Repository organization and depositing

Funders‘ Requirements: The DFG Guidelines

Informationen für Geistes- und Sozialwissenschaftler/innen Förderkriterien

- Use XML format such as DTABf: The data format for historic data developed by the CLARIN centre at the BBAW in Berlin
- Independent of specific software: CLARIN uses open interfaces
- Version of the publication: CLARIN supports
- Deposit in institutional or subject specific repository: The CLARIN centres have repositories
- Using persistent identifiers: CLARIN supports ISO 24619:2011
- Certified repositories: CLARIN repositories: DSA + centre assessment
- Free use for research and teaching: CLARIN-Supports authentication for the academic community

ten[4]

sierten Markup codiert.
jeweils aktuellen Version

sentlichen
p werden erwartet. Als
extGrid6 oder

erden die projektspezifisch
eader dokumentiert.
erden, sollte eine eeignete
e Unabhängigkeit von

entsprechen, die auch der
n sind zu begründen, etwa

fachlichen Repositorium

rtifikat8 oder ein

Die elektronische Ablage erfolgt ausdrücklich unter einer Lizenz, die eine freie Nachnutzung zu wissenschaftlichen Zwecken erlaubt. 9 Dem Antrag sollte eine Bestätigung der Institution beiliegen, welche die Langzeitarchivierung der Textdaten gewährt.

Repository of Trust: Data Seal of Approval

Implementation of the Data Seal of Approval

The Data Seal of Approval board hereby confirms that the Trusted Digital repository Tübingen CLARIN-D Repository complies with the guidelines version 2 of 2013 set by the Data Seal of Approval Board.
The afore-mentioned repository has therefore acquired the Data Seal of Approval of 2013 on May 19, 2015.

The Trusted Digital repository is allowed to place an image of the Data Seal of Approval logo corresponding to the guidelines version date on their website. This image must link to this file which is hosted on the Data Seal of Approval website.

Yours sincerely,

The Data Seal of Approval Board

DSA requirements*

- Organizational infrastructure
- Digital object management
- Technology

*With material from:
Peter Doorn, Director DANS
Ingrid Dillo, Deputy Director DANS
„FAIR Data in Trustworthy Data Repositories: Everybody wants to play FAIR, but how do we put the principles into practice?“
Licenced under CC-BY
See https://www.slideshare.net/EUDAT/fair-data-in-trustworthy-data-repositories-webinar-1213-december-2016

1. Organisational infrastructure

- R1. The repository has an explicit mission to provide access to and preserve data in its domain.
- R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.
- R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.
- R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
- R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.
- R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).

2. Digital object management

- R7. The repository guarantees the integrity and authenticity of the data.
- R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
- R9. The repository applies documented processes and procedures in managing archival storage of the data.
- R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

2. Digital object management

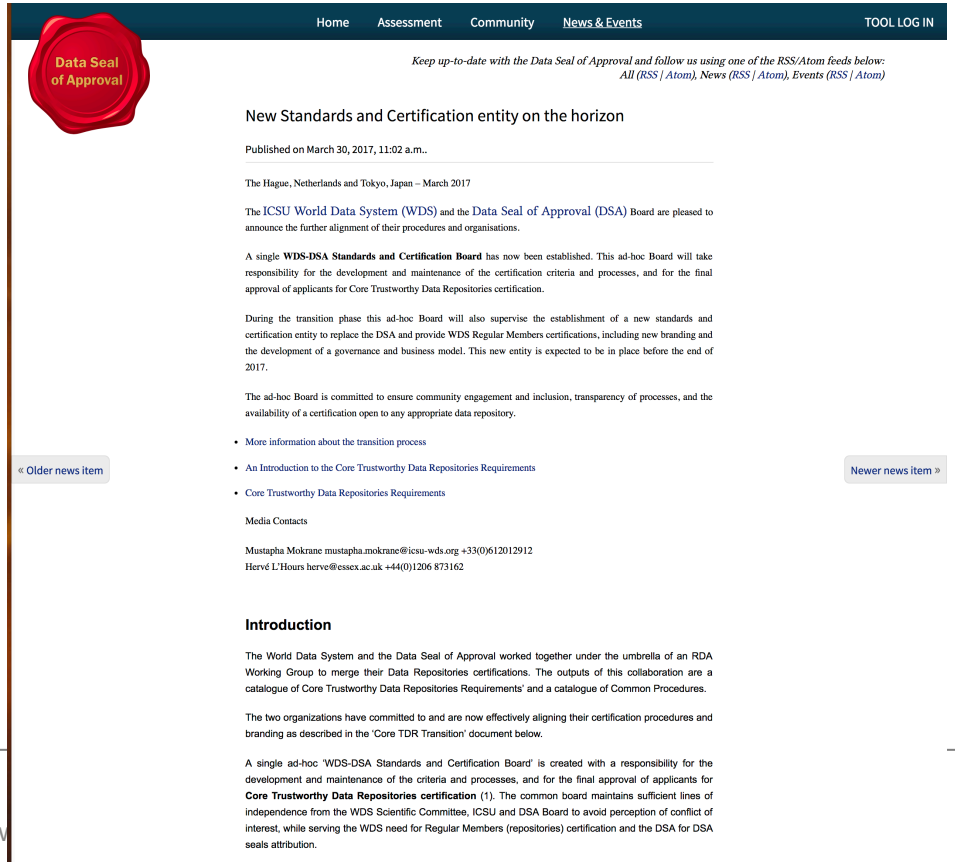
- R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.
- R12. Archiving takes place according to defined workflows from ingest to dissemination.
- R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.
- R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

3. Technical infrastructure

- R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.
- R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Get it now! -ehem

- https://www.datasealofapproval.org/en/news-and-events/news/2017/3/30/new-standards-and-certification-entity-horizon/



Home Assessment Community News & Events TOOL LOG IN

Data Seal of Approval

Keep up-to-date with the Data Seal of Approval and follow us using one of the RSS/Atom feeds below:
All (RSS | Atom), News (RSS | Atom), Events (RSS | Atom)

New Standards and Certification entity on the horizon

Published on March 30, 2017, 11:02 a.m..

The Hague, Netherlands and Tokyo, Japan – March 2017

The ICSU World Data System (WDS) and the Data Seal of Approval (DSA) Board are pleased to announce the further alignment of their procedures and organisations.

A single **WDS-DSA Standards and Certification Board** has now been established. This ad-hoc Board will take responsibility for the development and maintenance of the certification criteria and processes, and for the final approval of applicants for Core Trustworthy Data Repositories certification.

During the transition phase this ad-hoc Board will also supervise the establishment of a new standards and certification entity to replace the DSA and provide WDS Regular Members certifications, including new branding and the development of a governance and business model. This new entity is expected to be in place before the end of 2017.

The ad-hoc Board is committed to ensure community engagement and inclusion, transparency of processes, and the availability of a certification open to any appropriate data repository.

- More information about the transition process
- An Introduction to the Core Trustworthy Data Repositories Requirements
- Core Trustworthy Data Repositories Requirements

« Older news item Newer news item »

Media Contacts

Mustapha Mokrane mustapha.mokrane@icsu-wds.org +33(0)612012912
Hervé L'Hours herve@essex.ac.uk +44(0)1206 873162

Introduction

The World Data System and the Data Seal of Approval worked together under the umbrella of an RDA Working Group to merge their Data Repositories certifications. The outputs of this collaboration are a catalogue of Core Trustworthy Data Repositories Requirements' and a catalogue of Common Procedures.

The two organizations have committed to and are now effectively aligning their certification procedures and branding as described in the 'Core TDR Transition' document below.

A single ad-hoc 'WDS-DSA Standards and Certification Board' is created with a responsibility for the development and maintenance of the criteria and processes, and for the final approval of applicants for **Core Trustworthy Data Repositories certification** (1). The common board maintains sufficient lines of independence from the WDS Scientific Committee, ICSU and DSA Board to avoid perception of conflict of interest, while serving the WDS need for Regular Members (repositories) certification and the DSA for DSA seals attribution.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Foundation: FAIRness

- **Findable:** easy to find for both humans and computers, with metadata that facilitate searching for specific datasets,
- **Accessible:** stored for long term so that they can easily be accessed and/or downloaded with well-defined license and access conditions (open access *when possible*), whether at the level of metadata, or at the level of the actual data,
- **Interoperable:** ready to be combined with other datasets by humans or computers,
- **Reusable:** ready to be used for future research and to be further processed using computational methods.

