Letter of Intent

Proposal for a National Research Data Infrastructure (anticipated submission in 2020)

Name:

Nationale Forschungsdateninfrastruktur für das World Wide Web

Acronym:

NFDI Web

Applicant institution:

L3S Research Center
Appelstraße 9a
30167 Hanover

Head / Spokesperson:

Prof. Dr. Wolfgang Nejdl
L3S Research Center
Appelstraße 9a
30167 Hanover
nejdl@l3s.de

I Objectives, Work Program, and Research Environment

Many disciplines of computer science study particular aspects of information systems, ranging from scalable storage and analysis via information extraction, retrieval, and visualization to knowledge management and artificial intelligence. For these disciplines the World Wide Web has become the single most important resource. In fact, much of the wisdom that has been baked into today's commercial search engines, recommender systems, question answering systems, conversational agents, etc., was directly extracted from the web and can only be found there. Yet, almost everyone who carries out research on new information systems (or aspects thereof) starts from scratch each time, collecting raw data from the live web, whereas in particular past web data is mostly beyond reach. An exception from this rule are engineers and scientists working at a handful of large Internet corporations who have a ready-to-use copy of the web at their disposal. For them, analyzing the web and its history is an everyday task, whereas in academic realms it remains a major barrier to entry for most any research group.

Similarly, the social sciences and the humanities increasingly incorporate computational methodology into their methodological toolbox. As a medium of its own, the web is meanwhile tightly interwoven with the fabric of today's society and culture so that analyzing the web, as well as the way how people express themselves and communicate with each other via the web has become a key task in these disciplines. Almost every aspect of modern life has been touched by the web; new modes of personal interaction and cultural expression have emerged, old social contracts are questioned and new ones are formed, new societal challenges have arisen—in short: our way of life is and will stay heavily influenced by the web. In this regard, unlike most other digital media, the web is also freely available at scale. Yet, the difficulty of harnessing the web for analysis is paramount and actually constitutes a major limitation to advancing our understanding of today's society and culture.

Key objective of the National Research Data Infrastructure for the World Wide Web (NFDI Web) is to facilitate the utilization and analysis of the web at scale for academic computer science as well as for computational social science and for the digital humanities. Facilitation means empowering the aforementioned communities by providing services of the following and similar kinds:

- provisioning and maintaining a copy of the web and its history
- big data analytics on demand
- APIs and web services for frequently demanded analyses
- extraction, derivation, and provenance of task-specific data
- facilities to train large-scale machine learning models
- visual analytics tools to support specific analysis tasks
- project funding to carry out special-purpose data analyses
- community events to foster collaboration and shared work on the same task
- involvement into steering future infrastructure development

The NFDI Web consortium collaborates closely with the Internet Archive, San Francisco, and it hence can rely on the long lasting experience in building and maintaining large-scale infrastructures for web analytics. Basis of the NFDI Web is the existing Immersive Web Observatory, which hosts a copy of the web and its history, courtesy of the Internet Archive. To illustrate both the value of the data treasure and the need of its analysis, the following list shows a small selection of key questions that computer scientists, social scientists, or humanists may pursue to answer using our infrastructure:

- Who wrote the web?
- Did one-sidedness and bias change over time on the web?

- What digital traces of societal processes are captured by the Web?
- Hot can the monetary value of web platforms such as Wikipedia be measured?
- How can the web serve as a data source for historians?
- What social or technological innovations spread the fastest and why?
- How can knowledge extracted from web archives serve as an enabler for AI?
- Can distant and weak supervision be scaled to the web?

