Nationale Forschungsdateninfrastruktur für und mit Computer Science (NFDIxCS)

https://nfdixcs.org / Twitter: @nfdixcs

Sprecher: **Michael Goedicke,** Uni Duisburg-Essen **/** GI**,** mg@uni-due.de
CoSprecherin: Prof. Dr. Ulrike Lucke, Uni Potsdam **/** GI, ulucke@uni-potsdam.de

Forschungsgebiet und Ziele

1 Das Konsortium NFDIxCS möchte (a) die Umsetzung der FAIR Data Principles für Informatik-Forschungsdaten sowie Software-Artefakte fördern, (b) die Zitierbarkeit von Software und Informatik-Daten vereinfachen und damit (c) die Pubikationsprozesse und -kultur sowohl in der Informatik als auch in ihren Anwendungen modernisieren. NFDIxCS bietet ein Forum für die Diskussion über Formate von Informatik-Forschungsdaten, Metadatenformaten und Semantiken. Es wirkt auf allgemein akzeptierte Standards, insbesondere für die nachhaltige Speicherung, das Auffinden und Zurverfügungstellung von Informatik-Forschungsdaten hin.

2 Die Informatik ist dabei eine wichtige Schnittstelle in der Wissenschaft: Zum einen ist sie eine eigene Disziplin, die zunehmend große Datenmengen erzeugt und für die eigene Forschung benötigt. Dies gilt nicht nur für Bereiche wie Big Data oder Künstliche Intelligenz / Maschinelles Lernen, sondern auch für den Betrieb von HPC-Systemen, Rechnerarchitektur etc. Zum anderen spielt die Informatik eine wichtige Rolle in anderen wissenschaftlichen Disziplinen. Zahlreiche Verfahren und Prozesse, die neue Erkenntnisse liefern, wären, ohne die Informatik nicht zu realisieren. Auch hier spielen große und größte Datenmengen eine Rolle, welche wiederum eine großartige Möglichkeit bieten, auch genuine Informatikmethoden weiterzuentwickeln.

3 NFDIxCS möchte die Erfahrungen und das Wissen der Informatik-Community zu Systemarchitekturen, Prozessen, Standards für Interoperabilität, datenorientierte wissenschaftliche Publikations- und Kommunikationssysteme mit allen interessierten Wissenschaftsbereichen bzw. allen in der NFDI vertretenden Konsortien teilen. Diese bidirektionale Zielstellung wird durch das „x“ im Akronym repräsentiert. In diesem Sinne möchte das Konsortium auch weitere Partner gewinnen und auf Augenhöhe kooperieren, um die oben skizzierten Ziele zu erreichen und die zugehörigen Dienste nachhaltig zu betreiben.

Die Forschungsdaten der Informatik und ihre besonderen Anforderungen an FDM

Informatik-Forschung ist in den beiden oben genannten Rollen experimentell, beobachtend, analytisch, interpretierend, gestaltend und modellbildend tätig. Dementsprechend vielfältig sind die Forschungsdaten, die in der Informatik entstehen. Es existieren sowohl stark strukturierte (Beispiel: Messdaten in tabellarischer Form), semi-strukturierte (Beispiel: Logdateien aus dem Betrieb von verteilten Rechnersystemen, Netzwerken und IT-Security aber auch Aussagen in und über Software) als auch unstrukturierte Forschungsdaten (Beispiel: annotierte Bild- oder Textdatenbanken als Basis für Methodenentwicklung im Machine Learning, Daten aus Untersuchungen zu HCI und E-Learning aber auch technische und theoretische Informatik-Daten). Die fachspezifischen Prozesse zur Erzeugung, Verwaltung, Verarbeitung und Veröffentlichung dieser Daten reflektieren diese Vielfalt.

Eine Ende 2019 durchgeführte Umfrage unter den 14 Fachbereichen der Gesellschaft für Informatik e.V. (GI) und die intensive Diskussion für die erste Einreichung zeigt, dass die von