See: https://www.dtls.nl/fair-data/fair-data/
https://www.nature.com/articles/sdata201618

FAIR Guiding Principles

1. To be **Findable** any Data Object should be uniquely and persistently identifiable
    1. The same Data Object should be re-findable at any point in time, thus Data Objects should be **persistent**, with emphasis on their metadata,
    2. A Data Object should minimally contain basic machine actionable metadata that allows it to be distinguished from other Data Objects
    3. Identifiers for any concept used in Data Objects should therefore be **Unique** and **Persistent**

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

See: https://www.force11.org/node/6062

FAIR Guiding Principles

2. Data is **Accessible** in that it can be always obtained by machines and humans
    1. Upon appropriate authorization
    2. Through a well-defined protocol
    3. Thus, machines and humans alike will be able to judge the actual accessibilty of each Data Object.

See: https://www.force11.org/node/6062

FAIR Guiding Principles

3. Data Objects can be **Interoperable** only if:
    1. (Meta) data is machine-actionable
    2. (Meta) data formats utilize shared vocabularies and/or ontologies
    3. (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessible

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

See: https://www.force11.org/node/6062

FAIR Guiding Principles

4. For Data Objects to be **Re-usable** additional criteria are:
   1. Data Objects should be compliant with **principles 1-3**
   2. (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources
   3. Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

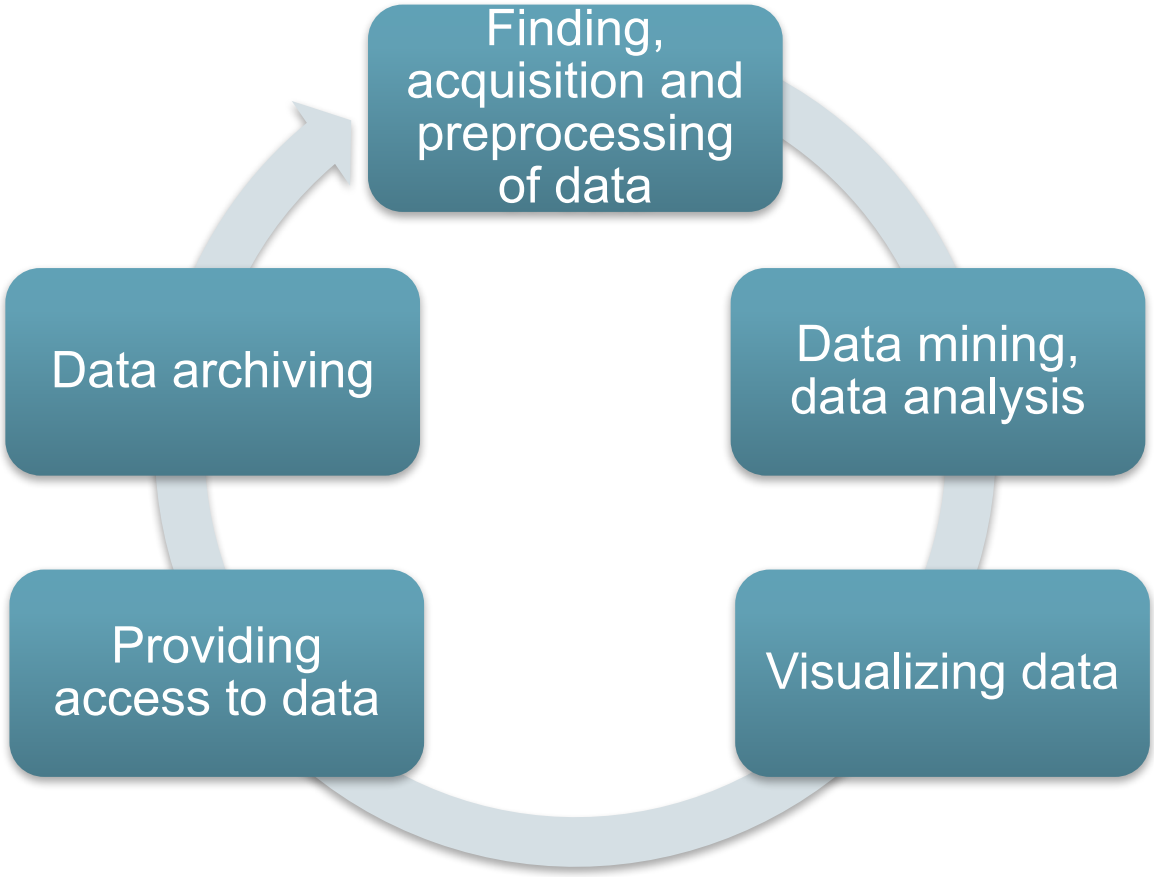
See: https://www.force11.org/node/6062

Aspects of FAIRness to data

- FAIR data management: posing requirements for new data creation
- FAIR data assessment: establishing the profile of existing data
- FAIR data technologies: transformation tools to make data FAIR

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

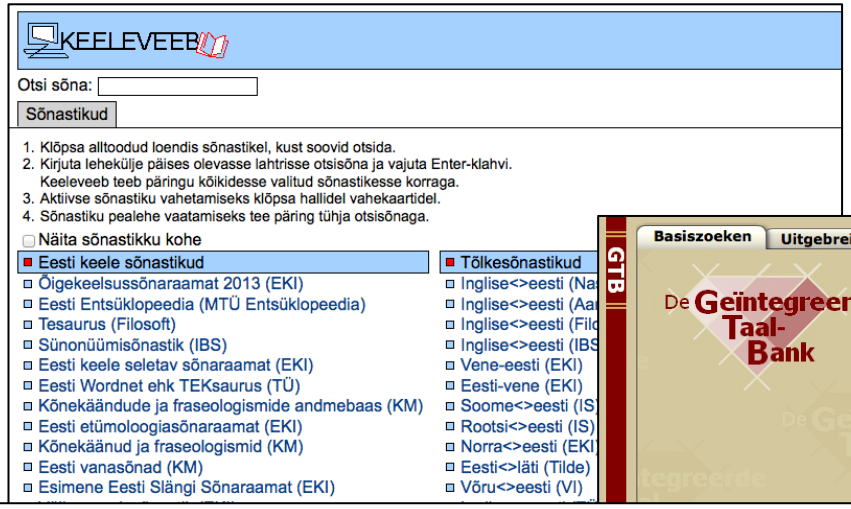
Looking at FAIR in terms of the phases of e-humanities projects

