RDMO – Catalogue for Text+

Version 1.1, 23.10.2024

The following questions in the "General" section serve as a general description of the project; all other sections refer to the description of the specific data sets.

Content:

General

Topic

What is the main research question of the project?

Describe briefly the project and its aims.

Please give some keywords describing the research question.

If possible, identify the keywords by using standardized vocabulary, for example, the GND or Wikidata. Example: GND: Sophocles, GND, 118615688

Please enter a keyword:

On which authority record is the keyword based?

☐ GND (authority file)
☐ Wikidata
☐ Other:

Research field

Which research field(s) does this project belong to?

Please select a suitable category from the main levels of the DFG subject classification system.

Project schedule

When does the project start?

When does the project end?

Funding

Who is the funder of the project?

Please also enter the respective funding code, if available.

☐ German Research Foundation (DFG): ______________________________
☐ European Commission (EC) - Horizon Europe: ______________________________
☐ Federal Ministry of Education and Research (BMBF): ______________________________
☐ Volkswagen Foundation: ______________________________
☐ Austrian Science Fund (FWF): ______________________________
☐ Swiss National Science Foundation (SNF): ______________________________
☐ Other: ☐

Does the funder have rules or recommendations for data management? If yes, please briefly outline them and refer to more detailed sources of information if necessary. Please also indicate, if the rules / guidelines are mandatory or optional.

„In principle, all funding programmes that describe a concrete work programme for a research project must include a description of how the research data will be handled.“
https://www.dfg.de/en/principles-dfg-funding/basics-and-principles-of-funding/research-data/research-fundingl

Examples for requirements regarding the management of research data in funded projects:

German Research Foundation (DFG):

- Guidelines on the Handling of Research Data (2015)
- Handling of research data. Checklist for planning and description of handling of research data in research projects (2021)

European Commission (EC):

- Guidelines on FAIR Data Management in Horizon 2020

Additional project data

At which URL is information about the project available?

Other requirements

Are there requirements regarding the data management from other parties (e.g. the scholarly/scientific community)?

☐ Yes
☐ No
☐ To be clarified

If yes, which are these additional requirements regarding data management?

Please refer to further information if necessary. Please also indicate the degree of binding nature of this information. Examples of subject-specific recommendations and guidelines from the **humanities and social sciences** are

Examples of discipline-specific requirements for the **humanities and social sciences** are:

☐ "Linguistics" Review Board on data standards and tools in the collection of language corpora (2019) (in German only)
☐ "Literary Studies" Review Board on funding criteria for scholarly editions in literary studies (2015) (in German only)
☐ "Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies" Review Board on the handling of research data (2019) (in German only)
☐ "Social Sciences" Review Board on the handling of research data in sociology, political science and communication science (2020) (in German only)
☐ "Economics" Review Board on the handling of research data (2019)
☐ "Ancient Cultures”. Review Board on the handling of research data (2020) (in German only)
☐ "Educational Research" Review Board on the handling of research data (2020) (in German only)
☐ "Psychology" Review Board on the handling of research data (2020)
☐ "Theology" Review Board (2022) (in German only)
☐ Other:

You can find the complete list of subject-specific recommendations on the handling of research data here.

Project coordination

Which institute(s) is responsible for the project?

Which persons (applicant / spoke person / project lead) are responsible for the project coordination?

Please state given name, surname, email and identificator (ORCID Search), if available.

Project partners

Please enter the detailed name of the institutional project partner(s).

Who is/are the contact person(s) for research data management at the partner institute?

Please state given name, surname, email and identificator (ORCID Search), if available.

Content classification

Datasets

Text+ offers a corresponding infrastructure for the long-term provision and archiving of reusable text- and language-based research data, in which various data and competence centres are involved, each with their own expertise on specific data types (collections, editions, lexical resources).

The inclusion of data in one of the institutions involved in Text+ ensures sustainable data storage (in accordance with FAIR/CARE) and increases reusability and retrievability via central reference systems (see also the sections on "Data usage and publication" and "Long-term preservation").

The following questions collect information on the data that is produced or used in the project. They also help to estimate the value of the data in terms of potential re-use and long-term preservation.

Before data are newly created, it is advisable to check if there are existing data that could be re-used. This way, redundant collection or creation of research data is prevented. This saves efforts and costs. If personal data are concerned, the principles of the General Data Protection Regulation (EU) (GDPR) like e.g. data minimisation (Art. 5 par. 1) and the Federal Data Protection Law (Bundesdatenschutzgesetz, BDSG).

The definition of what constitutes a dataset is an important conceptual decision that needs to be made and carefully balanced individually for every project. You yourself determine what is understood as a dataset within your project.

