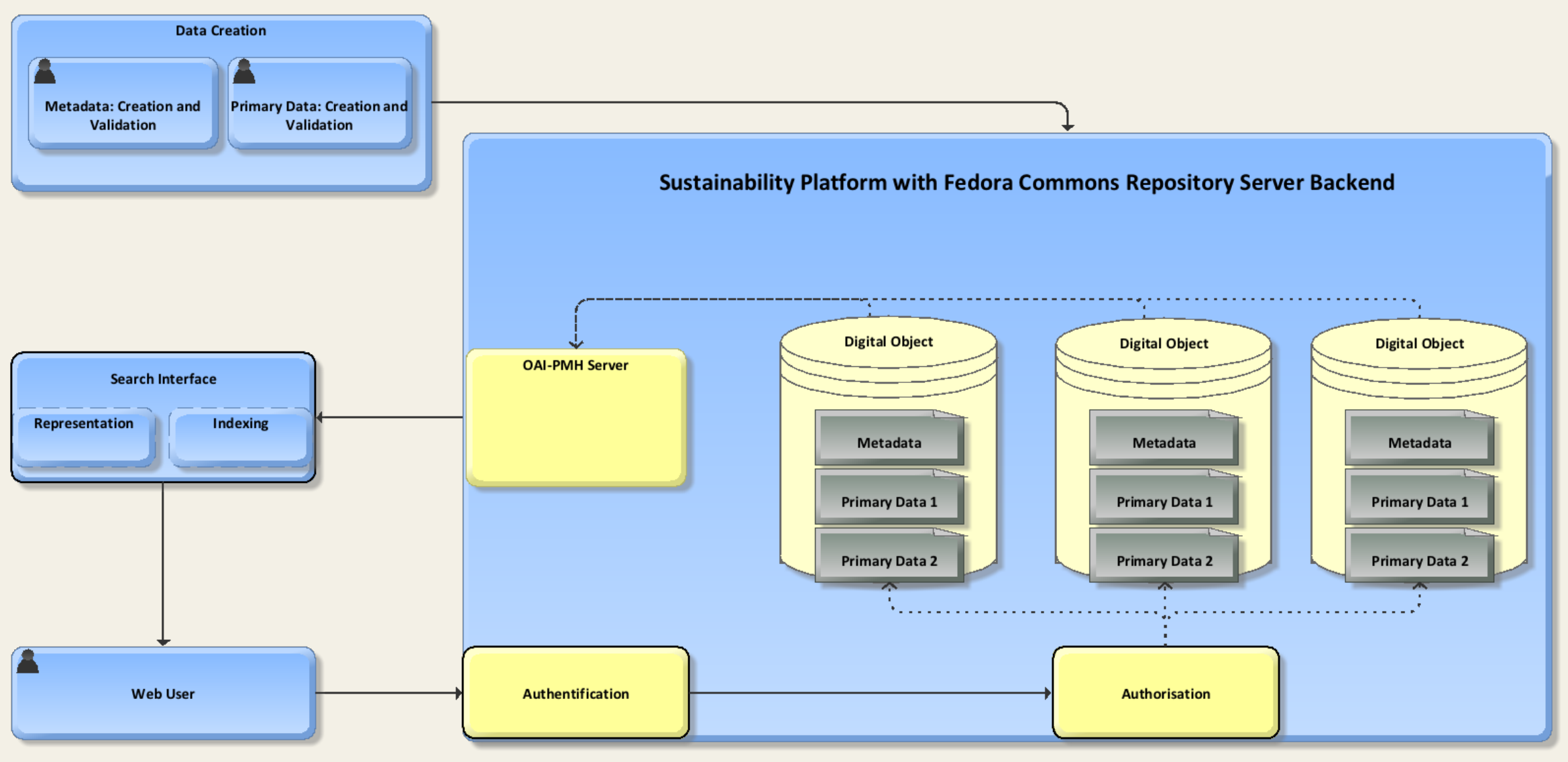
Linguistic resources such as corpora, lexicons, grammars, tools or experimental data become more and more important in the linguistic research community. Their creation is often complex, bits of information get lost or cannot be processed anymore. To avoid these circumstances, resources are archived in repositories. Thereby, they are made both searchable and (re-)usable on the long run.



As there are different types of resources, there is also a need of different approaches to describe these resources. A corpus requires other descriptive elements than a tool, for instance. The assignment of a type to a resource is not necessarily unique and depends on classification criteria.