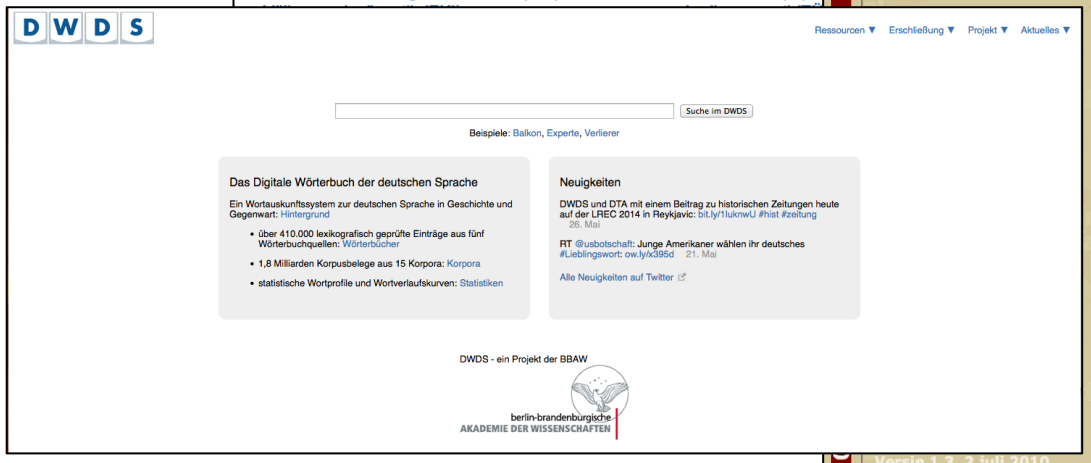


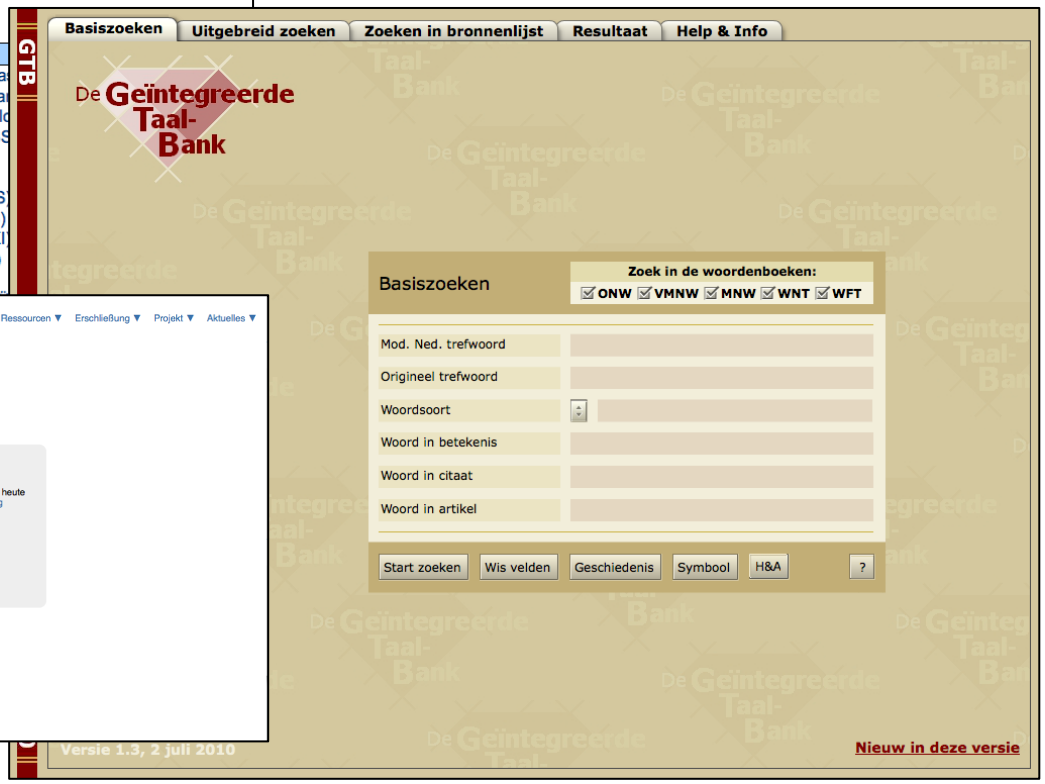
Corpora

Ústav Českého národního korpusu
Filozofická fakulta Univerzity Karlovy v Praze
hledat na stránkách
OK
Konference KL2014
Krátké zprávy
Co je korpus?
Kontakty
Dostupné korpusy
Projekt InterCorp
Naše publikace
Dohody a registrace
Hledat v ČNK
Pracovní kolektiv
Poskytovatelé textů
Granty, sponzoři
Jak citovat korpus
Knihovna
Výuka
Spolupráce
Jiné korpusy
Ke stažení
Manuál a instalace
Portál ČNK
Od začátku roku 2014 funguje nový portál pro práci s korpusy na adrese www.korpus.cz. V rámci této stránky najdete nejen všechny naše nástroje pro práci s korpusovými daty, zejména pak korpusový vyhledávač KonText, nástroj pro vyhodnocování variant SyD a další, ale také manuál ve formě Wiki a poradnu pro uživatele.
Co je korpus?
Korpus je soubor počítačově uložených textů (v případě mluveného jazyka - přepisů záznamu mluvy), který primárně slouží k jazykovému výzkumu. K práci s korpusy slouží speciální vyhledávací program. S jeho pomocí je možné vyhledávat slova a slovní spojení v kontextu a zjistit jejich frekvenci v korpuse i původní textový zdroj. Umožňuje i další zpracování nalezeného (např. abecední třídění apod.). U některých korpusů lze vyhledávat i podle slovních druhů.
Český národní korpus (ČNK) je akademický projekt zaměřený na budování rozsáhlého počítačového korpusu především psané češtiny. Pracuje na něm Ústav Českého národního korpusu na Filozofické fakultě Univerzity Karlovy v Praze (ÚČNK). Od svého založení roku 1994 má ÚČNK na starosti budování
Aktuality
Korpusy Aranea
V květnu 2014 byla v ČNK zveřejněna rodina nereferenčních srovnatelných webových korpusů Aranea sestavenýc Vladimírem Benkem. V současné dob pokrývá celkem 7 jazyků (de, en, fr, nl, ru, sk), další informace zde.
Proverbs: Their Lexical and Semant Features
Nedávno vyšla v USA anglicky psaná Františka Čermáka o paremiologii (přísloviectví) pojednávající o různých, hlavně však lexikálních aspektech více jazyků včetně jejich pragmatiky, paremiologických minim (češtiny, angli a němčiny). Je založená na rozsáhlé analýze dat, většinou korpusových. Kn nás nedostupnou, lze v omezeném roz pro vážné zájemce získat přímo od au
Více zde.
AUSTRIAN ACADEMY CORPUS
Text-Technology
Applications
Lab
Institution
Austrian Academy Corpus
Das AAC-Austrian Academy Corpus ist eine umfangreiche und komplex strukturierte Sammlung von digitalen Volltexten zur deutschen Sprache und Literatur im Untersuchungszeitraum 1848 bis 1989. Das AAC ist sowohl Basis für die Entwicklung innovativer texttechnologischer Anwendungen als auch für die Verfolgung philologischer und kulturwissenschaftlicher Forschungsinteressen. Das Programm erhebt den Anspruch, Schnittstellen zwischen kulturwissenschaftlichen Methoden und Prinzipien (u.a. philologische Genauigkeit, Editionstechniken, Textstrukturierung, Annotierung) und IKT-Anwendungen zu schaffen sowie Schnittstellen zwischen Grundlagenforschung und deren Umsetzungsmöglichkeiten.
Im AAC werden Texte digitalisiert, um sie philologisch zu bearbeiten, auszuwerten und zu interpretieren. Das Erkenntnisinteresse des AAC gilt den immer neu zu stellenden Fragen der Philologie: Wie funktionieren Wörter und Texte, woher kommen sie, wie sind sie gebaut, welche Inhalte vermitteln sie, warum werden sie geschrieben, warum werden sie gelesen, welchen Einfluss haben sie auf den Sprachgebrauch? Die EDV und texttechnologische Methoden machen es heute möglich, diese Fragen systematisch an äußerst große Mengen von Texten heranzutragen und sie in einem Team beschleunigt abzuarbeiten; durch diese geänderte Methodik in den Textwissenschaften gibt es neue Hypothesen und neue Antworten und das wiederum bringt Spannung, Kontroverse und neue Erkenntnisse in die Philologie.
INSTITUT FÜR DEUTSCHE SPRACHE
Mitglied der Leibniz-Gemeinschaft
Aktuelles
Organisationsstruktur
Forschung
Bibliothek
Service
Publikationen
Über uns
Direktion und zentrale Forschung
Grammatik
Lexik
Pragmatik
Öffentlichkeitsarbeit
Zentrale DV-Dienste
Verwaltung
Startseite : : Organisationsstruktur : : Direktion : : Korpuslinguistik Korpuslinguistik : : Projekte : : Korpusausbau
Suche
Direktion und zentrale Forschung
: : Forschungsinfrastrukturen
: : Korpuslinguistik
: : Projekte
: : Korpusausbau
: : Einsatzgebiete
: : Akquisition
: : IDS-Textmodell
: : Archiv
: : Ausgaben
: : Annotationen
: : Verfügbarkeit
: : Textgeber
: : Ehemalige Mitarbeiter
: : Kooperationen
: : Bibliographie (IDS)
: : Bibliographie
: : Analysemethodik
Ausbau und Pflege der Korpora geschriebener Gegenwartssprache
Kohlhammer
DAS DEUTSCHE REFERENZKORPUS – DEREKO
die weltweit größte Sammlung deutschsprachiger Korpora als empirische Basis für die linguistische Forschung
Die Korpora geschriebener Gegenwartssprache des IDS
bilden mit 6,1 Milliarden Wörtern (Stand 19.03.2013) die weltweit größte linguistisch motivierte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit.
enthalten belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Zahl von Zeitungstexten sowie eine breite Palette weiterer Textarten und werden kontinuierlich weiterentwickelt.
werden im Hinblick auf Umfang, Variabilität, Qualität und Aktualität akquiriert und erlauben in der Nutzungsphase über COSMAS II die Komposition virtueller Korpora, die repräsentativ oder auf spezielle Aufgabenstellungen zugeschnitten sind.
enthalten ausschließlich urheberrechtlich abgesichertes Material.
Kontakt:
<korpuslinguistik@ids-...>
Leitung:
Dr. Marc Kupietz
<kupietz@ids-...>
Wissenschaftliche Mitarbeiter:
Cyril Belica <belica@ids-...>
Dr. Harald Lüngen
<luengen@ids-...>
Rainer Perkuhn
<perkuhn@ids-...>
Kooperationen:
siehe hier
Ehemalige am Korpusaufbau beteiligte Mitarbeiter des IDS:
siehe hier