For example, a data set can consist of more data files of different types (numeric, image, text, ...) grouped together, which collect all results coming from an investigation series on a given research object.

What kind of data types are you expecting?

Please select a category.

Data type

☐ Tabular data (e.g. measurement series)
☐ Text (e.g. digital edition, transcription of interview)
☐ Database
☐ Audio file(s) (e.g. interview, voice recording)
☐ Visual file(s) (e.g. film, photography, figure)
☐ Geospatial data
☐ Software (e.g. developed within the project)
☐ 3d model (e.g. digital reconstruction of a stone age settlement)
☐ Quantitative online survey
☐ Other:

If possible, please indicate to which of the three Text+ data domains the respective dataset belongs.

☐ Collections (individual text corpora, text annotations, interviews, transcriptions, sensor data, surveys, etc.)
☐ Editions (OCR files, manuscripts and their transcriptions, text apparatuses, etc.)
☐ Lexical Resources (translations, terminologies, word nets, word lists, etc.)

Data origin

Is the dataset being created or re-used?

If the dataset is being re-used, please also consider the information under "Intellectual property rights".
☐ Created
☐ Re-used

(a) If re-used, who created the dataset?

Please name the respective persons or institutions.

(b) If re-used, under which address, PID or URL can the dataset be found?

(c) If re-used, which license (e.g. Creative Commons and/or terms of use) are subject of the data set?

(Hint: If the license/conditions for reuse are not yet clear: Which steps are you planning to undertake to clarify them?)

Which analog sources are the basis for the dataset?

(e.g. manuscript, language atlas, archival source; if available, please indicate archive signature! See also Bundesarchiv: Referencing)

Data collection

When does/did data collection or creation start?

When does/did data collection or creation end?

How was the dataset created?

Please describe the method used to create or collect the data. Additional information, for example, if voice recordings have been analyzed by grammar of phonetics, should also be provided, as well as self-developed software. Please also explain whether this is a fixed data set stored at a certain point in time or whether it is created dynamically over a certain period of time.

Technical classification

Formats

Which file formats will be used?

When choosing a data format, one should consider the consequences for collaborative use, long-term preservation as well as re-use. It is advisable to prefer formats that are standardised, open, non-proprietary, loss-free and well-established in the respective scholarly community. A good overview of this is provided by forschungsdaten.info (only in German). Further criteria and detailed explanations can be found here (only in German): WissGrid-Leitfaden, p. 22 f.

☐ Tagged Image File Format (TIFF, TIF) for images
☐ Waveform Audio File Format (WAV) for audio files
☐ Plain Text Document (TXT, ASC) for documents

☐ Portable Document Format/A (PDF/A) for documents
☐ Extensible Markup Language (XML) for documents
☐ Comma-Separated Values (CSV) for tabular data
☐ JavaScript Object Notation (JSON) for data exchange
☐ Other:

Data size

What is the actual or expected size of the dataset?

If large amounts of data are involved, financial resources for the provision of the infrastructure should be considered (see also section "Funding").

☐ less than 1 GB
☐ between 1 GB and 10 GB
☐ between 10 GB and 100 GB
☐ between 100 GB and 1 TB
☐ between 1 TB and 10 TB
☐ > 10 TB
☐ exact size:
☐ not yet determined

Tools

Which software, technologies or processes are used to generate or collect the data? Is proprietary software (e.g. Excel) used? Which? In which version?

This information is necessary to be able to reconstruct the process (e.g. via MAXQDA, MySQL-Datenbank, R, QGIS, Gephi) by which the data was generated. It is also a prerequisite to judge the objectivity, reliability and validity of the dataset.

For reproducible data, it is also required to re-generate the data if necessary. Therefore, all devices, software, software version and also information about the procedure necessary to be able to recreate the data must be preserved.

Metadata and referencing

Metadata

Which standard vocabularies are used for unequivocal identification entities within the data set?

(e.g. GND for persons, Wikidata for objects and concepts, GeoNames for geographica)

Which standards, ontologies, classifications etc. are used to describe the data and context information?

(e.g. CIDOC CRM, Europeana Data Model, TEI-XML, MEI-XML)
Consider already at this point that discipline-specific repositories for long-term storing and publication of data each have their own requirements to the metadata schema (see section "Data sharing and re-use").

☐ Discipline-specific standards, classifications etc. are used.

☐ A custom description system is used (please briefly outline and, if necessary, indicate where it is documented in more detail):

☐ Other:

☐ It has not yet been decided with which system the metadata and contextual information will be described.

☐ No fixed system for the description is used.

How will the metadata be collected?