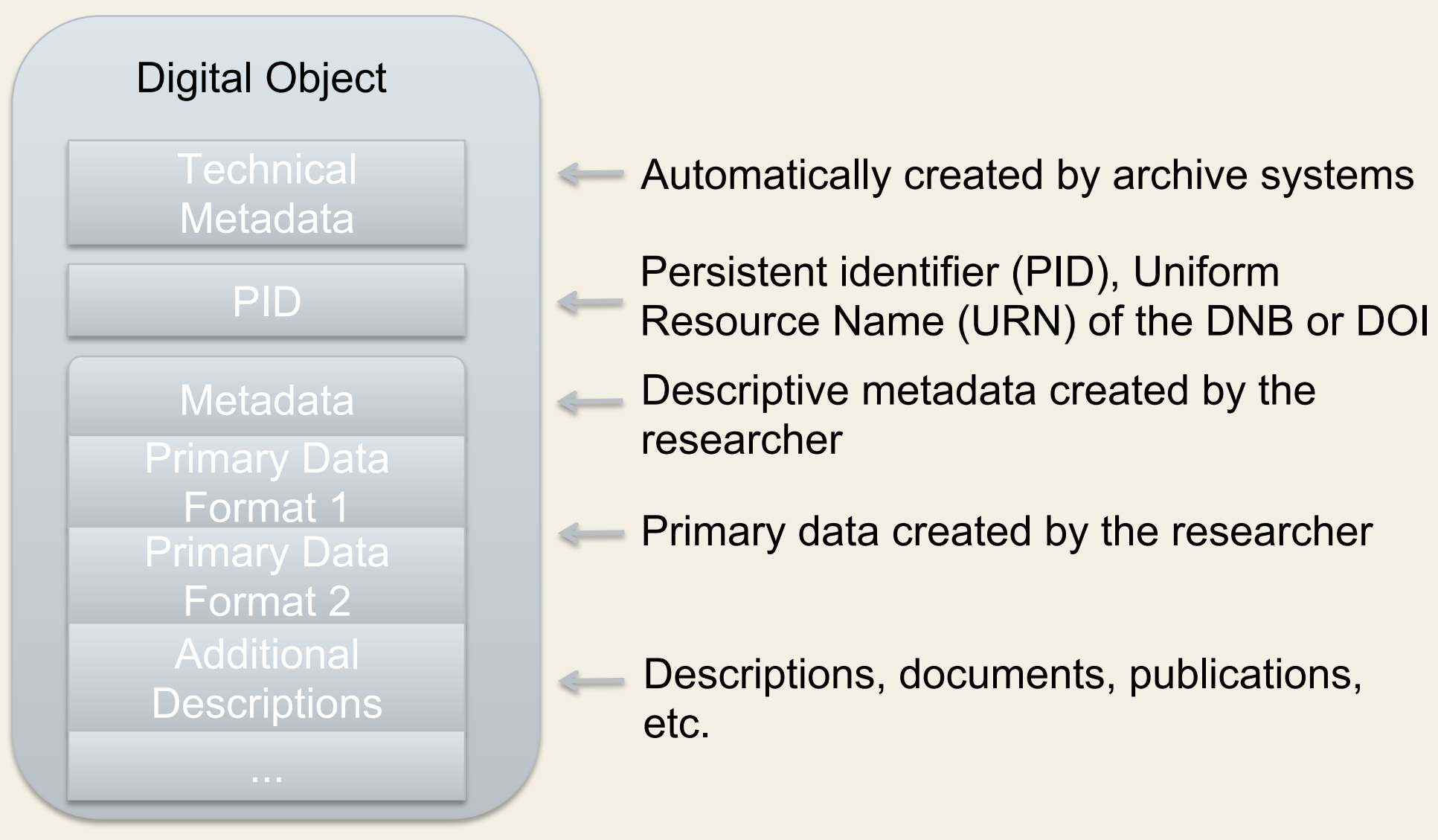
Independent of the resource type, the infrastructure needs to be able to handle and store these data types for the long term. For this purpose, there are sustainability platforms that can be used for data management, user access and as an interface to search engines. The users then only need to save their data to the platform.

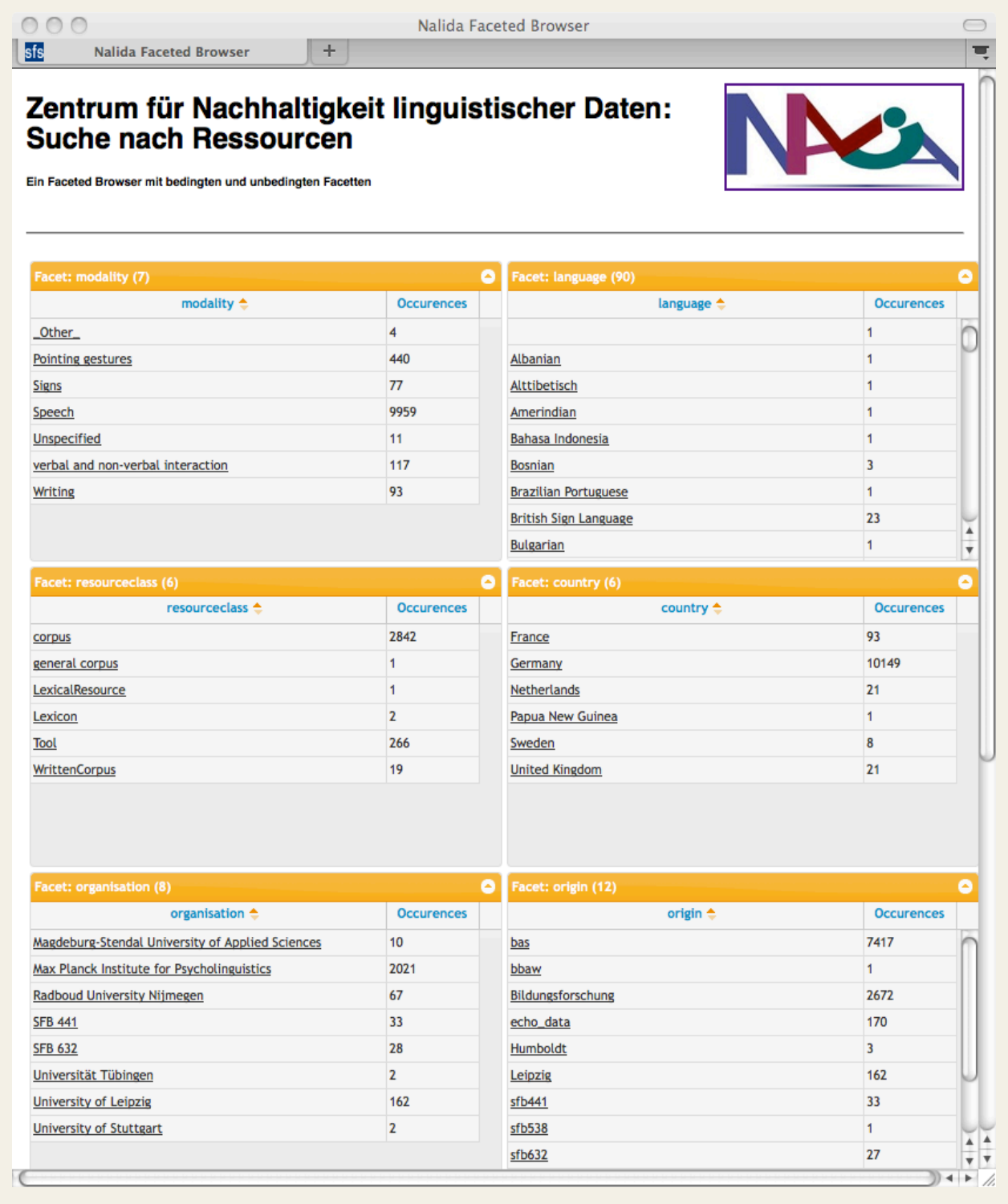
> *"Research infrastructures contribute significantly to gaining knowledge in the humanities, when it comes to problems in society or accessing cultural heritage."*
>
> *„Forschungsinfrastrukturen leisten in den Geistes- und Sozialwissenschaften einen wichtigen Beitrag zum Erkenntnisgewinn über gesellschaftliche Problemlagen und zur Erschließung unseres kulturellen Erbes."*
>
> (From: *Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften stärken,* Press release The German Council of Science and Humanities (Wissenschaftsrat), No. 3, Berlin, 31.01.2011)