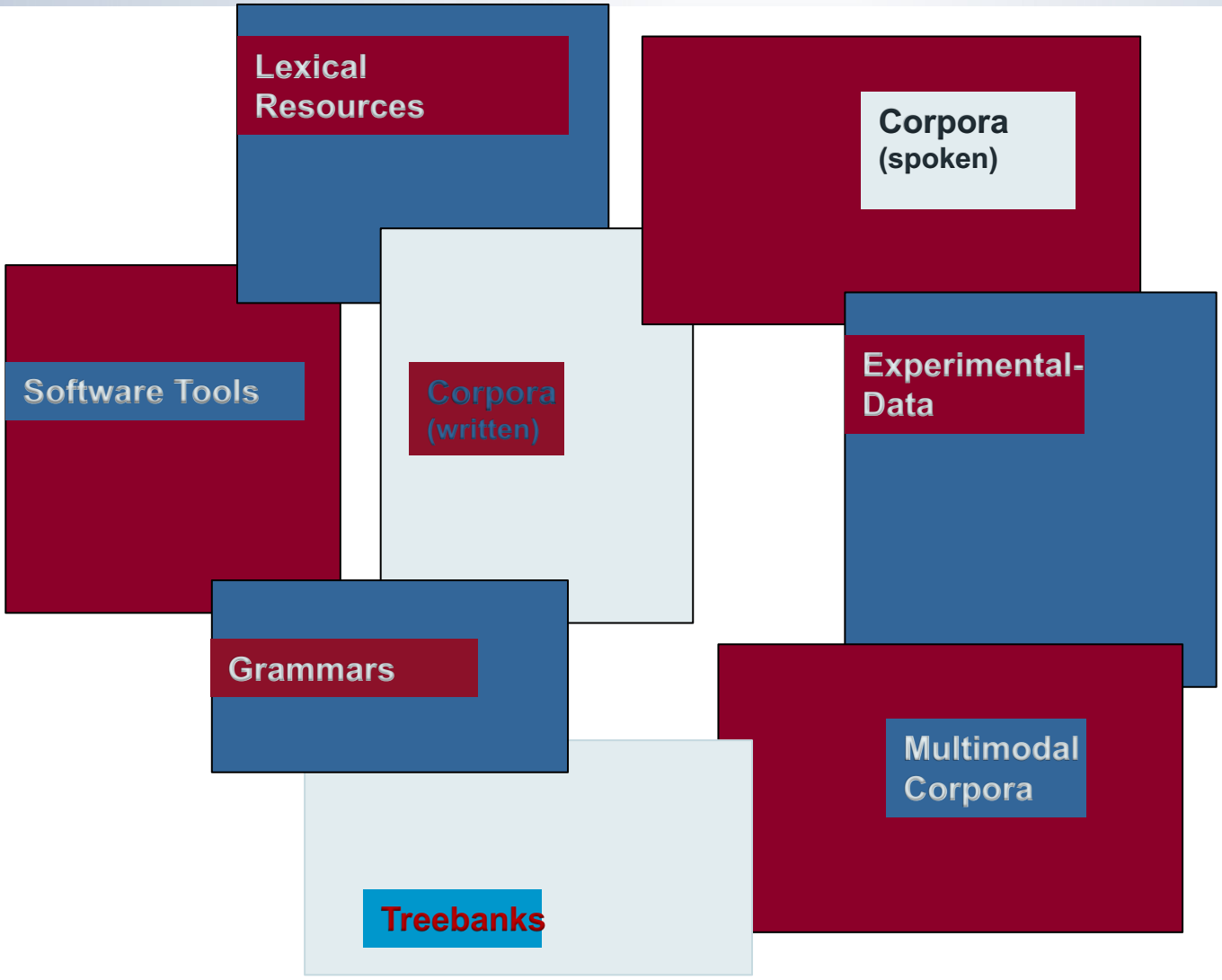
Lexical resources





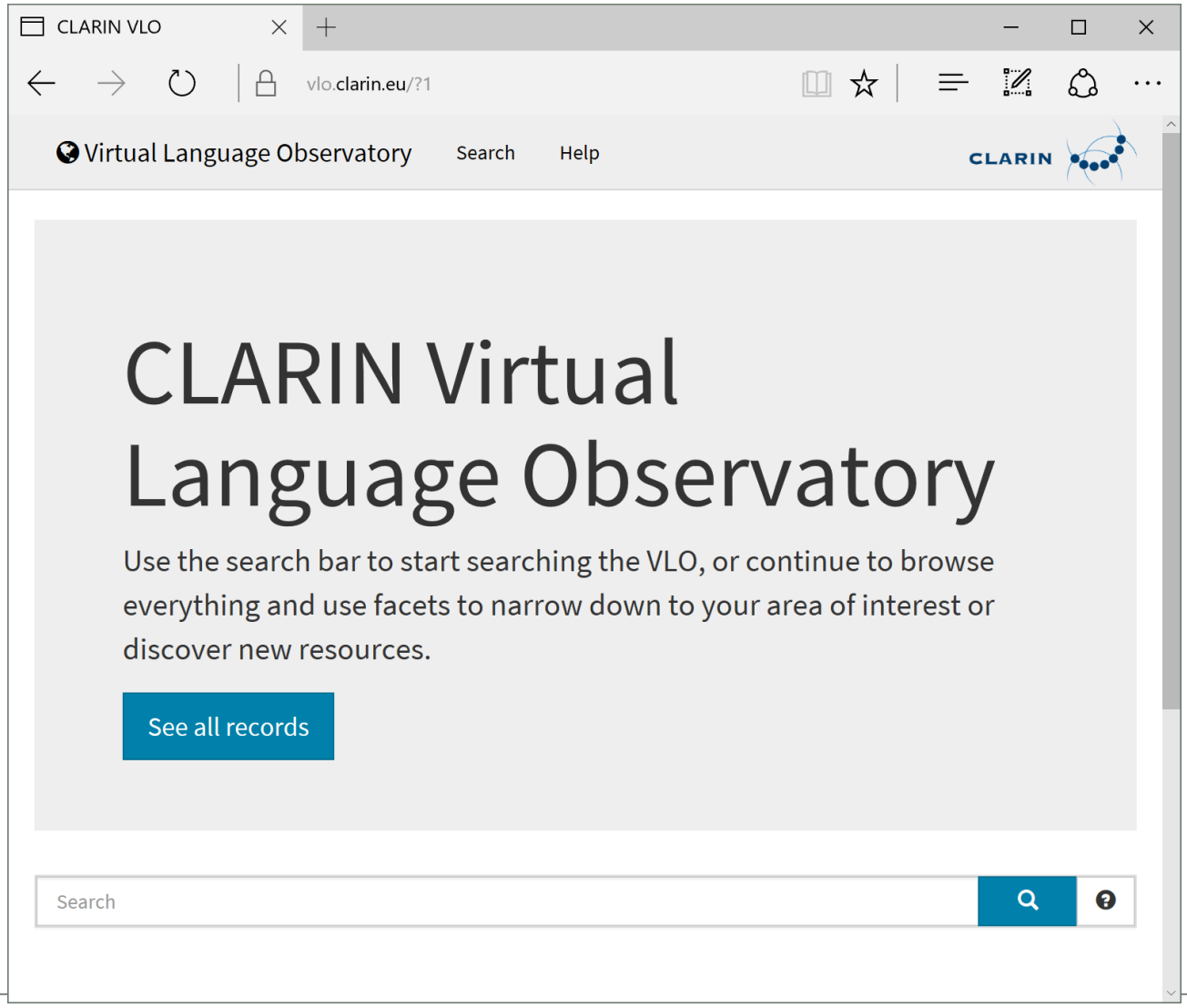


Variety of resources



Finding resources: Virtual Language Observatory (VLO)

- Searching for resources
- Based on descriptions of resources (metadata)
- Automatic processing of accessible data (OAI-PMH provider for CMDI)



VLO

CLARIN VLO

vlo.clarin.eu/;jsessionid=5CC33DB7A06082A43F4318A9A4CFA25A?0

Use the categories below to limit the search results to those matching the selected value (s).