☐ Automatic collection (e.g. timestamp of an interview, recording data of a photography):

☐ Semi-automatic collection (what metadata are automatically collected by the computer/software, but must be checked?):

☐ Manual creation (e.g. data on location, dates or persons that are manually extracted from historical documents. A metadata editor is suitable for this.):

Persistent Identifiers (PIDs)

Persistent identifiers (PIDs) are intended to enable the permanent referencing of (in particular) digital objects such as publications or research data. Instead of referring to the storage location of the object, as is usually the case when specifying a hyperlink as a reference, the PID acts as an intermediate instance from which the object is forwarded (this is called "resolving" the PID). The PID remains the same even if the storage location of the object changes. While a hyperlink would lead nowhere in this case, the object can still be reached via the PID. Further information on how PIDs work, how they are used and the different types of PIDs can be found, for example, in the Digital Preservation Coalition handbook or on the website of the PID Network Germany or PID4NFDI.

Will persistent identifiers (PIDs) be used for this data set?

☐ Yes

☐ No

Which system of persistent identifiers shall be used?

☐ Handle / DOI
☐ PURL
☐ ARK
☐ URN
☐ ISLRN
☐ Other:
☐ Commit tag on a code repository:
☐ None

Which (sub-) entities / sub units should be referenced using identifiers? Which of those identifiers should be persistent and citable?

Who is responsible for the maintenance of the PIDs and the object maintenance (i.e. who is responsible notifying the PID-Service about object relocation and the new address)?

A prerequisite for PIDs to work as promised is that they - as well as the objects they refer to - are maintained in a continuous and reliable way. This means, for example, that if the object location changes, this information is updated.

When the data are stored in a data centre or repository, these tasks are usually taken care of by the data centre / repository. However, to be sure, the responsibilities should be checked beforehand.

Legal and ethics

General legal issues

The website forschungsdaten.info provides a general overview of the legal aspects of research data. The Text+ Helpdesk also provides information on legal and ethical issues in text- and language-based research, such as copyright, but does not offer legal advice.

Does the legal situation of different countries have to be considered?

If you answer this question with "Yes", please get in touch with the legal department or a respective contact person at your institution to clarify if this has consequences for the project and its data management and if yes, what consequences these are.

☐ Yes
☐ No

Personal Data

Does this dataset contain personal data?

☐ Yes

☐ No

If so, please get in touch with the legal department or an appropriate contact person at your institution to clarify whether there are any consequences for your project and, if so, what these are.

Please also check whether you have completed the sections under "Data use and publication" with the appropriate measures to protect the data in accordance with the applicable data protection laws.

In principle, the following points must be documented for personal data:

- **Purpose of processing / consent management**
- **Time limit for storage**
- **Legal basis**

The handling and processing of personal data is regulated by law. The EU-wide standardised application is based on the EU General Data Protection Regulation (GDPR). It allows room for manoeuvre at national level. In Germany, this is regulated by the Federal Data Protection Act (BDSG). For universities, the individual data protection laws of the federal states apply to a large extent, e.g. the Data Protection Act of North Rhine-Westphalia (DSG NRW). The European GDPR defines personal data as any information relating to an identified or identifiable natural person (Art. 4 GDPR, para. 1). A person is identified if it is clearly recognisable to which person the data belongs. A person is identifiable if they can be identified by means of additional information.

You can find more information on this topic on the website of forschungsdaten.info.

Data protection

In addition to the European GDPR, which laws must be observed with regard to data protection issues for the project?

The EU General Data Protection Regulation (GDPR) takes precedence over national law. It has been applicable since 25 May 2018. For universities, the state data protection laws continue to apply, which may set out certain provisions of the GDPR, e.g. the Data Protection Act of North Rhine-Westphalia (DSG NRW).

In certain areas, specific laws apply that take precedence over the state data protection laws or the Federal Data Protection Act (BDSG).

For microcensus data, the Federal Statistics Act (BStatG).

- For research in schools in NRW, the NRW School Act (SchulG NRW)
- As a result, several laws may be affected depending on the research project.

