— NFDI Web —
Nationale Forschungsdateninfrastruktur für das World Wide Web

Sprecher: Prof. Dr. Wolfgang Nejdl, Forschungszentrum L3S, nejdl@l3s.de

Viele Disziplinen der Informatik befassen sich mit speziellen Aspekten von Informationssystemen, die von skalierbarer Speicherung und Analyse über Informationsextraktion, -abfrage und -visualisierung bis hin zu Wissensmanagement und künstlicher Intelligenz reichen. Für diese Disziplinen ist das World Wide Web zur wichtigsten Datenquelle avanciert. Tatsächlich wurde ein Großteil des Wissens, das in die heutigen kommerziellen Suchmaschinen, Empfehlungssysteme, Question-Answering-Systeme, Konversationsagenten usw. integriert wurde, direkt aus dem Web extrahiert und kann nur dort gefunden werden. Dennoch fangen fast alle Forscherinnen und Forscher, die neue Informationssysteme erforschen, immer wieder bei Null an und sammeln Rohdaten aus dem Live-Web, wobei insbesondere ältere Webdaten meist unerreichbar sind. Eine Ausnahme von dieser Regel bilden die Mitarbeiter, die bei einer Handvoll großer Internetfirmen arbeiten und über eine gebrauchsfertige Kopie des Webs verfügen. Für sie ist die Analyse des Netzes und seiner Geschichte eine alltägliche Aufgabe, während dies im akademischen Umfeld die meisten Forschungsgruppen nach wie vor eine große Eintrittsbarriere darstellt.

In ähnlicher Weise nehmen die Geistes- und Sozialwissenschaften die Methoden der Informatik zunehmend in ihren Werkzeugkasten auf. Als eigenständiges Medium ist das Netz inzwischen eng mit dem Gefüge der heutigen Gesellschaft und Kultur verwoben, so dass die Analyse des Netzes sowie der Art und Weise, wie Menschen sich über das Netz ausdrücken und miteinander kommunizieren, zu einer Schlüsselaufgabe in diesen Disziplinen geworden ist. Anders als die meisten anderen digitalen Medien, auch in seinem Umfang frei verfügbar. Dennoch ist die Schwierigkeit, das Netz für die Analyse nutzbar zu machen, von größter Bedeutung und stellt in der Tat eine große Einschränkung für die Förderung unseres Verständnisses der heutigen Gesellschaft und Kultur dar.

Hauptziel der Nationalen Forschungsdateninfrastruktur für das World Wide Web (NFDI Web) ist es, die Nutzung und Analyse des Webs im industriellen Maßstab für die akademische Informatik sowie für die rechnergestützte Sozialwissenschaft und für die digitalen Geisteswissenschaften durch die Bereitstellung von Dienstleistungen der folgenden und ähnlicher Art zu erleichtern:

- Bereitstellung und Pflege einer Kopie des Webs und seiner Historie
- Big Data Analytics on Demand
- APIs und Webdienste für häufig nachgefragte Analysen
- Extraktion, Ableitung und Provenienz von aufgabenspezifischen Daten
- Training von Deep Learning-Modellen auf Basis von umfangreichen Web-Daten
- visuelle Analysewerkzeuge zur Unterstützung spezifischer Analyseaufgaben
- Projektfinanzierung zur Durchführung von Datenanalysen für spezielle Zwecke
- Events zur Förderung der Zusammenarbeit an gemeinsamen Aufgaben

Das Konsortium der NFDI Web arbeitet eng mit dem Internet Archive, San Francisco, zusammen und kann sich daher auf die langjährige Erfahrung beim Aufbau und der Wartung großer Infrastrukturen für Web-Analytik stützen. Grundlage der NFDI Web ist das bestehende Immersive Web Observatory, das mit freundlicher Genehmigung des Internet Archives eine Kopie des Webs und seiner Historie vorhält. Um sowohl den Wert des Datenschatzes als auch die Notwendigkeit seiner Analyse zu veranschaulichen, zeigt die folgende Liste eine kleine Auswahl von Schlüsselfragen, die Informatiker, Sozial- oder Geisteswissenschaftler mit Hilfe unserer Infrastruktur beantworten können:

- Wer hat das Web geschrieben?

- Haben sich Einseitigkeit und Voreingenommenheit (Bias) im Laufe der Zeit im Web verändert?
- Welche digitalen Spuren gesellschaftlicher Prozesse werden durch das Web erfasst?
- Ist der monetäre Wert von Web-Plattformen wie Wikipedia messbar?
- Wie kann das Web als Datenquelle für Historiker dienen?
- Welche sozialen oder technologischen Innovationen verbreiten sich am schnellsten und warum?
- Wie kann aus Webarchiven extrahiertes Wissen als Enabler für KI dienen?
- Kann entfernte und schwache Überwachung auf das Web skaliert werden?