Informatikerinnen und Informatikern produzierten Forschungsdaten neben ihrer Struktur, Profilierung bzgl. Datensicherheit, Privatheit, ethischen Aspekten auch in Umfang stark variieren. Die hieraus resultierenden besonderen Herausforderungen für NFDIxCS liegen darin, dass neben den eigentlichen (Roh-)Daten auch die damit verbundenen – in einzelnen Fällen sehr viele unterschiedliche - Softwareartefakte zur Erzeugung, Vorverarbeitung, Analyse und Darstellung der Daten sowie die Umgebungs- und Ausführungsinformation der jeweiligen Experimente mit abgespeichert werden müssen. Dies muss in einer Form geschehen, in der auch nach vielen Jahren und neuen Versionen verwendeter Betriebssysteme, Programmierumgebungen, Datenbanken etc. die Daten wiederhergestellt und möglichst die in Frage stehenden Experimente wiederholt werden können. Vor allem für die Validität und Qualität von Publikationsprozessen muss ein „Einfrieren“ der Daten und der dazugehörigen Umgebungs- und Ausführungsinformation gewährleistet sein inkl. der verschiedenen. Unter Einbeziehung der Anforderungen aus Heterogenität in Mengengerüst und Struktur und der Forderung nach Einbeziehung lokaler Lösungen entsteht das Bild eines komplexen verteilten Informationssystems, das aus lose gekoppelten erweiterbaren Teilarchitekturen der Daten- und Infrastruktur-Anbieter bestehen wird.
Es besteht daher der Bedarf an einem auf Informatik-Aspekte fokussierten NFDI-Konsortium, da die bisherigen Konsortien die spezifischen Anforderungen, welche Forschungsdaten aus der Informatik mit sich bringen, nicht ausreichend berücksichtigen. Es hat sich auch gezeigt, dass für unterschiedliche Bereiche der Informatik gemeinsame Aspekte für Forschungsdaten abgeleitet werden können, was die Effizienz der eingesetzten Ressourcen stark erhöhen wird.

Forschungsdatenmanagement für NFDIxCS

Das Datenmanagement hat für NFDIxCS drei wichtige Dimensionen. (1) Zum einen ist organisatorisch durch entsprechende Gremien sicher zu stellen, dass geeignete Daten, Meta-Daten und zugehörige Qualitäten u.a. der Ethik, Sicherheit, Privatheit etc. fachspezifisch definiert und sichergestellt werden. Ein übergeordneter Prozess sorgt dafür, dass Sub-Disziplin übergreifende Formate, Standards und Qualitäten abgeglichen und ein hoher Grad an Wiederverwendung stattfindet. (2) Zum anderen wird auf einer technischen Ebene dafür gesorgt, dass eine verteilte auf P2P-Prinzipien basierende Architektur aus Software- und Datenbank-Infrastruktur etabliert wird. Diese basiert auf dem Konzept und Framework eines RDMC (Research Data Management Container), der diese wiederverwendbaren durch Plugins erweiterbare Einheiten für Daten und Software definieren und realisieren. Es werden Arbeitsgruppen aus den Fachgebieten der Informatik etabliert, die die Prinzipien des jeweiligen fachspezifischen Datenmanagement für das jeweilige Fachgebiet definieren und mit den übergeordneten Steuerungsgremien abstimmen. Daraus ergeben sich dann spezifisch konfigurierte RDMC für das jeweilige Informatik-Fachgebiet / Informatik-Sub-Disziplin. (3) Service Provider realisieren das Hosting solcher (vor)konfigurierten RDMCs die sich automatisiert in eine Gesamtstruktur eines Portals einbinden. Dieses Portal realisiert allgemeine Services und APIs und Benutzerschnittstellen zum Wiederauffinden der Forschungsdaten auf der Basis von Meta-Daten effizient und präzise. Außerdem wird durch solche standardisierten Formate und Frameworks die Ausfallsicherheit durch Verteiltheit und Vermeidung von Technologie und Provider-Lockin erreicht.

Datenmanagement, Communities, Standards

NFDIxCS bietet mit seinen Partnern und Teilnehmern die notwendigen Erfahrungen direkt aus dem Bereich FDM – einerseits durch Beteiligung an institutionellem und übergreifendem FDM-Engagement und Beteiligungen in SFBs (INF-Projekte oder thematisch orientierte SFBs wie FONDA) andererseits im Aufbau und Betrieb komplexer verteilter Hardware/ Softwaresysteme durch Beteiligung / Leitung von institutionellen Rechen-/Datazentren. Desweiteren werden die

entsprechenden Konzepte und Technologien bereits in Lehrveranstaltungen aufgenommen, so dass entsprechender Nachwuchs für diesen Bereich ausgebildet wird.
Darüber hinaus bietet die GI als Partner die Strukturen und die organisationale Unterstützung zur Bildung und Betrieb von übergreifenden Steuerungsgremien, Arbeitsgruppen und Managementstrukturen für die Etablierung von Regelstrukturen und Organisation des Outreach zur nationalen Community als auch für die Liason zu internationalen Gruppen und Organisationen. Als größte und wichtigste Fachgesellschaft für Informatik im deutschsprachigen Raum bietet die GI diese Strukturen, um Diskussionen zu bündeln und in einen intensiven Austausch mit den intendierten Nutzendengruppen zu treten. Hier existieren seit langem Gremien, die sich mit FDM beschäftigen, deren Arbeit im Präsidiums-Arbeitskreis NFDIxCS gebündelt wird. In den 14 Fachbereichen und zahlreichen Fachgruppen der GI können Besonderheiten der einzelnen Disziplinen in einem Bottom-Up-Prozess diskutiert und anschließend im PAK NFDIxCS aggregiert werden. Außerdem wird das Thema Kompetenzaufbau in Form von Aus- und Weiterbildung adressiert. Weitere informatiknahe Fachgesellschaften und Vereinigungen und Partner sollen auf Augenhöhe eingebunden werden. In dieser Konstellation bietet NFDIxCS auch die Plattform für die Entwicklung von allgemein akzeptierten Standards für die Speicherung von Metadaten, Daten-Austausch, Speicherung und Archivierung.
Auch soll der Dialog mit fachspezifischen Verlagen, Journalen und Konferenzen über die Publikationspraxis von Daten und Software bzw. Ergänzungen der Publikationsverfahren um Reviews für Daten und Software und die persistente Referenzierung von Daten und Software geführt werden. Die GI ist ein langjährig erfahrener Träger für solche Aushandlungsprozesse.

