Please address the following aspects in your letter of intent

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

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

2 Formal details

- Planned name of the consortium
  - AI services for Natural Language Research Data
- Acronym of the planned consortium
  - NFDI4Language
- Applicant institution
  - German Research Center for Artificial Intelligence
    (Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, DFKI)
    Trippstadter Str. 122
    67663 Kaiserslautern
- Spokesperson
  - Dr. Georg Rehm, georg.rehm@dfki.de, DFKI: Speech and Language Technology Lab

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
  - 4 – 44 (Engineering Sciences)
    - 409 (Computer Science)
      - 409-05 Interactive and Intelligent Systems, Image and Language Processing, Computer Graphics and Visualisation
      - 409-06 Information Systems, Process and Knowledge Management
  - 1 – 11 (humanities)
    - 101 (Ancient Cultures)
      - 101-01 Prehistory
      - 101-02 Classical Philology
      - 101-03 Ancient History
      - 101-04 Classical Archaeology
      - 101-05 Egyptology and Ancient Near Eastern Studies
    - 102 (History)
    - 103 (Fine Arts, Music, Theatre and Media Studies)
    - 104 (Linguistics)
      - 104-01 General and Comparative Linguistics, Typology, Non-European Languages
      - 104-04 Applied Linguistics, Experimental Linguistics, Computational Linguistics
    - 105 (Literary Studies)
    - 106 (Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies)
      - 106-01 Social and Cultural Anthropology and Ethnology
      - 106-02 Asian Studies
      - 106-03 African, American and Oceania Studies
      - 106-04 Islamic Studies, Arabian Studies, Semitic Studies
      - 106-05 Religious Studies and Jewish Studies

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

Scholarly research has become increasingly more data-intensive in all disciplines. This profound methodological change has been accelerated with the progress made in Artificial Intelligence, Big Data and Machine Learning (hereafter summarized as AI). Decades of research in Digital Humanities have shown that many methods in Natural-Language Processing (NLP), Language Technology (LT) and Natural-Language Understanding (NLU) can be applied to research questions in a meaningful way (including, among others, morphological analysis, parsing, named entity recognition, concept detection, sentiment analysis, opinion mining, text mining, summarization, machine translation, stylometrics, text comparison etc.).

In the past, the application of such methods required weeks or months of software development. We are now approaching a technical level that will enable researchers to pick and choose their tools from a portfolio of processing services, and to form individualized, adaptive and sustainable processing workflows, while addressing the FAIR principles Reusability and Interoperability.

The multilingual nature of metadata challenges the two other FAIR principles - Findability and Accessibility, especially because of the different writing systems (e.g. Arabic, Chinese, etc.) or different semantic processing approaches (e.g. Law). As scholars more regularly use data and, in some cases, AI infrastructures in their research practice, there is a particular urgency to not only provide such an infrastructure on a wider scale, but also to make their operation and results interpretable for non-technical experts. By using such an AI infrastructure, traditional human labour-centered research practices can be complemented by LT, NLP and AI techniques. This requires that such platforms act as both technological enabler and educational mediator.

At the time of writing (the group of stakeholders will be substantially extended in the next year), NFDI4Language consists of three groups of users and providers of research data: (1) research areas concerning the Non-Western World, e.g., Asian, African, Arabic, Jewish Studies but also Ancient Studies with Archaeology, Ancient Languages and Egyptology; (2) researchers from linguistics and computational linguistics; (3) researchers from the field of computer science, web-based knowledge generation and digital learning as well as AI-researchers.

NFDI4Language will connect these three fields of research with information infrastructure providers like data centers and libraries to interact in terms of research data, methods and infrastructure:

- The research data from Ancient Cultures, Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies

with common challenges like Unicode, (inter)national authority files, multilingual metadata, semantic mapping, multilingual Named Entity Recognition, OCR, etc. and the need to explore, investigate, curate, structure, annotate, summarize, translate and analyze their research data in a Non-Western Language.

- The research methods from Linguistics (including Corpus Linguistics) and Computational Linguistics: annotation, abstraction, mapping, analysis, querying, data evaluation, corpus construction, model generation, etc.
- The AI community and individual areas of computer science for the development of required technologies as well as software usability and research on existing interdependencies between people's participation processes (the social system), the employed software (the technical system) and the collectively created artifact (the knowledge system).
- Data centers and libraries as strong partners for digital infrastructure and services with deep knowledge about metadata, digitalization, information retrieval, information services and support and training services.