The essential elements of the sustainability platform are digital objects. A digital object is a collection that contains a resource as well as its metadata description or any additional information belonging to it. This information may include scientific research articles that either describe the resource or use it as a basis, internal (project) reports, notes, etc. For authorisation restrictions on user groups, the platform supports access limitations.

Digital objects can be identified and addressed by unique persistent identifiers (PIDs). Libraries, such as the German National Library (DNB), have developed ranges and formats for these identifiers. Other examples are DOIs.


Digital Object
Technical Metadata
PID
Metadata
Primary Data Format 1
Primary Data Format 2
Additional Descriptions
...
Automatically created by archive systems
Persistent identifier (PID), Uniform Resource Name (URN) of the DNB or DOI
Descriptive metadata created by the researcher
Primary data created by the researcher
Descriptions, documents, publications, etc.



Zentrum für Nachhaltigkeit linguistischer Daten: Suche nach Ressourcen

Ein Faceted Browser mit bedingten und unbedingten Facetten

Facet: modality (7)

modality	Occurences
Other	4
Pointing gestures	440
Signs	77
Speech	9959
Unspecified	11
verbal and non-verbal interaction	117
Writing	93

Facet: language (90)

language	Occurences
	1
Albanian	1
Alttibetisch	1
Amerindian	1
Bahasa Indonesia	1
Bosnian	3
Brazilian Portuguese	1
British Sign Language	23
Bulgarian	1

Facet: resourceclass (6)

resourceclass	Occurences
corpus	2842
general corpus	1
LexicalResource	1
Lexicon	2
Tool	266
WrittenCorpus	19

Facet: country (6)

country	Occurences
France	93
Germany	10149
Netherlands	21
Papua New Guinea	1
Sweden	8
United Kingdom	21

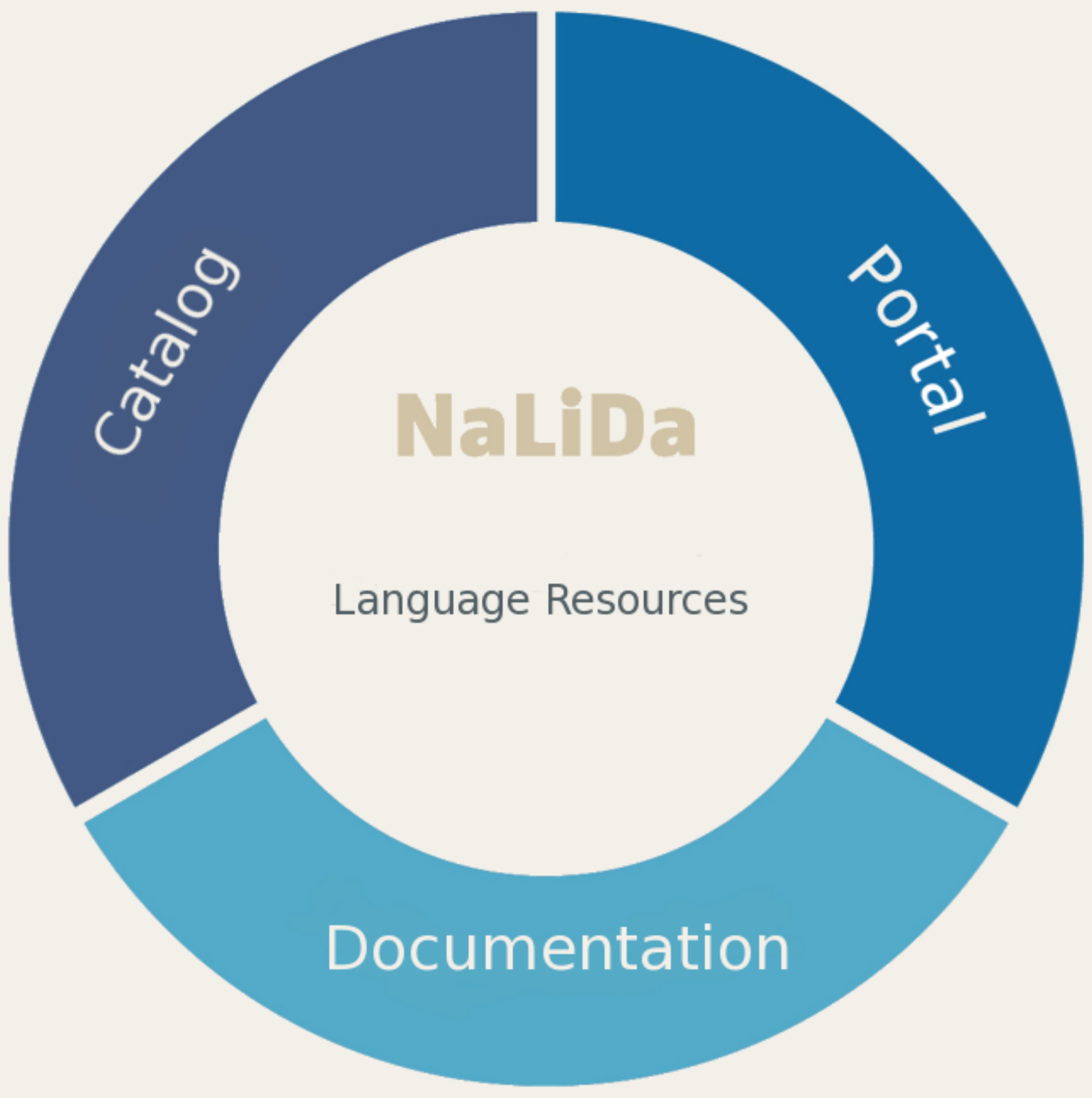
Facet: organisation (8)