Kooperation & Infrastrukturen

Die Zusammenarbeit mit weiteren existierenden NFDI-Konsortien u.a. NFDI4ING, NFDI4DS, Mardi ist vereinbart sowie wird NFDIxCS mit neuen entstehenden Konsortien wie z.B. NFDI4Mobiletech suchen. International ist mit der Software Heritage Foundation und der Eclipse Foundation die Zusammenarbeit vereinbart, weitere wie die ReSA[1] sind angestrebt.
Um den Aufbau einer offenen Informatik-Forschungsinfrastruktur (Open Access, Open Science, Open Data) zu beschleunigen, sollen die Rechenzentren u.a. der GDWG, RWTH, Leibnizrechenzentrum und weiteren Infrastruktur bereitstellen. Eine Partnerschaft mit dem DFN ist besprochen, um Identitäten und ein adäquates Rechte- und Rollenkonzept zur Verfügung zu stellen.

Querschnittsthemen und Schnittstellen

NFDIxCS hat sechs relevante Querschnittsthemen identifiziert, zu denen die Informatik wichtige Beiträge liefern kann: Security & Privacy; Usability; Distribution, Liason & Connectivity and Interoperability; Persistent Storage & Scalability; Evolution. Als Mitzeichner der Leipzig-Berlin-Erklärung2 ist NFDIxCS so aufgestellt, dass es Querschnittsthemen für die verschiedenen Aspekte des NFDIxCS-Konsortiums definiert, Lösungen in den fachspezifischen Gremien konfiguiert und in den angebotenen Services dann realisiert werden. Die Lösungen, die übergreifende Bedeutungen haben wie z.B. Anonymisierungs- / Pseudonymisierung- oder Verschlüsselungstechniken (Privatheit/ Datensicherheit) oder Meta-Daten basierte Suchverfahren sollen im Rahmen des NFDI-Vereins, den assoziierten NFDI-Konsortien und weiteren Partnern vor allem auch im internationalen Bereich gefunden und standardisiert werden.

---

[1] https.//www.researchsoft.org

[2] Maik Bierwirth, Frank Oliver Glöckner, Christian Grimm, Sonja Schimmler, Franziska Boehm, Christian Busse, … Heike Neuroth. (2020, June 15). Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung. Zenodo. http://doi.org/10.5281/zenodo.3895209

Vorgesehene Mitglieder des Konsortiums (Co-Sprecherinnen/Co-Sprecher und die weiteren, beteiligten Institutionen):

Prof. Michael Goedicke GI Vizepräsident	Uni Duisburg-Essen
Prof. Ulrike Lucke GI Vizepräsidentin	Uni Potsdam
Daniel Krupka Geschäftsführer GI	GI Berlin
Prof. Dr. Anne Koziolek	KIT Karlsruhe
Prof. Dr. Ralf Reussner	KIT Karlsruhe
Prof. Dr. Hannes Federrath GI Präsident	Uni Hamburg
Prof. Dr. Albrecht Schmidt	LMU München
Prof. Dr. Tobias Nipkow	TU München
Prof. Dr. Martin Schulz	TU München
Prof. Dr. Ramin Yahyapour Direktor	GWDG Göttingen
Prof. Dr.-Ing. André Brinkmann	JGU Mainz
Prof. Dr. Wolfgang Nagel	TU Dresden
Prof. Dr. Nicolas Gauger	TU Kaiserslautern
Prof. Dr. Dr. Thomas Lippert	FZ Jülich
Prof. Dr. Michael Resch	Uni Stuttgart
Prof. Dr. Christian Plessl	Uni Paderborn
Prof. Dr. Matthias Müller	RWTH Aachen
Prof. Dr. Christian Bischof	TU Darmstadt
Prof. Raimund Seidel, Ph.D. Wiss. Direktor	Schloss Dagstuhl - Leibniz Center for Informatics Wadern / Saarbrücken
Dr. Marcel R. Ackermann	Schloss Dagstuhl - Leibniz Center for Informatics Wadern
Dr. Michael Wagner	Schloss Dagstuhl - Leibniz Center for Informatics Wadern
Prof. Dr. Franziska Boehm	FIZ Karlsruhe