Gefördert durch
DFG Deutsche Forschungsgemeinschaft

# User-oriented Guidelines for Data Ingest

**Juli 8, 2024**

| | |
|---|---|
| Version | Version 1.0 (final) |
| Redaktion | 8.7.2024 |
| Redaktionsteam | Axel Herold mit Beiträgen von Mitarbeiterinnen und Mitarbeitern aus der TA Lexical Ressources |
| Projekt | Text+ - Sprach- und textbasierte Forschungsdateninfrastruktur |
| Bezeichnung | LR 1.1 User-oriented Guidelines for Data Ingest |
| Förderung | DFG Förderkennzeichen 460033370 |
| Laufzeit | 01.10.2021 bis 30.09.2026 |
| Berichtszeitraum | – |



*Publikationsserver des Leibniz-Instituts für Deutsche Sprache*
*URN: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-130134*

# Summary

In this deliverable, partner institutions describe their established workflows for data ingest into their data repositories. As these repositories predate the attempts at creating a consolidated national research data infrastructure and the Text+ project in particular, the approaches and procedures vary to some degree.

# 1 BBAW

The Centre of the German Language (Zentrum Sprache) at the Berlin-Brandenburg Academy of Science and Humanities (BBAW) has operated a research data repository for lexical data since 2013. The repository was created in the context of the CLARIN-D infrastructure project[1] following requirements of the European CLARIN initiative[2]. It is part of the CLARIN-D Centre (Type B)[3] at the BBAW. After the completion of the CLARIN-D project, operation and maintainance of the repository was taken over by Text+ and continues under the established name of the repository as "Clarin Center at the BBAW".

The Clarin Center at the BBAW focuses on two different categories of data:

1. **lexical resources** (e.g. dictionaries provided by the "Digitales Wörterbuch der Deutschen Sprache"[4] (DWDS, Digital Dictionary of the German Language), and
2. **historical text corpora** (predominantly provided by the "Deutsches Textarchiv"[5] (DTA, German Text Archive).

The Clarin Center at the BBAW meets the technical requirements for securing long-term availability of linguistic and lexicographic services and resources. The repository as the technical core of the centre is certified by the "Core Trust Seal"[6] (CTS) which is granted by an independent international agency. In particular, the CTS covers organizational aspects as well as issues concerning data workflow, quality management and sustainability.

The centre's repository is open to the publication of research data from external institutions if the data fall into one of the two main data categories. As data hosting for external bodies requires prior mutual agreement, the ingest process is initiated by contacting the Clarin Center, e.g. directly using its contact form at https://clarin.bbaw.de/kontakt, via the Text+ helpdesk or by contacting staff at the centre directly.

There are three main sets of constraints that need to be considered and assessed before a resource can be ingested into the repository: technical constraints as to the actual data storage formats, legal constraints, and data quality and consistency. All of these constraints apply to the primary data as well as to the metadata descriptions.

## 1.1 Technical constraints

Data storage in the repository is restricted to textual data (both primary data and annotations). Storage of sound our image data is generally not provided. Accepted formats must be based on TEI[7], preferably TEI Lex-0[8] and be accompanied by a schema (e.g. in XSD, Relax NG, Schematron) that describes the format and enables automatic validation of the data. For historical data, the use of the DTA Base Format (DTABf)[9] is strongly encouraged.

---

[1] https://www.clarin-d.net/
[2] https://www.clarin.eu/
[3] https://hdl.handle.net/11372/DOC-93
[4] https://www.dwds.de/
[5] https://www.deutschestextarchiv.de/
[6] https://www.coretrustseal.org/
[7] https://www.tei-c.org/
[8] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html
[9] https://www.deutschestextarchiv.de/doku/basisformat/

Metadata descriptions must be provided in the CMDI (Component Metadata Infrastructure, ISO 24622-1/2) format[10] before data can be ingested into the repository.

## 1.2 Legal constraints

Resources and metadata in the repository must be open for free and unrestricted publication. The repository does not provide any means of access restriction or user authentication. It is possible, however, to only publish a metadata description of a dataset. This metadata description may then include pointers to storage locations of the primary data at arbitrary external data hosting facilities which may implement access restrictions.