organisation	Occurences
Magdeburg-Stendal University of Applied Sciences	10
Max Planck Institute for Psycholinguistics	2021
Radboud University Nijmegen	67
SFB 441	33
SFB 632	28
Universität Tübingen	2
University of Leipzig	162
University of Stuttgart	2

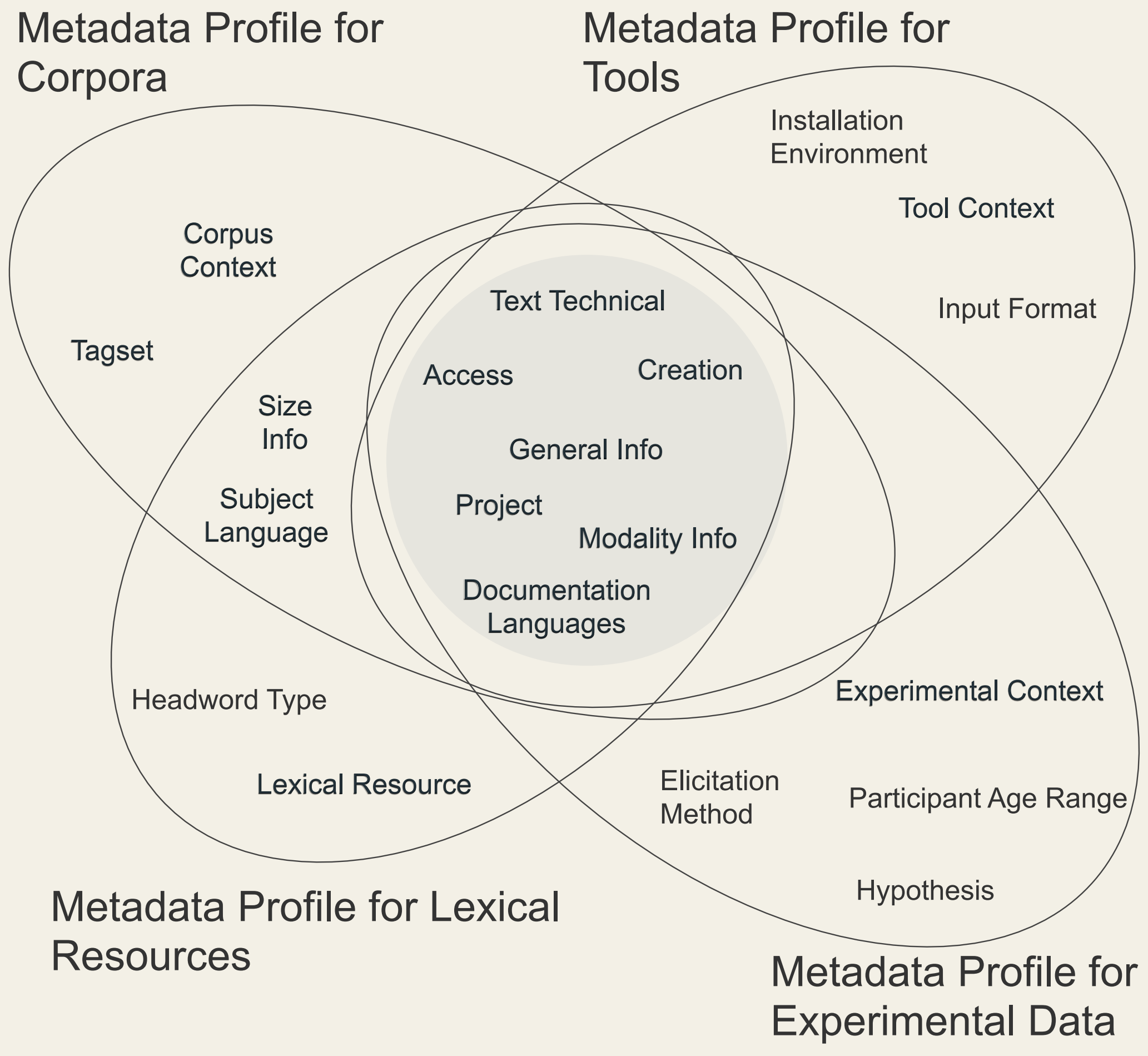
Facet: origin (12)

origin	Occurences
bas	7417
bbaw	1
Bildungsforschung	2672
echo_data	170
Humboldt	3
Leipzig	162
sfb441	33
sfb538	1
sfb632	27

The classification of resources is reflected by the search options that are based on the structures of the resource descriptions, i.e. the metadata. By this, a search for resources is restricted by characteristics such as modality, language or genre, reducing the size of the search space. Depending on selections, new structures may also appear. Here for example, *annotation schemes* were introduced based on the selection of the resource type *corpora,* a structure not relevant for many resource types.