Existing Infrastructure: The Immersive Web Observatory

The Bauhaus-Universität Weimar and the universities of Halle and Leipzig, project and infrastructure funding was acquired to build cluster computers for the purpose of web analytics: As part of the BMBF-funded InnoProfile projects “Intelligent Learning”, “Big Data Analytics”, and “Provenance Analytics”, as well as a DFG large-scale research facility grant for a cluster computer for the “Digital Bauhaus Lab” major research instrumentation (Art. 91b, German Basic Law), we installed the cluster computers Alphaweb (2009), Betaweb (2015), and Gammaweb (2016). This setup allowed for developing and sustaining large-scale experimental web search engines, such as ChatNoir, Netspeak, and Args. Based on the decade-long experience with this infrastructure, another large-scale infrastructure funding from the BMBF was successfully acquired to build the “Immersive Web Observatory”: a web-scale data center. As part of this initiative, the Deltaweb cluster was recently (2018) installed. In addition, to allow for web mining at scale on this combined infrastructure, the Web Archive collection of the Internet Archive was licensed, a web crawl of more than 750 billion web pages and versions thereof collected since 1996. The Web Archive is the only publicly available web crawl that is comparable to Google’s proprietary one.

The NFDI Web is based on the outlined infrastructure: The archived web data are hosted within the “Immersive Web Observatory” at the Digital Bauhaus Lab of the Bauhaus-Universität Weimar. The currently available storage capacity of about 16 PB total suffices to redundantly store a portion of 8 PB of web archive data from the Internet Archive. A large indexing cluster providing search functionality over the NFDI Web’s data will be hosted at the Martin-Luther-Universität Halle-Wittenberg, utilizing hardware with large main memory and fast SSDs. The services and APIs for easy access and analysis of the web data as well as evaluation as a service are hosted and maintained at the Universität Leipzig. Web archival technology is developed at L3S Hanover, and complementary acquisition and standardization of web data is organized at the RWTH Aachen. As a foundation, this data center accommodates the initial load expected during the bootstrapping of the NFDI Web. We plan to extend the data center in the future in order to sustain an increasing adoption and growth of the NFDI Web by the respective research communities. This may include further hardware investments at the aforementioned sites, but also the incorporation of hardware hosted at national and international partner sites forming the basis for a distributed network of data centers providing more resilience, larger compute capacities as well as larger storage.

Community and Collaboration Policy

We expect strong community involvement in building and operating the NFDI Web: The goal of the infrastructure is not to remain a passive data host, but to actively foster the utilization of the data provided. A key component of involving the respective communities into this process will be small and large projects for which the NFDI Web will provide funding for individuals as well as for groups. It is the community members who understand best how raw web data needs to be processed to derive research datasets that are useful to a wider international community tailored to a given task, problem, or question of interest. In this regard, the NFDI Web will provide both the interface to derive data as well as the platform to expose derived data to the community in a standardized way.

Furthermore, a number of prospective NFDI consortia presented at the first NFDI conference are interested in a collaboration. Currently, we are talking to NFDI Text+, to 2linkNFDI, and to KonsortsWD, among others. Given our plans to submit a proposal no earlier than 2020, we have not yet reached detailed collaboration agreements at this time. Throughout the ongoing process of rallying our community and taking into account the state, progress, and plans of the other NFDIs, we envisage to reach agreements by the next submission deadline.

II Cross-Cutting Topics

We identify the following topics and tasks, whcih will pertain to not just our NFDI, but many if not even all NFDIs:

- *Data Derivation, Versioning.* The data that are provisioned by the NFDI Web are raw data to many of its users. I.e., our goal is to serve a copy of the web as it is, whereas users may require the data in a refined, cleaned, or otherwise altered form. At the very least, users will want to sample from the web specific resources that fulfill criteria relevant to solving a given task or answering a given research question. We therefore expect users to create numerous derived datasets, large and small, through some kind of processing which are then published to a wider community. Moreover, these processes will be subject to change, for instance, when errors are identified and fixed, or when after some time has passed, a new version of the extraction process is deployed. It is the goal of NFDI Web to keep track of derived datasets and how they are used by the community, maintaining a registry of the datasets and its versions. We believe that this task will also pertain to a number of other NFDIs, especially those where large datasets are maintained which users will sample from.