The technical core of NFDI4Language consists of a platform and its integrated services including defined scenarios that offer easy-to-apply/adapt solutions for non-technical disciplines. The platform has a special focus on providing and, crucially, combining existing and adaptable monolingual, cross-lingual and multilingual language technologies with AI approaches to enable powerful and flexible research data processing workflows that can be tailored to the specific research question, setting and data at hand. The system will allow for the combination and orchestration of individual processing services into sustainable, reproducible and verifiable workflows. Their analytic results can be fed into machine learning algorithms for further processing and experimenting. Thus, besides providing the processing services and workflow orchestration features, the consortia will enable scholars to develop strong mental models of current algorithms' inner workings in order to carry out meaningful research in their own disciplines based on current AI methods.

Equally important is the expansion, connection and development of the communities of practices of humanities with language data and linguistics, computational linguistics, and computer science.

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

- Anonym Classic (ERC-Projekt 2018-2022, FU: The Arabic Anonymous in a World Classic) with
    - Knowledge, Information Technology, and the Arabic Book (KITAB): http://kitab-project.org/
    - LERA, Martin Luther University Halle-Wittenberg, an interactive, digital tool for analyzing variations between multiple versions of a text in a synoptic manner: https://sada.uzi.uni-halle.de/
- BOP, Berlin Open Science Platform (internal project at TU, Open Science, curation, sustainability, research data and repositories)
- DKT – Digital Curation Technologies (2015-2017 at DFKI; funded by BMBF – services for text-based data collections, curation, Linked Open Data)
- European Language Grid (2019-2021 at DFKI; funded by EU, Horizon 2020 – sustainability, platform functionality, RD repositories, international compatibility and interoperability, large, Europe-wide community)
- FDMentor (2017-2019, funded by BMBF; FU, HU, TU, U Potsdam & Viadrina; among other aspects: RDM training and TtT workshops – proposal for succeeding programme 'FDNext' is submitted to DFG LIS funding programme)
- I2M (2017-2020 at HCC, funded by BMBF - Designing human-machine workflows in NLP-based idea augmentation processes)
- IKON (2016-2020 at HCC; funded by BMBF - Interpretability techniques for NLP-based visualization pipeline for research projects)
- META-NET und META-SHARE (2010-2017 at DKFI; EU, FP7 and Horizon 2020 – research data and repositories
- Neonion (seit 2010 at HCC, funded by MPIWG, EXC BWG) – Collaborative human-machine annotation tool for the humanities)
- QURATOR (2018-2021 at DFKI; funded by BMBF – services for text-based data collections, curation, Linked Open Data, development of workflows for curating, sustainability, platform functionality)

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

- We will have an agreement with all known consortia with cross-cutting topics to coordinate and discuss our contributions to NFDI. This includes 2linkNFDI, BRIDGE4NFDI, GeRDI4NFDI, ForumX, NFDI4HPC, NFDI Web, CompeNDI, NFDI4CS4NFDI, RSE4NFDI and Interdisciplinary NFDI.
- With CompeNDI there will be a deeper collaboration because a Train-the-Trainer concept is crucial for NFDI4Language.

- Though their alignment, solutions and in part subject areas differ greatly from NFDI4Language, we expect a close collaboration with NFDI4Culture, TEXT+, NFDI4Collections, NFDI4Memory due to the overlap concerning methodology, types of research data and subject areas.

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.
  - In the fields of research data management and data management in general there are many cross-cutting topics and generic services which are important for most NFDI consortia. Intertwined with the FAIR principles this includes standards, metadata, persistent identifiers, APIs; machine learning, Semantic Web and Linked Open Data, Interoperability; copyrights, licensing.
  - All NFDI consortia have to address knowledge management and exchange as well as expertise and training.
- Please indicate which of these cross-cutting topics your consortium could contribute to and how.
  - NFDI4Language can and will contribute to all topics mentioned above. The FAIR principles and common rules of data management and software development are integral part of network collaboration in our digital age. The technical core of NFDI4Language and the consortia itself are designed as Open Science.