- Language
- Collection
- Resource type
- Modality
- Format
- Keyword

Availability

Search options

<< < 1 2 3 4 5 6 7 8 9 > >>

WordNet

1

⊞ plWordNet is a lexico-semantic network which reflects the lexical system of the Polish language. There are at present ca. 144,000 nouns, verbs and adjectives in plWordNet, ca. 203,000 word senses and ca. 500,000 relations. It is already the second-largest wordnet in the world, and it keeps growing.

*

Finnish WordNet

⊞ The language resource is available in Kielipankki - the Language Bank of Finland at http://urn.fi/urn:nbn:fi:lb-2016042205; download: http://urn.fi/urn:nbn:fi:lb-2014052713. The Finnish WordNet is a lexical database for Finnish. It is a part of the FIN-CLARIN infrastructure project. FinnWordNet is lic…

Ⓟ

Czech WordNet

1

⊞ Multilingual Lexicons; The Czech WordNet was developed by the Centre of Natural Language Processing at the Faculty of Informatics, Masaryk University, Czech Republic. The Czech WordNet captures nouns, verbs, adjectives, and partly adverbs, and contains 28,201 word senses (synsets). Every synset encodes the equivalenc…

?

Bulgarian WordNet

1

⊞ Multilingual Lexicons; The Bulgarian WordNet is a network of lexical-semantic

?

Motivation

- Persistent archiving of resources
  - Data is stored and maintained in a consistent form
  - Within dedicated infrastructures
  - With no interference by other processes
- Addressing the resources from outside of the repository
  - For citability
  - Increasing visibility to the academic public
  - Using resolvable identifiers
- Describing the resources consistently
  - Searches over their descriptions
  - Metadata
- Sharing of resources
  - Providing download options for the general public
  - Access for authorized users

Extrinsic motivation for archiving: Funding

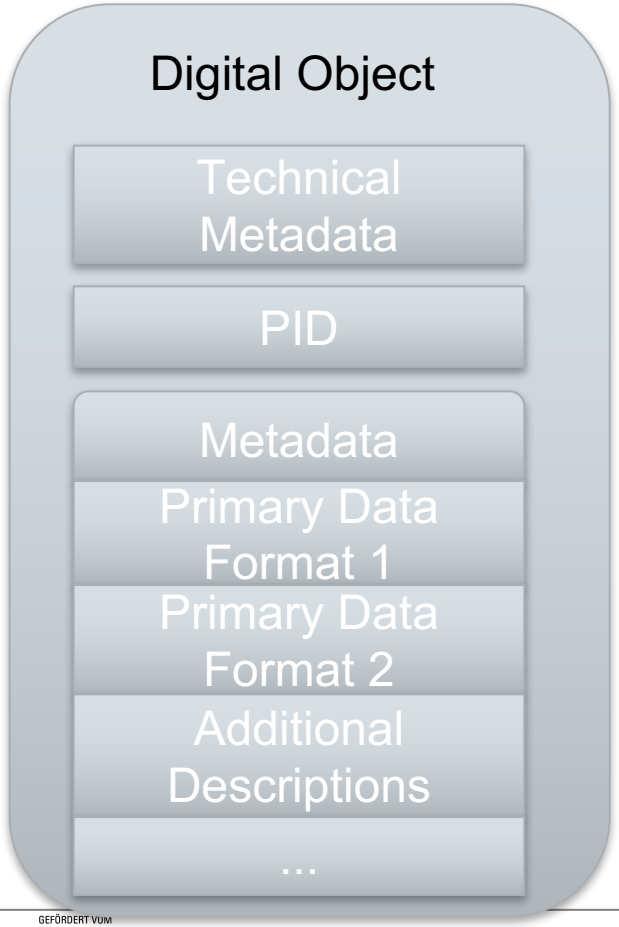
- HORIZON 2020 requires FAIR
- DFG (German Research Foundation):
  - min 10 years availability
  - Originating institution
- Wissenschaftsrat (Consulting council for the German Government):
  - In general publicly accessible
  - QA in research (anti plagiarism and fraud)
- Requests by other researchers

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Archiving: How? Three phases:

- Preparing a resource for archiving
- Inserting the resource into the archive system
- Archive acceptance: Sanity check and PID assignment

Technical Infrastructure: Fedora-Commons backend



← Automatically created by archive systems

Persistent identifier (PID), Uniform Resource Name (URN) of the DNB or DOI

← Descriptive metadata created by the researcher

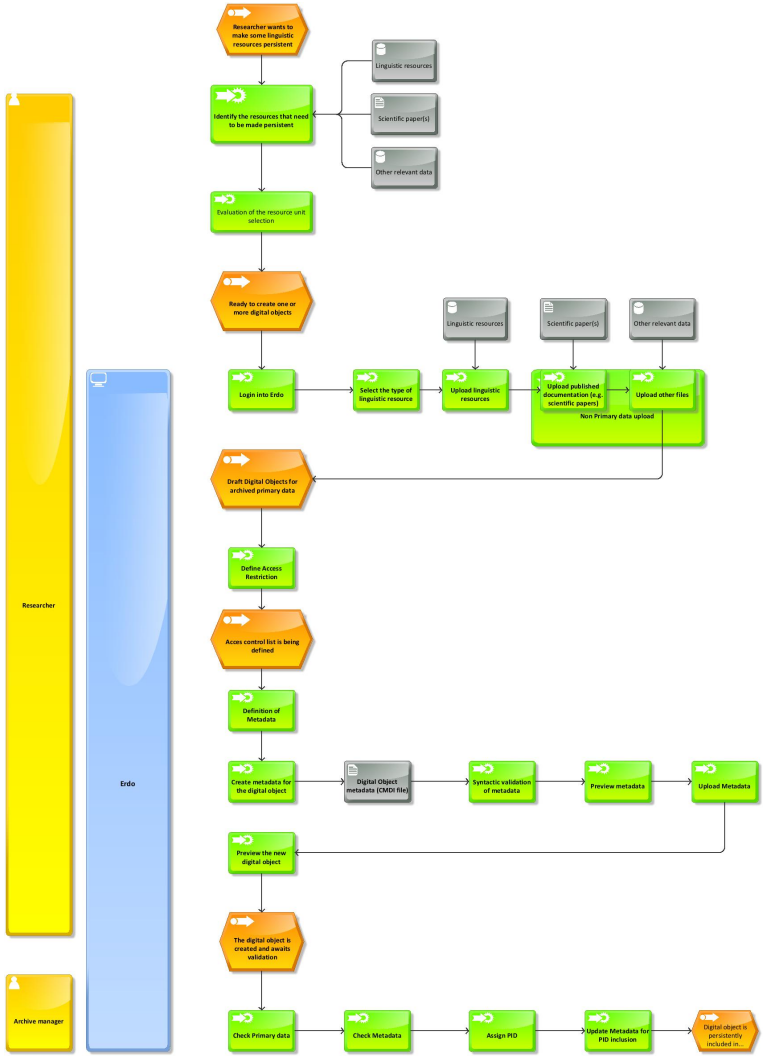
← Primary data created by the researcher

