Please address the following aspects in your letter of intent

1 Binding letter of intent as advance notification or non-binding letter of intent

[Please indicate clearly whether your document is a binding letter of intent as advance notification or a non-binding letter of intent.]

X	Binding letter of intent (required as advance notification for proposals in 2019)
□	Non-binding letter of intent (anticipated submission in 2020)
□	Non-binding letter of intent (anticipated submission in 2021)

2 Formal details

- Planned name of the consortium

Text+: Language- and Text-Based Research Data Infrastructure

- Acronym of the planned consortium

Text+

- Applicant institution

Leibniz Institute for the German Language, Mannheim
R5, 6-13, 68161 Mannheim
Prof. Dr. Henning Lobin

- Spokesperson

Prof. Dr. Erhard Hinrichs, hinrichs@ids-mannheim.de, Leibniz Institute for the German Language, Mannheim

- Co-applicant institution
- [Name and address]
- [Name of the head of the co-applicant institution]
- Co-spokesperson
- [Name, e-mail address, institutional affiliation]

[Please repeat for all co-applicant institutions/co-spokespersons within the planned consortium.]

- Participant
- [Name, institutional affiliation]

[Please repeat for all participants within the planned consortium.]

3 Objectives, work programme and research environment

- Research area of the proposed consortium (according to the DFG classification system)
  www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/amtsperiode_2016_2019/fachsystematik_2016-2019_en_grafik.pdf

Humanities

- Concise summary of the planned consortium’s main objectives and task areas

Language and text are the basis of human communication and, as such, they underlie all aspects of society: ethics, economy, education and research, and even the internet. “Text as data” is a rather new phenomenon but is instrumental to be addressed specifically in a national research data infrastructure. Text+ will tackle this challenge primarily from a Humanities research perspective but strives to engage with other research fields over time. Language and text also have great potential to build a bridge between data-oriented research – sometimes viewed as abstract and incomprehensible – and the general public, as the vast digital resources available for literature and other collections can easily relate to people's everyday lives.

Initially, the Text+ consortium will focus on three types of research data: language- and text-based collections, dictionaries/lexical resources, and editions – henceforth abbreviated ‘Collections’, ‘Dictionaries’ and ‘Editions’. These three types of data have a long tradition in the Humanities and have given rise to mature methodological paradigms that require distinctive, yet cross-cutting practices of research data creation, curation and management. Collections, Dictionaries, and Editions depend on one another and therefore need to be linked in a common infrastructure in terms of metadata and digital objects. On the one hand, the annotation of text-based collections and editions is highly dependent on high-quality dictionaries of various kinds. On the other hand, dictionaries and other lexical resources are greatly enhanced by references to authentic language data that can be derived from collections and editions. Thus, Collections, Dictionaries and Editions constitute the empirical basis for a wide range of disciplines, including, but not limited to, linguistics, literary studies, cultural studies, history, theology, and philosophy. They are also essential for qualitative studies in the social and political sciences.

The importance of Collections, Dictionaries and Editions for Humanities research and teaching has been underscored in a continuous dialogue with a broad range of Humanities scholars, learned societies and professional organizations. This dialogue is essential for building up a user-oriented research infrastructure and has been ongoing for more than ten years in the context of the research infrastructure initiatives CLARIN and DARIAH – for Text+ amplified through intense collaboration with the Academies of Sciences and Humanities with centuries of experience in long-term research. In preparation of the NFDI, the dialogue between user communities and infrastructure providers has been broadened by a series of workshops and coordination activities in which Text+ has taken a pro-active role.

Research data in the Humanities are typically housed in geographically distributed data and research centres, which often hinders findability, access, interoperability and reusability of data and services. The main objective of Text+ is therefore to create a FAIR-compliant research data infrastructure that overcomes these obstacles. The development of this research infrastructure will be driven by research: the portfolio of Text+ is managed predominantly by three Scientific Committees, for Collections, Dictionaries and Editions, respectively, representing individual subject experts and learned societies (professional associations). The members of the Scientific Committees are involved as Participants in the proposal, while the co-applicant institutions will provide a liaison, who are also the co-spokespersons of the consortium, to the operations of Text+ in each of these three committees. The operation liaisons (co-spokespersons) are part of a fourth committee, the Operations Committee. For implementing a balance between scientific aspects and operational aspects, the operation liaisons should not chair a Scientific Committee. The Scientific Speaker and Speaker of Operations will oversee the three Scientific Committees and the Operations Committee.