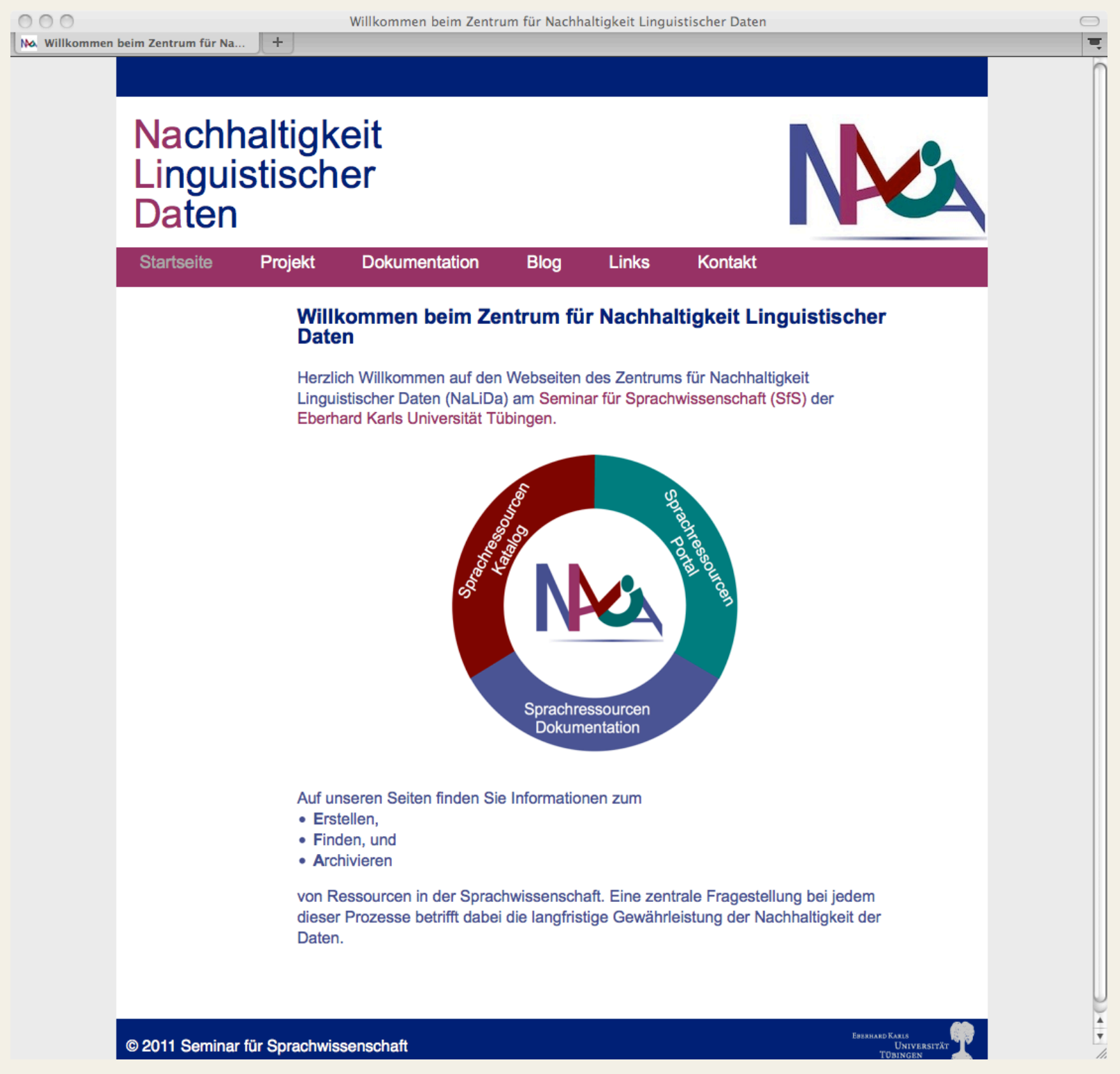


> *"Primary data as the basis for publications are to be kept on persistent and secured media within the institution where they were created for ten years."*
>
> *„Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden."*
>
> (From: Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten (Recomendations for safekeeping and providing digital primary research data), DFG: Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss für Informationsmanagement, German Research Foundation p. 2, January 2009)



The long-term archiving does not only include the technical storage, i.e. the storing of primary data in so-called data streams, but also the descriptions of the resources, i.e. the metadata. Within the NaLiDa project, CMDI (Component Meta Data Infrastructure) is used as the underlying metadata schema. For each resource class there is a metadata profile available for describing a resource. The smallest building blocks, i.e. the data categories (DatCats), are collected in components that can be reused in different profiles independent of the resource class. Often-used components build a core which is comparable to Dublin Core, but includes more data categories relevant for linguistic resources.

For the purpose of archiving and searching, metadata are available in structured XML formats. These XML formats are also dependent on the class of a resource. A few levels of description are kept as general as possible and can thus be applied to a number of linguistic resources. Nevertheless, these general descriptions also surpass traditional bibliographies, as bibliographic metadata such as Dublin Core are of limited use for linguistic resources.