← Descriptions, documents, publications, etc.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Phase 1: Workflow

- Decide on the resource: What is one resource?
  - One experiment?
  - All parallel experiments in a study?
  - One lexicon?
  - One lexicon article?
  - One corpus?
  - One annotation layer?
- Rules:
  - One resource, one citable PID
  - A resource should "make sense" without other resources
  - See also ISO 24619:2011



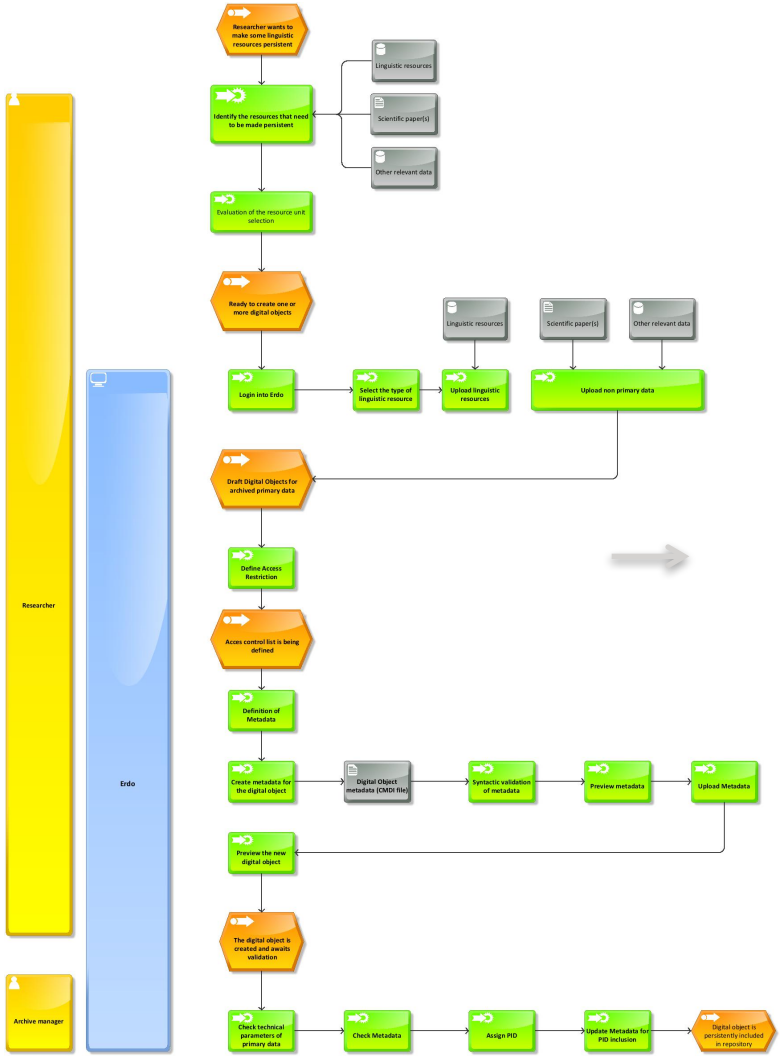
Granularity according to ISO 24619:2011

- Individual IDs for resources in these cases:
- If there is an existing identifier scheme for a type of resources, for instance, ISBN, this level of granularity should be retained[...]
- If the resource is associated with the complete content of a digital file.
- If the resource is autonomous and exists outside a larger context.
- If a resource should be citable apart from any containing resource.

- "Subject to the needs of resource creators with respect to the level of granularity they deem suitable to the specific resource environment."

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Phase 2: Entering the archive (data upload)



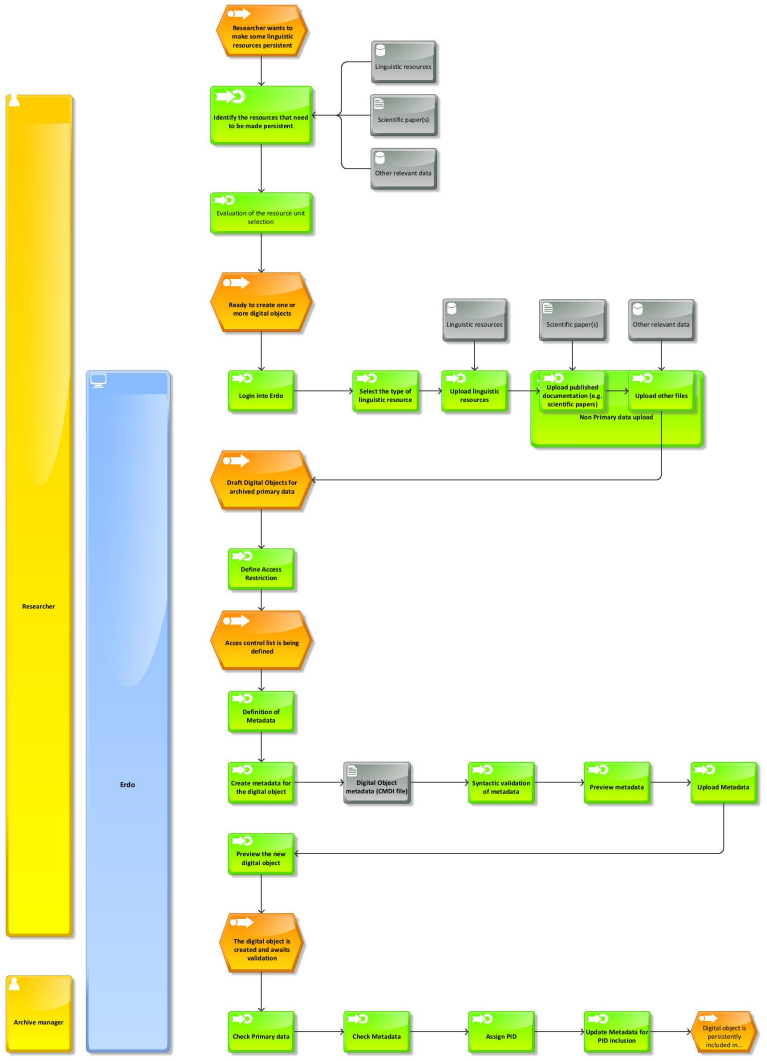
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Documentation beyond the primary data

- Some primary research data: not self explanatory
- Existing documentation:
  - Publications
  - Technical documents
  - README-files
- Access restrictions to documentation
  - Publishers
  - "unfinished"
  - Internal information

Phase 2: Entering the archive (access control)

- Project internal: all read
- Allow other individuals to read
- Allow public to read

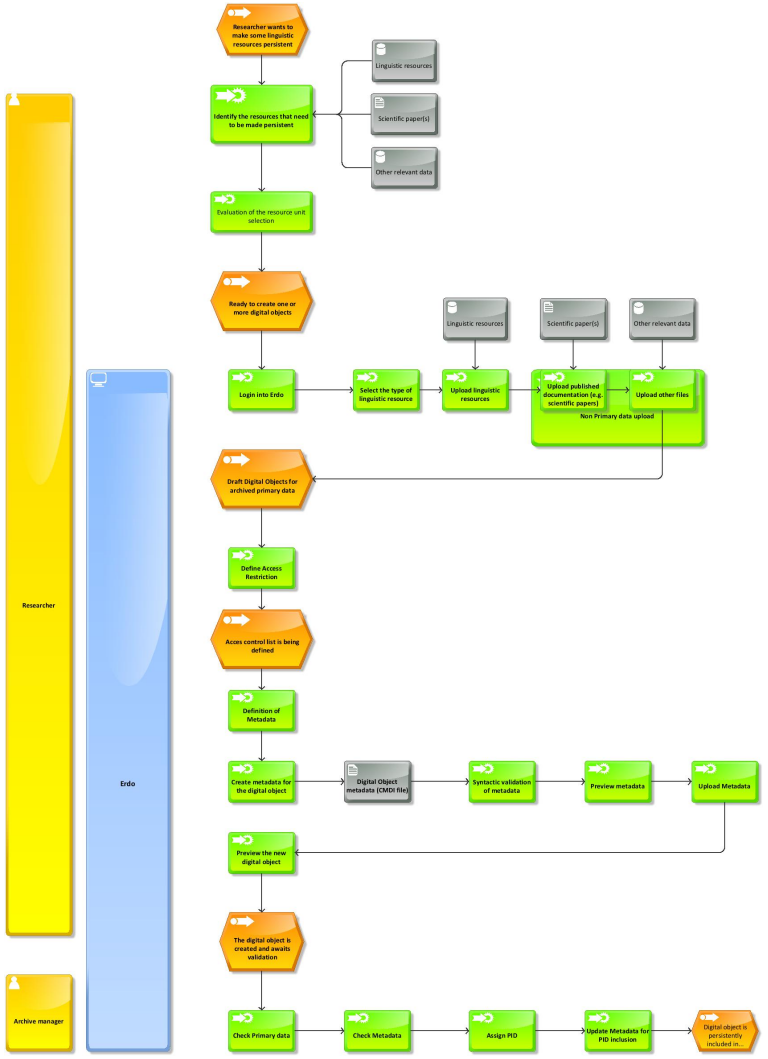


Phase 2: Entering the archive (access control)