Von der Bauhaus-Universität Weimar und den Universitäten Halle und Leipzig wurden Projekt- und Infrastrukturmittel zum Aufbau von Cluster-Rechnern für die Zwecke der Web-Analytik eingeworben: Im Rahmen der vom BMBF geförderten InnoProfile-Projekte “Intelligentes Lernen”, “Big Data Analytics” und “Provenance Analytics” sowie einer DFG-Großgeräteförderung für einen Clusterrechner für die Großgeräteausstattung “Digital Bauhaus Lab” (Art. 91b GG) haben wir die Clusterrechner Alphaweb (2009), Betaweb (2015), Gammaweb (2016), Deltaweb (2018) und Epsilonweb (2020) installiert. Die Cluster ermöglichen die Bereitstellung groß angelegter experimenteller Web-Suchmaschinen wie Chat-Noir, Netspeak und Args. Um auf dieser kombinierten Infrastruktur ein Web-Mining in großem Maßstab zu ermöglichen, wurde außerdem das Web-Archiv des Internet Archives lizenziert, ein Web-Crawl von mehr als 800 Milliarden Webseiten und deren Versionen, das seit 1996 aufgebaut wurde. Das Webarchiv ist das einzige öffentlich zugängliche Web-Crawl, das mit dem Googles vergleichbar ist.

Die NFDI Web basiert auf der skizzierten Infrastruktur: Die archivierten Webdaten werden innerhalb des “Immersive Web Observatory” am Digital Bauhaus Lab der Bauhaus-Universität Weimar gehostet. Die derzeit verfügbare Speicherkapazität von insgesamt ca. 16 PB reicht aus, um einen Teil von 8 PB der Webarchivdaten aus dem Internet-Archiv redundant zu speichern. Ein großer Indexierungscluster, der Suchfunktionalität über die Daten der NFDI Web bietet, wird an der Martin-Luther-Universität Halle-Wittenberg unter Verwendung von Hardware mit großem Hauptspeicher und schnellen SSDs gehostet. Die Dienste und APIs für den einfachen Zugriff und die Analyse der Webdaten sowie die Auswertung als Service werden an der Universität Leipzig gehostet und gepflegt. Die Web-Archivierungstechnologie wird am L3S Hannover entwickelt, und die ergänzende Erfassung und Standardisierung von Web-Daten wird an der RWTH Aachen organisiert.

Wir erwarten eine starke Beteiligung der Forschungsgemeinde am Aufbau und Betrieb der NFDI Web: Ziel der Infrastruktur ist es nicht, ein passiver Datenhost zu bleiben, sondern die Nutzung der bereitgestellten Daten aktiv zu fördern. Eine Schlüsselkomponente bei der Einbeziehung der jeweiligen Forschungsgemeinden in diesen Prozess werden kleine und große Projekte sein, für die die NFDI Web sowohl für Einzelpersonen als auch für Gruppen Mittel bereitstellen wird. Es sind die Mitglieder der Forschungsgemeinden, die am besten verstehen, wie Web-Rohdaten verarbeitet werden müssen, um Forschungsdatensätze abzuleiten, die für die internationale Forschungsgemeinde nützlich sind und auf eine bestimmte Aufgabe, ein Problem oder eine Frage von Interesse zugeschnitten sind. In dieser Hinsicht wird die NFDI Web sowohl die Schnittstelle zur Ableitung von Daten als auch die Plattform bieten, um abgeleitete Daten der Gemeinschaft auf standardisierte Weise zugänglich zu machen.

Als Plattform für das Web als Korpus, integriert das NFDI Web sich nahtlos in den Kanon anderer NFDI-Konsortien, die aufgrund ihrer inhaltlichen Ausrichtung auch Web-Daten aufbereiten, unter anderen die Konsortien Text+, NFDI4Culture, NFDI4Language, NFDIxCS, KonsortSWD, und NFDI4Memory. In diesem Zusammenhang versteht sich das NFDI Web gleichsam als Fachkonsortium wie auch als Querschnittskonsortium. Auf Initiative des 2linkNFDI-Konsortiums hat sich das NFDI-Web-Konsortium auch an der Formulierung der Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen beteiligt. Ausgehend von den Anforderungen an die NFDI Web betrachten wir die folgenden drei Querschnittsthemen als von allgemeinem Interesse für die NFDI:

Dataset Search. Das NFDI Web ermöglicht die Erstellung von web-basierten Datensätzen, die für ein bestimmtes Forschungsprojekt benötigt werden. Der Aufbau und die Pflege eines entsprechenden Such- und Retrievalsystems ist für einen FAIRen Zugriff auf die Daten wichtig. Dies gilt für fast alle anderen NFDIen. Wenn ein bestimmter Datensatz von denjenigen, die ihn benötigen, nicht gefunden werden kann, lohnt es sich auch nicht ihn zu archivieren. Der Aufbau und die Aufrechterhaltung einer verteilten NFDI-übergreifenden Abrufinfrastruktur könnte daher durchaus einer der Hauptfaktoren sein, die den Erfolg des gesamten NFDI-Programms bestimmen. Die Partner der NFDI Web forschen und lehren im Information Retrieval. Der Aufbau skalierbarer Suchmaschinen ist Teil unserer täglichen Arbeit. Zusammen vereinen wir jahrzehntelange Erfahrung auf diesem Forschungsgebiet und unterhalten eine Reihe von großen Suchmaschinen, die weltweit genutzt werden. Als Teil der NFDI Web schlagen wir vor, eine verteilte Suchinfrastruktur ("NFDI Search") aufzubauen, die als Einstiegspunkt zu den von den von NFDIen bereitgestellten Datensätzen dienen soll.