Willkommen beim Zentrum für Nachhaltigkeit Linguistischer Daten
Nachhaltigkeit Linguistischer Daten
Startseite
Projekt
Dokumentation
Blog
Links
Kontakt
Sprachressourcen Katalog
Sprachressourcen Portal
Sprachressourcen Dokumentation

Willkommen beim Zentrum für Nachhaltigkeit Linguistischer Daten

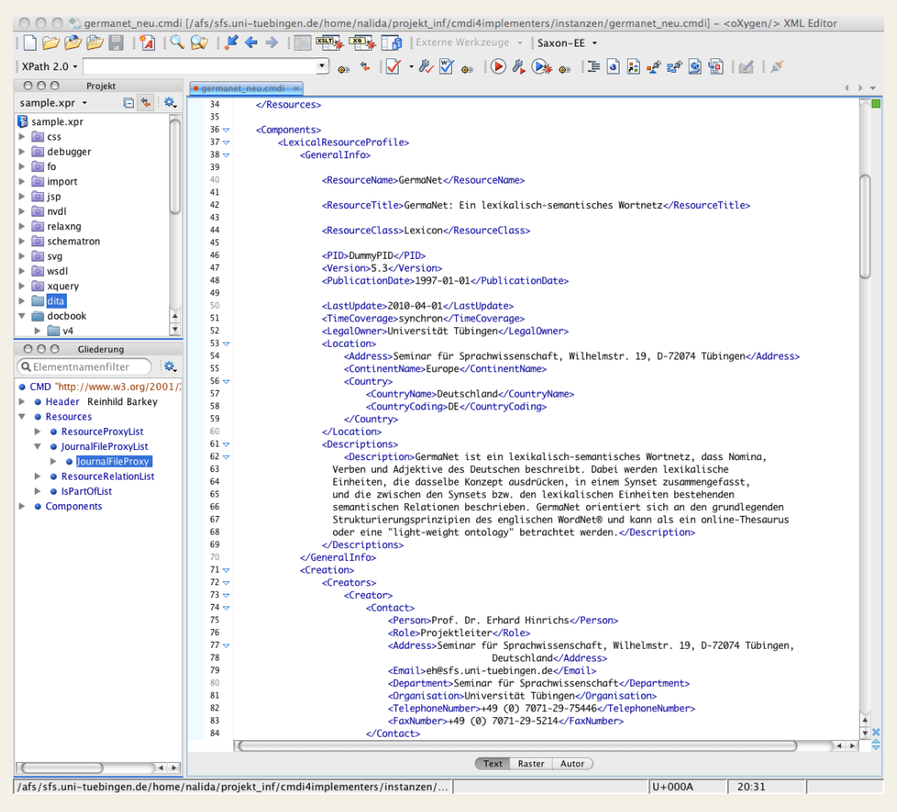
Herzlich Willkommen auf den Webseiten des Zentrums für Nachhaltigkeit Linguistischer Daten (NaLiDa) am Seminar für Sprachwissenschaft (SfS) der Eberhard Karls Universität Tübingen.

Auf unseren Seiten finden Sie Informationen zum
- Erstellen,
- Finden, und
- Archivieren

von Ressourcen in der Sprachwissenschaft. Eine zentrale Fragestellung bei jedem dieser Prozesse betrifft dabei die langfristige Gewährleistung der Nachhaltigkeit der Daten.