☐ Bundesdatenschutzgesetz (BDSG, Federal Data Protection Act)

☐ Landesdatenschutzgesetz Baden-Württemberg (State Data Protection Act of Baden-Württemberg)

☐ Landesdatenschutzgesetz Bayern (State Data Protection Act of Bavaria)

☐ Landesdatenschutzgesetz Berlin (State Data Protection Act of Berlin)

□ Landesdatenschutzgesetz Bremen (State Data Protection Act of Bremen)

□ Landesdatenschutzgesetz Brandenburg (State Data Protection Act of Brandenburg)

□ Landesdatenschutzgesetz Hamburg (State Data Protection Act of Hamburg)

□ Landesdatenschutzgesetz Mecklenburg-Vorpommern (State Data Protection Act of Mecklenburg-Vorpommern)

□ Landesdatenschutzgesetz Hessen (State Data Protection Act of Hesse)

□ Landesdatenschutzgesetz Nordrhein-Westfalen (State Data Protection Act of North Rhine-Westphalia)

□ Landesdatenschutzgesetz Rheinland-Pfalz (State Data Protection Act of Rhineland-Palatinate)

□ Landesdatenschutzgesetz Niedersachsen (State Data Protection Act of Lower Saxony)

□ Landesdatenschutzgesetz Saarland (State Data Protection Act of Saarland)

□ Landesdatenschutzgesetz Sachsen (State Data Protection Act of Saxony)

□ Landesdatenschutzgesetz Sachsen-Anhalt (State Data Protection Act of Saxony-Anhalt)

□ Landesdatenschutzgesetz Schleswig-Holstein (State Data Protection Act of Schleswig-Holstein)

□ Landesdatenschutzgesetz Thüringen (State Data Protection Act of Thuringia)

□ Sozialgesetzbuch X (Volume X of the German Social Insurance Code), e.g. for medical data

□ Bundesstatistikgesetz (German Federal Statistics Act), e.g. for census data

□ Other:

Sensitive Data

Does the data set contain special categories of personal data?

□ Yes

□ No

According to the GDPR, this includes personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation (Art. 9 GDPR, para. 1) (General Data Protection Regulation). Such data is considered particularly sensitive and requires even stricter protective measures than those already required for personal data in general.

If you answer this question with "Yes", please get in touch with the data protection officer of your institution to check which additional protection measures are necessary.

Will the data be anonymised or pseudonymised?

□ Yes, during the collection

□ Yes, before / at the beginning of the data analysis

□ Yes, after the data analysis / before publication

□ No

To what extent is the "informed consent" obtained from the persons concerned?

Basically, the collection, processing, archiving and publication of personal data is only admissible, when the "informed consent" of the person in question has been obtained.
If necessary, please contact the data protection officer at your institution to find out about the necessary conditions for consent.

☐ For analysis / use of the data within the project as well as for re-use
☐ Only for analysis / use of the data within the project
☐ For analysis / use of the data within the project as well as for long-term storage

Where and how is the "informed consent" documented?

Deletion management: By when will the (not anonymised or not pseudonymised) original data be safely deleted? Who is responsible for this?

Other sensitive data

Does this dataset contain sensitive data other than personal data?

Examples are data that contain trade or business secrets or geoinformation on endangered species.
☐ Yes
☐ No

(a) If yes, please describe the non-personal sensitive data used in the project?

Official approval

Certain areas of research require an official approval in order to protect humans, animals and the environment. A compilation of authority contacts and information on the subject of research can be found here.

Has the project been approved by a research ethics committee?

☐ Yes, reviewed and approved by the following committee: ____________________
☐ Yes, approved under obligations which will be complied in the following way:

☐ Not yet, but it is already in the review process.
☐ Not yet, it will be handed in for review by: ______________________________
☐ No, a review is not necessary, because: ______________________________

Is a legal approval / permit needed for the research?

☐ No
☐ Yes. The permit has been received.
☐ Yes. The permit has been applied for on: ______________________________
☐ Yes. The permit will be applied for by: ______________________________

If yes, which permit?

If yes, which is the responsible agency?

Is a data access committee needed to handle access requests to the published data of the project?

☐ Yes
☐ No

Intellectual property rights I

Does the project use and/or produce data that is protected by intellectual or industrial property rights?

In this case, you should at an early stage contact respective support (e.g. your institution's legal department or a legal consulting services of a library).

Data or software may be subject to copyright or other intellectual property rights. The legal situation can vary considerably from country to country, even within the EU. In Germany, works of literature, science and art that constitute a "personal intellectual creation" are protected by copyright under the Copyright Act (UrhG). Copyright protection expires 70 years after the death of the author.

Pure data, e.g. measurement data or survey data, but also metadata (except for any "descriptive" metadata) are not eligible for protection.

Section 2 of the Copyright Act lists the following protected types of work, although the list is not exhaustive:

- linguistic works, such as written works, speeches and computer programmes
- musical works

- pantomime works, including dance works
- works of visual arts, including works of architecture and applied arts and designs of such works
- photographic works, including works similar to photographic works
- representations of a scientific or technical nature such as drawings, plans, maps, sketches, tables and plastic representations.

According to § 3, translations and other adaptations of works which are personal intellectual creations of the adaptor are also protected.