High-Performance Computing und Big Data Analytics. Wenn Datensätze wachsen, werden sie weniger mobil. In diesem Fall ist das Algorithmus-zu-Daten-Paradigma erforderlich, bei dem die Kapazitäten zur Verarbeitung der Daten innerhalb der Infrastruktur, die die Daten beherbergt, bereitgestellt werden. Eine Cloud-Infrastruktur wird von allen NFDIen benötigt, bei denen Daten in größerem Umfang erhoben werden. Dies kann sich auch auf abgeleitete Daten beziehen. Wir planen, unser Datenzentrum in der Zukunft zu erweitern, um das Wachstum der NFDI Web und die Angebote an die jeweiligen Forschungsgemeinschaften zu stützen. Dazu können weitere Hardware-Investitionen an den genannten Standorten gehören, aber auch die Einbindung von Hardware, die an nationalen und internationalen Partnerstandorten gehostet wird und die Grundlage für ein verteiltes Netzwerk von Rechenzentren bildet, die mehr Ausfallsicherheit, größere Rechenkapazitäten sowie größeren Speicherplatz bieten.

Reproduzierbarkeit. Wenn Experimente am Standort der Infrastruktur durchgeführt werden, indem Daten verarbeitet werden, die zu groß sind, um sie lokal zu übertragen und zu verarbeiten, ist die Reproduzierbarkeit der Experimente besonders wichtig. Darüber hinaus kann ein bestimmter (abgeleiteter) Datensatz von für viele Mitglieder einer Forschungsgemeinde unabhängig voneinander von Interesse sein. Die NFDI Web wird eine Verarbeitungsinfrastruktur bereitstellen, die das Experimentieren vor Ort unterstützt. Auch hier ist es wahrscheinlich, dass bei vielen anderen NFDIen ähnliche Anforderungen auftreten werden. Die Partner der NFDI Web haben in den vergangenen Jahren Technologien zur Steigerung der Reproduzierbarkeit von Experimenten in der Informatik gearbeitet. Wir betreiben den ersten funktionsfähigen Prototyp, genannt TIRA Integrated Research Architecture, der die Cloud-basierte Evaluierung nach dem Evaluation-as-a-Service-Paradigma implementiert. Das Ziel der TIRA ist es, Forscherinnen und Forscher in die Lage zu versetzen, an klar definierten Aufgaben zu arbeiten, die auf einem oder mehreren Datensätzen mit entsprechenden Probleminstanzen basieren. Eine gemeinsame Nutzung durch alle NFDIen soll ermöglicht werden.

Der Ausbau der vorhanden Infrastruktur zu einer nationalen Forschungsdateninfrastruktur für das Web ist eine strategische Investition in die Zukunft des Forschungsstandorts Deutschland und Europa. Die ständige Verfügbarkeit von Web-Daten in industriellen Größenordnungen wird es der akademischen Forschung ermöglichen, Fragen zu beantworten, die von gesellschaftlicher, geschichtlicher, kultureller und praktischer Relevanz sind sowie die sich daraus ergebenden Technologien und Erkenntnisse frei verfügbar zu machen. Anders als die industrielle Forschung wird dies ein Multiplikator für Innovation sein. Von der Teilnahme an der kommenden NFDI-Konferenz erhoffen wir uns Rege Diskussionen und Austausch mit Vertretern anderer Konsortien sowie der NFDI-Geschäftsstelle darüber, wie diese Vision Realität werden kann.

Vorgesehene Mitglieder des Konsortiums:

Institution:

L3S Research Center
Appelstraße 9a
30167 Hanover

Co-Sprecher:

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

Institution:

Bauhaus-Universität Weimar
Geschwister-Scholl-Straße 8
99423 Weimar

Co-Sprecher:
Prof. Dr. Benno Stein
Web Technology and Information Systems
benno.stein@uni-weimar.de

Institution:

Leipzig University
Ritterstraße 26
04109 Leipzig

Co-Sprecher:
Prof. Dr. Martin Potthast
Text Mining and Retrieval
martin.potthast@uni-leipzig.de

Institution:

Martin-Luther-Universtät Halle-Wittenberg
Universitätsplatz 10
06108 Halle

Co-Sprecher:
Prof. Dr. Matthias Hagen
Big Data Analytics
matthias.hagen@informatik.uni-halle.de

Institution:

RWTH Aachen
Templergraben 55
52062 Aachen

Co-Sprecher:
Prof. Dr. Markus Strohmaier
Computational Social Sciences and Humanities
markus.strohmaier@humtec.rwth-aachen.de