© 2011 Seminar für Sprachwissenschaft

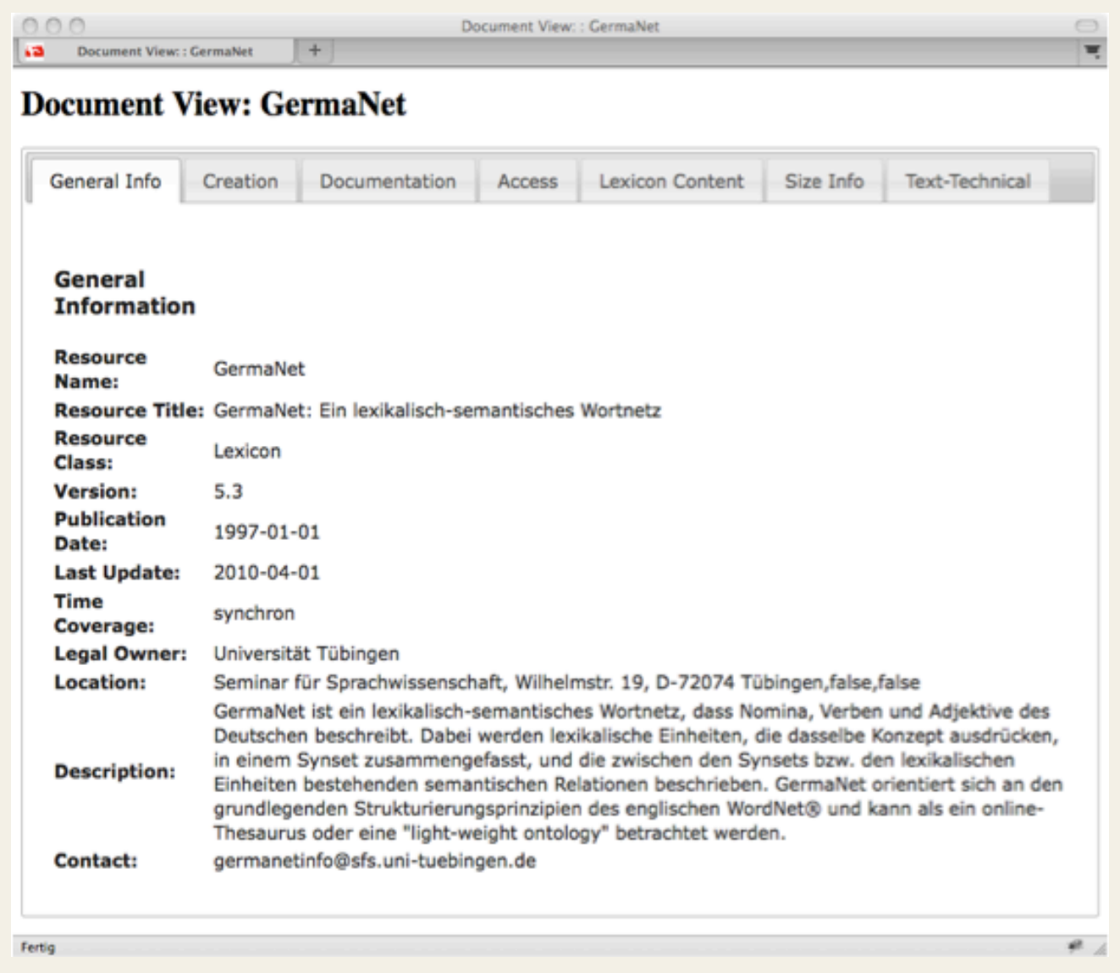
http://www.sfs.uni-tuebingen.de/nalida


germanet_neu.cmdi [/afs/sfs.uni-tuebingen.de/home/nalida/projekt_inf/cmdi4implementers/instanzen/germanet_neu.cmdi] - <oXygen/> XML Editor
Projekt
sample.xpr
Gliederung
Text
Raster
Autor

```
</Resources>

<Components>
    <LexicalResourceProfile>
        <GeneralInfo>

            <ResourceName>GermaNet</ResourceName>

            <ResourceTitle>GermaNet: Ein lexikalisch-semantisches Wortnetz</ResourceTitle>

            <ResourceClass>Lexicon</ResourceClass>

            <PID>DummyPID</PID>
            <Version>5.3</Version>
            <PublicationDate>1997-01-01</PublicationDate>

            <LastUpdate>2010-04-01</LastUpdate>
            <TimeCoverage>synchron</TimeCoverage>
            <LegalOwner>Universität Tübingen</LegalOwner>
            <Location>
                <Address>Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen</Address>
                <ContinentName>Europe</ContinentName>
                <Country>
                    <CountryName>Deutschland</CountryName>
                    <CountryCoding>DE</CountryCoding>
                </Country>
            </Location>
            <Descriptions>
                <Description>GermaNet ist ein lexikalisch-semantisches Wortnetz, dass 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>
            </Descriptions>
        </GeneralInfo>
        <Creation>
            <Creators>
                <Creator>
                    <Contact>
                        <Person>Prof. Dr. Erhard Hinrichs</Person>
                        <Role>Projektleiter</Role>
                        <Address>Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen,
                                   Deutschland</Address>
                        <Email>eh@sfs.uni-tuebingen.de</Email>
                        <Department>Seminar für Sprachwissenschaft</Department>
                        <Organisation>Universität Tübingen</Organisation>
                        <TelephoneNumber>+49 (0) 7071-29-75446</TelephoneNumber>
                        <FaxNumber>+49 (0) 7071-29-5214</FaxNumber>
                    </Contact>
```

Technical views to metadata are only relevant for computer programs (and programmers). The needs of linguists cannot be satisfied with them. The NaLiDa project offers a web portal for language resources with linguists as its target user group. Users, who are interested in language resources or their sustainable provision, can get an overview over existing resources represented by a language catalog. This catalog is based on the structures of the metadata and bundles different perspectives and aspects of the descriptions. One way of a user-friendly visualization offers the use of tabs.

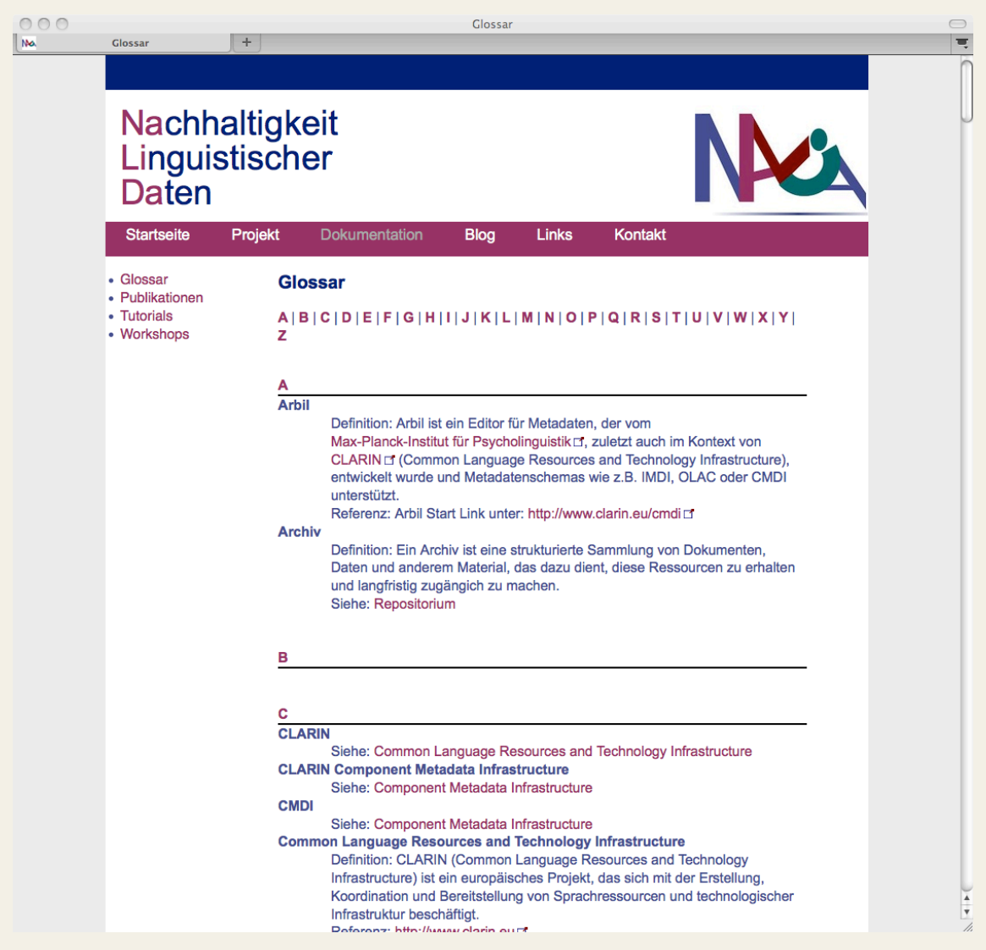
> *"The German Council of Science and Humanities sees the extensive public collection and provision of primary research data also as an appropriate means for quality assurance in academic practice[...]"*
>
> *„Der Wissenschaftsrat begreift die umfassende öffentliche Sammlung und Bereitstellung von Forschungsprimärdaten auch als ein probates Mittel der Qualitätssicherung in der wissenschaftlichen Praxis [...]."*
>
> (From: Empfehlung zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften (Recommendations for research infrastructures in the humanities), The German Council of Science and Humanities (Wissenschaftsrat), Berlin, 28.01.2011, p.59)