Finally, collective works and database works are also protected under Section 4, which may well be relevant in the field of research data. Collective works are defined as collections of works, data or other independent elements which, because of the selection or arrangement of the elements, constitute a personal intellectual creation.

A database work within the meaning of the law is a collective work whose elements are systematically or methodically organised and individually accessible by electronic means or otherwise.

Other relevant intellectual property rights can be industrial property rights such as patents, utility models, plant variety protection [for plant breeding], semiconductor protection, trademarks, geographical indications, registered designs or business names.

☐ Yes
☐ No
☐ Unknown

Intellectual property rights II

Does copyright law apply to this dataset?

Please consider also the hints in the sections "Content classification/Data origin" and "Data usage and publication/Data sharing and re-use".

☐ Work of literature, scholarship or the arts
☐ Translation or other edition of a work
☐ Collected edition or database work
☐ Other:
☐ No

Was investigated who the rights owner is?

☐ Yes
☐ No

Data usage and publication

Data storage and security

Was investigated who the rights owner is?

☐ Project members
☐ Project partners
☐ Only internals
☐ External partners

Which measures or provisions are in place to ensure data security?

☐ Protection against unauthorized access to the server of the institute
☐ Data recovery / Backup

Who is responsible for the backups?

This question refers to backups while the data are being worked with. Questions of long-term preservation will be addressed in the respective section.

Data sharing and re-use

Will the data resulting out of the project be published or shared?

☐ Yes, externally for everyone
☐ Yes, externally limited with individual approval
☐ No

If no, please explain why not. Please differentiate between legal and contractual reasons and voluntary restrictions.

If no, will the project staff be able to access the data after the project end (if they leave the institution)? If so, by which regulation?

Under which terms of use or license will the dataset be published?

The options refer to the licenses of the Creative Commons family. Multiple answers possible, e.g. BY SA).

☐ Attribution (BY)
☐ NonCommercial (NC)
☐ NoDerivatives (ND)
☐ ShareAlike (SA)
☐ Public domain (CC0)
☐ MIT License for software
☐ Other, or exact designation - if known:

Where will the data (including metadata, documentation, and relevant code or software, if applicable) published?

Name the database/repository, with a link if necessary.

When deciding on a suitable repository for publishing the data, the following features are of central importance:

- Public infrastructure, accessible free of charge
- Permanent existence guarantee of the repository
- Possibility of open licensing
- Unique referencing of resources using persistent identifiers (PIDs)
- Support of relevant metadata schemas
- Integration into higher-level search infrastructures

Repository registers: Re3data, DataCite, RIsources (DFG), OpenDOAR. In particular, the NFDI consortia of the humanities (Text+, NFDI4Culture, NFDI4Memory, NFDI4Objects) maintain up-to-date curated lists of repositories or offer suitable repositories for research data themselves for free and open use by their partners (see Text+ overview and curated repository list of 4Culture).

Examples for generic repositories are

- Zenodo (offers among other things a connection to github and is organized in communities, e.g. Text+)
- DARIAH-DE Repository (provides a separate area for the management, organization and publication of the data as collections by the DARIAH-DE Publikator
- RADAR4Culture (format-independent repository for the subjects of architecture, art and music, theatre, dance, film and media studies.)

Examples for discipline-specific repositories are

- TextGrid Repository (focus on textual data in TEI XML with a background in literary studies)
- CrossAsia Access Repository (publication server for Asian Studies, open for research data)
- GESIS (Data Services for the Social Sciences)

- CLARIN-D Centers (accepts linguistic text and language data on request; all CLARIN-D centers are also part of Text+; contact: Text+ Helpdesk

The Text+ Helpdesk provides detailed consultation to all researchers and selects together with you a suitable repository along established criteria.

See also

- Hug et al. (2023): Wohin damit? Storing and reusing my language data
- Draxler et al. (2023a): Wohin damit? So kommen Ihre Forschungsdaten in die Text+ Infrastruktur
- Draxler et al. (2023b): Wohin damit? Wir geben Ihren Daten ein Zuhause
- Leitlinie für das Integrieren von Daten in Text+/NFDI

Long-term preservation

Long-term storage

Where will the data (including metadata, documentation and, if applicable, relevant code) be stored or archived after the end of the project?

Unlike repositories, which hold their data for immediate use, a long-term archive is for long-term storage of data that may not be frequently accessed or actively used. A solution for long term archiving with bitstream preservation is currently being developed in the NFDI consortium Text+, see also Dogaru, G. (2023). Das Text+ Langzeitarchiv. (only in German)

☐ Own institution
☐ Another partner institution: ____________________
☐ Other:
☐ To be decided