Acceptable data licences are primarily Creative Commons (CC) licences in accordance with the Open Science Strategy[11] of the BBAW. NC (non-commercial) and ND (no derivatives) restrictions are acceptable but discouraged because they restrict some reuse scenarios and thus do not fully align with the FAIR principles[12].

## 1.3 Data quality and consistency

Data quality and data consistency are assessed individually for each resource together with the data provider. Transcription guidelines must be provided and the workflow for data acquisition, processing and annotation must be made transparent in detail.

# 2 IDS

The IDS, as part of the Lexical Resources Task Area of Text+, accepts data to be ingested that comply with the following guidelines:

- The data comprise a lexical resource, for example, a dictionary.
- The data are lexicographically compiled data, i.e. data/texts collected and created by scholars in a multi-part work process following quality assurance procedures. Data generated exclusively by algorithmic or AI procedures without further processing and quality assurance cannot be considered for ingest.
- German is the primary object language of the resource, i.e. the “language being described”.
- The resource is submitted in an open, system and platform-independent format such as XML, JSON, or CSV, preferably encoded using established lexicographic data specifications such as TEI Lex-0.
- The data schema and input conventions are documented (the documentation must also be provided in an open format, as above).
- The XML is successfully validated against a referenced schema language, e.g. DTD or RELAX NG.
- A CMDI-compliant metadata scheme describing the data is submitted alongside the data.[13]

[10] https://www.clarin.eu/content/cmdi-component-metadata-infrastructure
[11] https://www.bbaw.de/bbaw-digital/open-science
[12] https://www.go-fair.org/fair-principles/
[13] http://repos.ids-mannheim.de/resources/LZA_IDS_Depositing_Policy.pdf

- The data, in its entirety, follows the Open Access guidelines laid out by the IDS.[14] This requires a Creative Commons Attribution International license[15] (CC-BY), and that the licence holder is listed in the metadata document.
- The resource has been appraised by two senior members of the Department of Lexical Studies of the IDS ("Lexik") for quality and potential for contribution to the field of lexicology, and approved for possible ingest. This includes adherence of the data to CARE principles[16].
- The acceptance of a project for ingest remains dependent on the available capacity of the Long-Term Archive of the IDS ("Langzeitarchiv").
- The resource is to be stored and can be accessed in the "Langzeitarchiv", where it will comply with the in-house standards for data management.[17] The IDS is under no obligation to integrate the resource into any tool or work, either existing or future, but the resource will be accessible for download by third parties.

If a resource does not fully comply with the above guidelines, the matter may be discuss with the Text+ Lexical Resources Data Controller at the IDS.

# 3 SAW

The Saxon Academy of Sciences and Humanities in Leipzig (SAW) is operating a repository for the long-term preservation of digital resources, along with their descriptive metadata, and ensure their continuous availability.[18]

The repository's focus lies on written text corpora, reference corpora, general lexical resources and linguistic resources for "under-resourced" languages. Preferably resources from these fields are integrated into the repository. Yet, the repository might also accept language-related resources from other fields, as long as they are of high scientific value for the respective communities. In any other case, the repository is glad to help finding an adequate archiving facility at another institution.

## 3.1 Conditions for the acceptance of a new resource

- The resource is the result of one or multiple research projects.
- Exhaustive Metadata is available, in CMDI or at least Dublin Core[19] (Metadata Requirements).[20]
- The data follow an established and standardized format[21] or extensive documentation for the format is available or provided alongside the resource.
- Information about the creation and legal situation of the resource is available.
- Resources are either freely available or come with a license allowing access for members of research institutions. Access to the metadata must not be limited in any way.

[14] https://www.ids-mannheim.de/bibliothek/open-access-am-ids/leitlinie
[15] https://creativecommons.org/licenses/by/4.0
[16] https://www.gida-global.org/care
[17] https://www.ids-mannheim.de/fileadmin/org/Richtlinien/Leitlinie_FDM_IDS_version_2023-01-18.pdf
[18] https://www.clarin.eu/centres/saw
[19] https://www.dublincore.org/
[20] https://repo.data.saw-leipzig.de/depositing-metadata/en
[21] https://clarin.ids-mannheim.de/standards/views/view-centre.xq?id=SAW

## 3.2 Depositing procedure

The SAW's depositing procedure is described extensively on the website of the SAW's repository.[22] The key steps are summarized here:

1. Submission of a filled out resource deposition request form (RDRF), to clarin@saw-leipzig.de, by the depositor, with possible rounds of feedback and resubmission.
2. After the request form was accepted, acknowledgement and signing of the depositor's agreement by the depositor.
3. Preparation of a Submission Information Package (SIP) by the depositor, consisting of:
    a. the signed depositor's agreement
    b. metadata corresponding to the submitted resource
    c. an archive file containing the data and adhering to the BagIt format[23]
4. Appraisal and verification of the SIP by the data centre, with possible rounds of feedback and resubmission.
5. Preparation and ingestion of the Archival Information Package (AIP), consisting of the BagIt archive file and the resource's metadata.

# 4 UniK

UniK's repository focuses on the collection of two different types of data: audio-visual corpora, and lexical resources. It follows the CLARIN centre requirements and is CTS certified.

User guides to the repository[24] are provided online with one guide aiming specifically at the submission of resources[25]. Inquiries are possible via the help-desk of the Language Archive Cologne[26].

For the submission of lexical resources, the Kosh API service[27] is used. Technical requirements of the service are documented online.[28]

All data must be under a free and open license in order to be considered for ingest at UniK. There are no format restrictions as long as valid XML data is provided by the data provider.

# 5 UniTü

Within it's CTS-certified TALAR repository, UniTü provides standardized submission and ingest workflows for research data. A web-based GUI helps users to enter some basic metadata and to upload their data. The back-end of the repository generates CMDI files based on the metadata provided and puts all data into a BagIt archive. This archive is then forwarded to the archive manager for manual inspection.

Subsequent (human) communication between data provider and archivist assures mutual engagement in metadata refinement. All metadata is encoded according to ISO 24622-1/2 (CMDI). The data provider is required to sign a data sharing contract ("Datenüberlassungsvertrag").

---

[22] https://repo.data.saw-leipzig.de/depositing/en
[23] https://tools.ietf.org/id/draft-kunze-bagit-16.html
[24] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides
[25] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/submission-guidelines
[26] https://lac.uni-koeln.de/
[27] https://kosh.uni-koeln.de/
[28] https://kosh-docs.vercel.app/deployment/backend

Different resource types (e.g., Lexical Resource, Text Corpus) have different CMDI profiles. Prior to the ingest, the CMDI metadata descriptions are finalized and validated.

The TALAR repository has a strong preference for non-proprietary formats. Data may have to be converted before the actual ingest (e.g., conversion from Excel format to CSV). The repository focuses on the collection of wordnet type data in the GermaNet XML format[29].

# 6 UniTr

The Trier Center for Digital Humanities (TCDH) does not maintain a data repository. However, it does offer the option to include data for integration and publication in the Trierer Wörterbuchnetz[30]. As a public platform, the Trierer Wörterbuchnetz offers access to a large number of dictionary resources that are linked to each other. The focus is on dictionaries on the German language (general language dictionaries, historical language dictionaries, regional or dialect dictionaries, technical dictionaries and authors' dictionaries).

The inclusion of a resource is dependent on various requirements:

- The data should be scientifically compiled dictionary resources that match the orientation of the Wörterbuchnetz.
- Mainly text data is published, but image and map material can be embedded if required.
- The data must be under a free and open license.
- The required data format is XML, preferably TEI Lex-0, and the data is successfully validated against a referenced schema.

Data that is suitable in terms of content and legal status but that does not meet the format requirements, may still be prepared for integration depending on the capacities of the TCDH. In those cases, cooperation projects will be set up together with the data provider.

# 7 DSA

The Research Center »Deutscher Sprachatlas« (DSA) at the Philipps University Marburg mostly collects data on physical mediums. It operates a central "check-in" folder for data ingest. Basic metadata are collected prior to data depostition:

- record data type
- contact adress
- data description
- check-in date
- in-house employee responsible

After the collection of basic metadata, the appropriate in-house work-group is notified and starts the inspection of the data. Further metadata are generated throughout this process:

- employee responsible
- checkout-out date
- new location

---

[29] https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/lexica/germanet-1/datenformate/xml-files/

[30] https://woerterbuchnetz.de/

Further procedures are dependent on the results of the inspection.

In the near future, DSA will start transitioning to DSpace's Simple Archive Format[31].

[31] https://dspace.lyrasis.org/