The broad participation of learned societies (professional associations) and subject experts will be guaranteed through their substantial involvement as Participants in this proposal, also serving as chairs of the three Scientific Coordination Committees for Collections, Dictionaries and Editions. The three chairs of the Scientific Committees, the Operations Speaker and the Scientific Speaker will form the Executive Board for final decision-making. In this way, the consortium is compact in that only few co-applicant institutions are necessary, while a broad representation of the research communities is realized through Participants. The representation of researchers by their majority of votes in the Executive Board guarantees that research aspects drive the agenda of Text+. Text+ is also committed to diversity for the composition of its committees and particularly for the composition of its Executive Board, e.g. gender balance, variety of disciplines, etc.

In accordance with the objectives and governance, Text+ will be organized in terms of the following task areas, each of which will be managed by one of the four co-applicant institutions and co-spokespersons. These ensure the long-term viability of the infrastructure and contribute many years of experience in research data collection, management, and provisioning.

In the Task area Collections, the German National Library (DNB) in Leipzig and Frankfurt a. M. will serve as co-applicant institution, with Dr. Peter Leinen acting as its co-spokesperson. The co-applicant institution Berlin-Brandenburg Academy of Sciences and Humanities and its co-spokesperson Dr. phil. habil. Alexander Geyken will coordinate the Task area Dictionaries/Lexical Resources. In the Task area Editions, Prof. Dr. Andreas Speer, who is representing the co-applicant institution(s) University of Cologne/North Rhine-Westphalian Academy of Sciences, Humanities and the Arts will serve as coordinator. The co-applicant institution University of Göttingen (Göttingen State and University Library, SUB) will coordinate the Task area Operations, with Prof. Dr. Wolfram Horstmann acting as co-spokesperson, i.e. Speaker of Operations. Spokesperson of the consortium, i.e. Scientific Speaker, will be Prof. Dr. Erhard Hinrichs for the Leibniz Institute for the German Language in Mannheim, which acts as applicant institution for Text+. The Leibniz Institute for the German Language is a key centre for language- and text-based research and related research data provisioning, and has significant experience in national and European research collaboration and management.

- Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfil the planned consortium's objectives

Due to the distributed nature of the resources, software tools, and services that Text+ will provide for the NFDI, the consortium will put in place a common technical backbone for a federation of geographically distributed data centres and competence centres. This backbone will consist of a layered set of infrastructure protocols and services. These protocols and services will ensure the seamless integration of research data and services from different data and service providers within the Text+ consortium.

A set of basic infrastructure services, which include persistent identifier services, uptime monitoring services for data repositories and software tools, authentication and authorization services (AAI), will provide a solid foundation. A set of extended services will include a federated metadata infrastructure, federated metadata and research data search services, and services for the long-term archiving of research data. Furthermore, a set of extended services will support different stages of the research data lifecycle, including platforms for the annotation, analysis, visualization, and publication of research data.

Text+ will closely collaborate with the NFDI directorate in order to ensure that the technical infrastructure of Text+ will be fully compliant with the technical requirements of the NFDI as a whole.

International collaboration of Text+ includes active participation in the Research Data Alliance, in Standardization organizations such as relevant technical sub-committees of ISO, TEI and the W3C consortium, in order to be able to contribute to and promote implementation of internationally accepted standards for research data and research data management. At the

European level, Text+ will continue to coordinate its infrastructure activities with the ESFRI European Research Infrastructure Consortia CLARIN ERIC and DARIAH ERIC. Members of the Text+ consortium will also continue to actively contribute to the construction of a European Open Science Cloud for the Humanities and Social Sciences via HORIZON2020 projects, such as EOSC-Hub and SSHOC.