- Project internal: all read
- Allow other individuals to read
- Allow public to read

The pay-off of persistency:

- Digital objects are not deletable
- Editing follows strict procedure
- Yes but no: workarounds

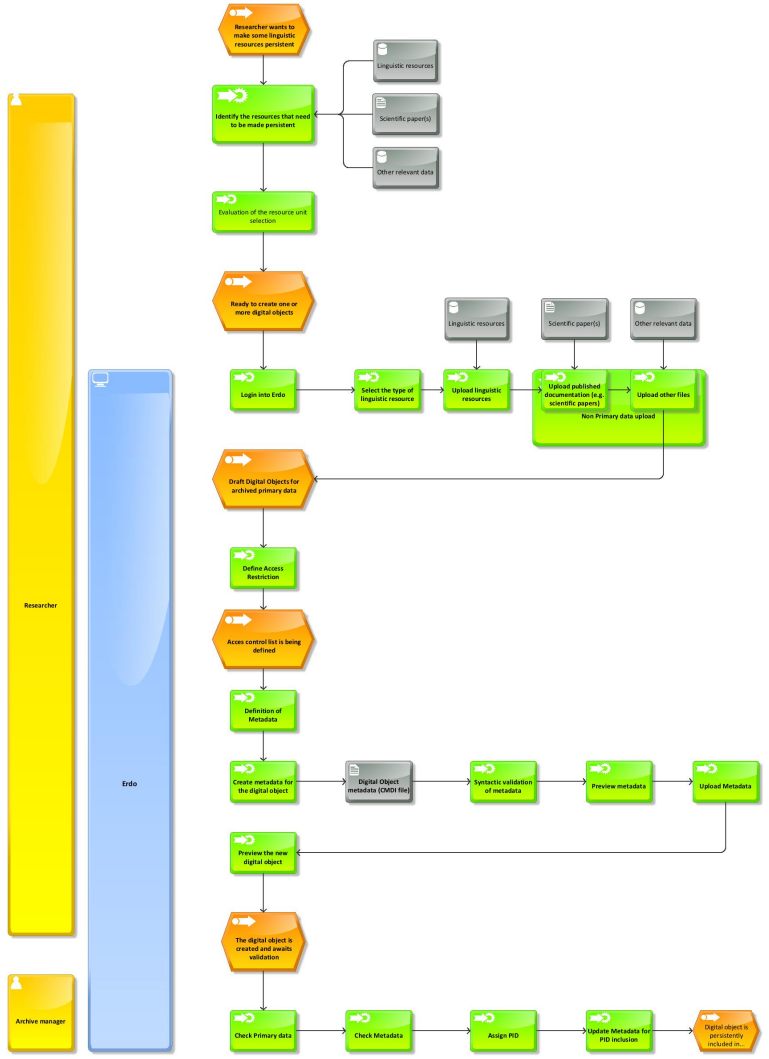


Editing and deleting persistent/archived data?

- User requirement
  - No finished LR
  - Versioning too techy
  - No changes, no data (!)
- Legal requirement
  - Cases of copyright infringement
  - Dichotomy: delete at project end vs. keep at least 10 years
- GIF vs. JPEG compared to txt-, vs. txt+,

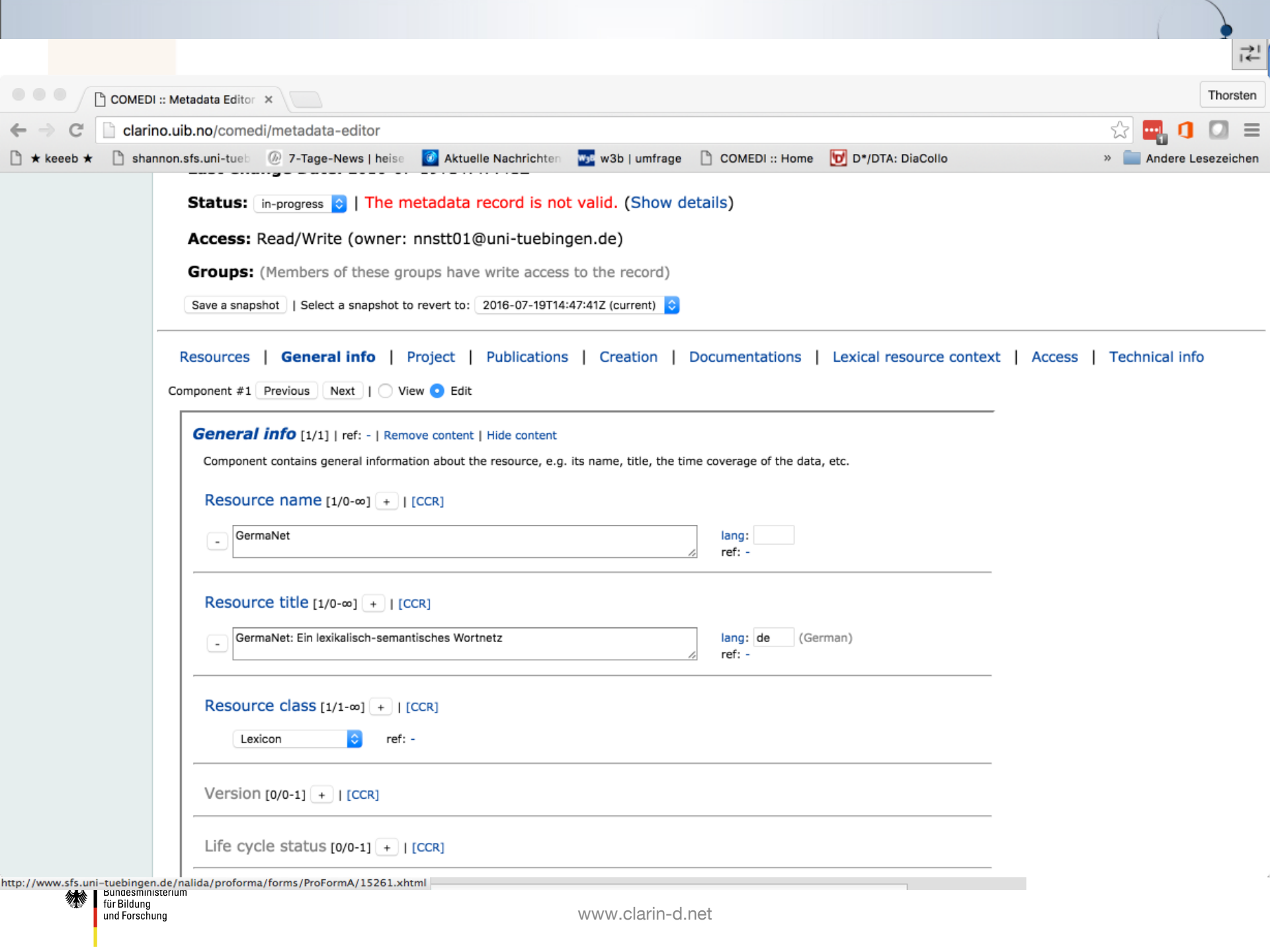
Phase 2: Entering the archive (metadata creation)

- Metadata is structured data for describing and finding resources
- Essential for archives and catalogues