Document View: GermaNet

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
Description:	GermaNet ist ein lexikalisch-semantisches Wortnetz, dass 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.
Contact:	germanetinfo@sfs.uni-tuebingen.de


Glossar
Nachhaltigkeit Linguistischer Daten
Startseite
Projekt
Dokumentation
Blog
Links
Kontakt
Glossar
Publikationen
Tutorials
Workshops
A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Glossar

A

Arbil
Definition: Arbil ist ein Editor für Metadaten, der vom Max-Planck-Institut für Psycholinguistik, zuletzt auch im Kontext von CLARIN (Common Language Resources and Technology Infrastructure), entwickelt wurde und Metadatenschemas wie z.B. IMDI, OLAC oder CMDI unterstützt.
Referenz: Arbil Start Link unter: http://www.clarin.eu/cmdi

Archiv
Definition: Ein Archiv ist eine strukturierte Sammlung von Dokumenten, Daten und anderem Material, das dazu dient, diese Ressourcen zu erhalten und langfristig zugängich zu machen.
Siehe: Repositorium

B

C

CLARIN
Siehe: Common Language Resources and Technology Infrastructure
CLARIN Component Metadata Infrastructure
Siehe: Component Metadata Infrastructure
CMDI
Siehe: Component Metadata Infrastructure
Common Language Resources and Technology Infrastructure
Definition: CLARIN (Common Language Resources and Technology Infrastructure) ist ein europäisches Projekt, das sich mit der Erstellung, Koordination und Bereitstellung von Sprachressourcen und technologischer Infrastruktur beschäftigt.

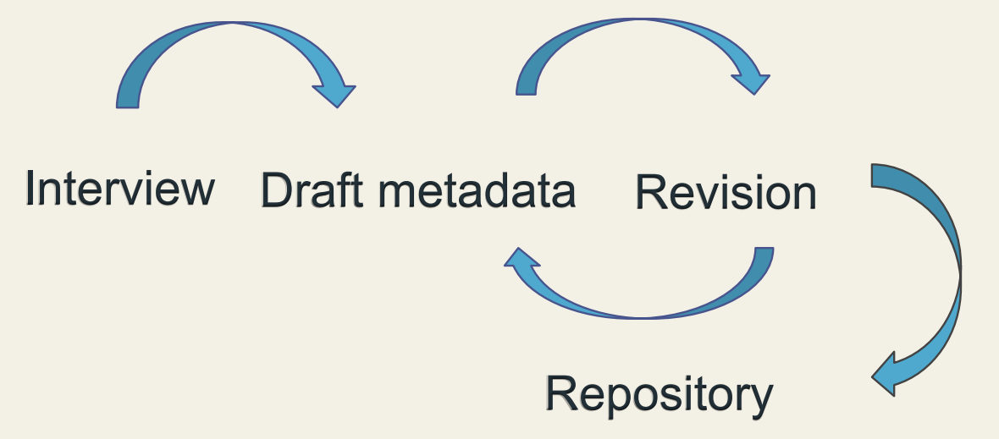
The search functions and metadata views within the web portal provide access to primary research data such as corpora, lexical resources, experimental data, etc. Privacy and property restrictions may require limited access to the public, hence authorization is part of the archive system. Reasons for restrictions can include: pending publication approval, ethical reasons, etc. Nonetheless, such resources can still be found on the basis of metadata so that it is at least possible to approach the resource creators and learn about the conditions for accessing the resource.

The provision of linguistic resources and the search over these, is a major part of the web portal, also serving the purpose of exchanging information on the creation and archiving of language resources. In order to support the communication between different research branches, there is - apart from other documentations and tutorials - a glossary which makes terms accessible that are relevant for linguistic resources. This is also part of the sustainability aspect.

Within the NaLiDa project, the creation of metadata generally uses an iterative process. This mainly serves the purpose of training resource creators enabling them to create metadata on their own, but also of including new resource types. On the basis of first interviews and resource descriptions, a draft of a metadata schema is created together with a metadata instance. If needed, further components and profiles are also developed.

This metadata instance can then be corrected by the resource creator without requiring any further technical knowledge. Adjustments can be made at any time.

If the metadata instance is complete and correct, it can be stored in the sustainability platform together with the primary data. After this process, it is still possible to edit the instance.



University of Tübingen· Faculty of Humanities· Department of Linguistics
Wilhelmstraße 19 · 72074 Tübingen · Germany
www.sfs.uni-tuebingen.de/nalida
nalida@sfs.uni-tuebingen.de