- Interfaces to other proposed NFDI consortia: brief description of existing agreements for collaboration and/or plans for future collaboration

In the area of the Humanities, extensive consultations between infrastructure providers and infrastructure users have taken place in preparation of the NFDI. These consultations took place in a workshop series and in the context of an NFDI Forum in order to identify the actual needs of Humanities and Social Science scholars across a wide range of disciplines. In addition, interfaces to three other NFDI consortia (NFDI4Culture, NFDI4Memory, and NFDI4Objects) have been established in the course of several coordination meetings. These interfaces and the planned collaboration are described in a Memorandum of Understanding (MoU). This MoU has been published electronically at https://doi.org/10.5281/zenodo.3265763 . Moreover, a cooperation on quantitative and qualitative language- and text-based research methodologies will also be established with KonsortSWD, where Text+ focuses on metadata and text resources while KonsortSWD focuses on anonymization procedures.

Textual resources are used in a variety of scientific disciplines. For example, scholarly communication in virtually every scientific discipline is language- and text-based. Therefore, interfaces to cross-disciplinary services such as text data mining, terminologies, and authority data are planned. In addition to interfaces to the above-mentioned NFDI consortia, Text+ identified common interfaces with NFDI consortia as diverse as NFDI4Ing and NFDI4Earth or NFDI4BioDiv.

NFDI4Ing and Text+ will collaborate in scoping out and leveraging the potential of text mining techniques for the extraction of research data from pertinent digital scholarly literature and documents.

In addition, we are in contact with other NFDI initiatives who intend to submit proposals in the submission rounds of the years 2020 and 2021.

4 Cross-cutting topics

- Please identify cross-cutting topics that are relevant for your consortium and that need to be designed and developed by several or all NFDI consortia.

Cross-cutting topics relevant for Text+ concern technical infrastructure services, particularly for ensuring compliance with FAIR principles, and services for infrastructure users. Technical infrastructure services include Authentication and Authorisation Services, Persistent Identifiers for research data, protocols for metadata creation and harvesting in a distributed infrastructure, certification services for data centres, as well as easy access to data storage and archiving facilities as well as High Performance Computing.

Cross-cutting services for infrastructure users include search and interoperability solutions for federated metadata and research data, text data mining, training activities, help desk support, especially for best practices and standards relevant to research data encoding and research data management, as well as access to expertise on legal and ethical aspects of research data.

- Please indicate which of these cross-cutting topics your consortium could contribute to and how.

Text+ will contribute to cross-cutting topics by sharing expertise and practical experience with other NFDI consortia in the following areas:

Authentication and Authorisation Services:
Identity Management (IdM) and Authentication-/Authorisation Infrastructure (AAI) solutions for distributed research data infrastructures leveraging Open Source Software and standard protocols (e.g. SAML/Shibboleth, OpenID Connect) as well as Identity- and Service-Provider proxy mechanisms for existing research federations (e.g. DFN-AAI, eduGAIN).

Persistent Identifiers for research data:
Usage of Persistent Identifiers for ensuring findability and accessibility of research data and aggregating information about research data and tools in resource registries.

Protocols for metadata creation and harvesting:
Active participation in standardization efforts for language resource management (ISO TC 37 SC4) and promoting the use of standards for metadata harvesting.

Metadata and authority data:

Promoting the use of common metadata models and contributing to cross-disciplinary authority data (e.g. GND – Gemeinsame Normdatei) for ensuring findability and interoperability of research data.

Digital preservation:

Active contribution to solutions for digital long-term preservation of research data.

Data quality:

Active contribution to the development of curation criteria and quality standards for research data in the Humanities and related quality management processes.

Service and software quality:

Active contribution to solutions for quality management of software and services, lifecycle management, and data centre certification.

Search and interoperability solutions:

Search and data modelling solutions that enable and enhance findability across potentially heterogeneous metadata and data sources.

Text Data Mining:

Development and deployment of Text Data Mining Techniques (DMT) for language- and text-based research data.

Training activities:

Organization of training events and summer schools for junior researchers at the national and international level, providing and contributing to learning and teaching support platforms, as well as contributing to curriculum development.

Help desk support:

Development and operation of a ticket-based help desk with a distributed network of domain experts.

Knowledge sharing of best practices and standards relevant to research data encoding and research data management, as well as access to expertise on legal and ethical aspects of research data.