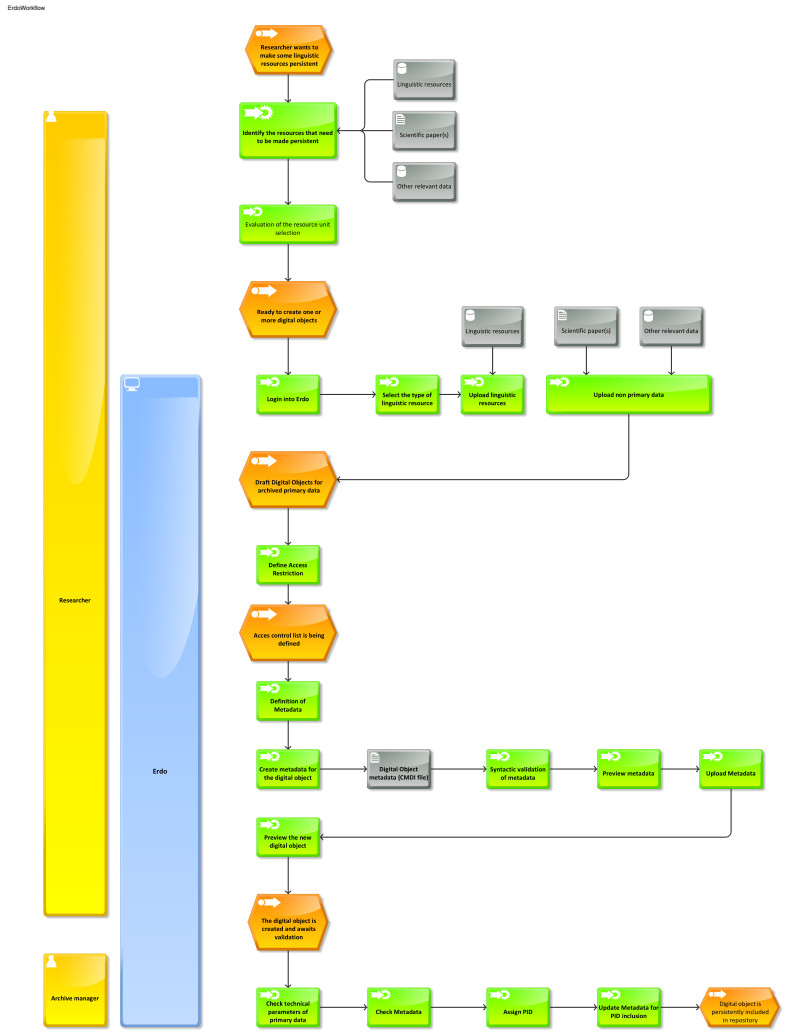


Metadata creation process

- Major concerns
  - Nobody wants to contribute metadata
  - Nobody wants to spend time on archiving
  - Tools are too cumbersome
  - Not all bits of information are available
- But:
  - Everybody wants their data too be found
  - Everybody wants to have the most correct representation of their data
  - Purpose dependent editors for the technophobe


COMEDI :: Metadata Editor
Thorsten
clarino.uib.no/comedi/metadata-editor
★ keeeb ★
shannon.sfs.uni-tueb
7-Tage-News | heise
Aktuelle Nachrichten
w3b | umfrage
COMEDI :: Home
D*/DTA: DiaCollo
Andere Lesezeichen
Status: in-progress | The metadata record is not valid. (Show details)
Access: Read/Write (owner: nnstt01@uni-tuebingen.de)
Groups: (Members of these groups have write access to the record)
Save a snapshot | Select a snapshot to revert to: 2016-07-19T14:47:41Z (current)
Resources | General info | Project | Publications | Creation | Documentations | Lexical resource context | Access | Technical info
Component #1 Previous Next | View Edit
General info [1/1] | ref: - | Remove content | Hide content
Component contains general information about the resource, e.g. its name, title, the time coverage of the data, etc.
Resource name [1/0-∞] + | [CCR]
GermaNet
lang:
ref: -
Resource title [1/0-∞] + | [CCR]
GermaNet: Ein lexikalisch-semantisches Wortnetz
lang: de (German)
ref: -
Resource class [1/1-∞] + | [CCR]
Lexicon
ref: -
Version [0/0-1] + | [CCR]
Life cycle status [0/0-1] + | [CCR]
http://www.sfs.uni-tuebingen.de/nalida/proforma/forms/ProFormA/15261.xhtml

Phase 3: Archivists take over

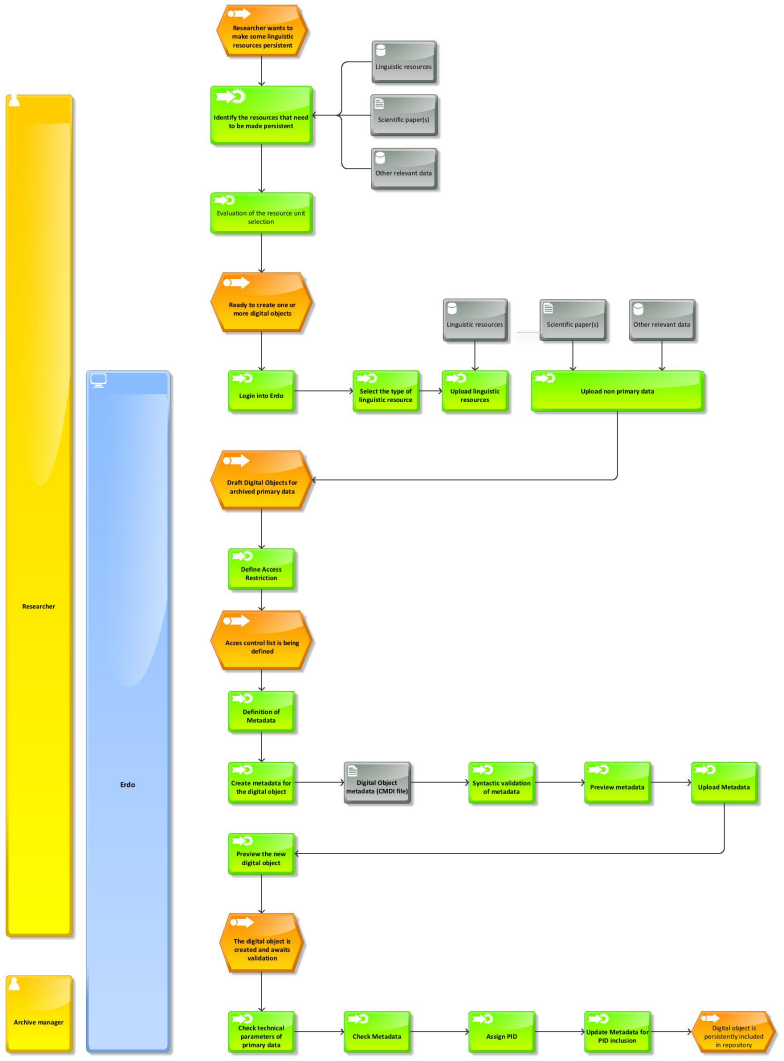


Checking the data

- Access restrictions (!)
- But:
  - File size
  - Automized processes (XML syntax parsing)
  - Metadata checks
    - Number of files
    - Interrelation of files
    - Amount of filled in data

Whole workflow

No exceptions, revisions and options shown



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Reading list

- Emanuel Dima, Verena Henrich, Erhard Hinrichs, Marie Hinrichs, Christina Hoppermann, Thorsten Trippel, Thomas Zastrow and Claus Zinn (2014). A Repository for the Sustainable Management of Research Data. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul. http://www.lrec-conf.org/proceedings/lrec2012/pdf/470_Paper.pdf
- S. Strathmann. 2009. Einfuhrung. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, and K. Huth, editors, nestor Handbuch: Eine kleine Enzyklopadie der digitalen Langzeitarchivierung, chapter 3. Rahmenbedingungen fur die LZA digitaler Objekte. Verlag Werner Hülsbusch, 2. edition.
- Report on Guidelines for Common Policies Implementation (1) (2017): http://www.parthenos-project.eu/Download/Deliverables/D3.1_Guidelines_for_Common_Policies_Implementation.pdf