- *Data Provenance.* The original sources of a given piece of data will have to be recorded and kept as meta data for every piece of data provided by NFDI Web. Web Archives as they are built today are no homogeneous collections of crawls, but comprise many specifically created crawls that focus on certain aspects like genres, websites, or crawling strategies. In this regard, keeping track of who created a given web crawl for what purposes is important. The Internet Archive maintains a complete record of its currently available collections, however, at NFDI Web, we plan to also create and contribute our own up-to-date crawls. Provenance will also be an important subject at most any other NFDI which maintains data for users.

- *Data Search.* Key to the success of NFDI Web will be to enable its users to quickly and accurately search and retrieve web datasets and subsets of web datasets that are needed for a given research project. As the datasets provided by NFDI Web are massive, building and maintaining a corresponding search and retrieval system will in itself be challenging. This also holds for nearly all other NFDIs, where as the number of datasets and the amount of data provided grows, maintaining an overview of what is there and ensuring the findability of datasets will become a key issue to ensure the FAIR principles. If a given dataset cannot be found by those who need it, it may as well not be archived. Therefore, building and maintaining a cross-NFDI distributed retrieval infrastructure may well be one of the main factors that determine the success of the entire NFDI program.

- *High-Performance Computing and Big Data Analytics.* As datasets grow, they become less mobile. The entire dataset of NFDI Web will very likely be tied to the distributed infrastructure that hosts it and therefore be an immovable property. In this case, the algorithms-to-data paradigm is required,

where the computing facilities are provided to process the data within the infrastructure hosting it. The “Immersive Web Observatory” provides sufficient capacity to process the web data stored. For this purpose, we have built and maintained a cloud infrastructure based on an S3-compatible storage layer as well as a big data analytics stack using well-known big data processing as well as microservice orchestration frameworks. A similar infrastructure will be required by all NFDIs where data are collected at scales that go beyond the typical resources available to their users. This may also pertain to derived data as long as their size are on the same order of magnitude as the original data; otherwise, small derived data may be freely shared.

- *Replicability and Reproducibility.* Where experiments are carried out at infrastructure site by processing data that are too big to be transferred and processed locally, the replicability and reproducibility of experiments is especially important, since users may have not as fine-grained control over the experiment process as would be the case when working entirely locally. Furthermore, a given (derived) dataset may be subject to a computational task that is addressed not only by one community member but many, independently of each other. In that case, the NFDI Web shall provide for fair access to the dataset to all community members as well as fair comparative evaluation procedures. Here, also the reproducibility and replicability of experiments is important. At the same time, a central repository of what analyses already have been conducted can help to avoid running the same pipelines on the same data again and again. NFDI Web plans on providing a processing infrastructure that especially supports experimentation in this way. Again, it is likely that at many other NFDIs, where the data provided are subject to computational exploitation by its users, similar situations will arise.

From the aforementioned cross-cutting topics, there are two for which the consortium of NFDI Web is particularly qualified to contribute to:

- *Data Search.* Most of the NFDI Web (co-)applicants work in information retrieval. Building scalable search engines is part of our everyday work and research. Together, we combine decades of experience in that research field, having published at all major venues of our trade as well as maintaining a number of large-scale search engines that are used worldwide. The field of data search has recently gained some notoriety after the release of Google’s dataset search prototype; it is an active sub-discipline of information retrieval. As part of NFDI Web, we propose to built a distributed search infrastructure (“NFDI Search”) that will serve as an entry point to all datasets provided by all NFDIs. This will require some coordination in terms of meta data formats as well as indexing capabilities, however, this is a small overhead for the capability to search in milliseconds through all of the probably millions of datasets that will be collected throughout the coming years.

- *Replicability and Reproducibility.* Again, members of the NFDI have been working on the reproducibility and replicability of computer science experiments for the past decade. We have built and operate the first working prototype that implements cloud-based evaluation under the evaluation-as-a-service paradigms, called TIRA Integrated Research Architecture. The goal of TIRA is to enable researchers to work on clearly defined tasks that are based on one or more dataset comprising corresponding problem instances. The architecture is freely available. Sharing it with all NFDIs will be straightforward.