

HPC als Bestandteil einer Forschungsinfrastruktur zusammen mit dem Forschungsdatenmanagement gedacht

Annette Strauch-Davey

„HPC ist eine Wissenschaft und Schlüsseltechnologie in unserer heutigen Gesellschaft und viele Fragestellungen könnten ohne HPC nicht gelöst werden.“¹



Abb. 1: Webseite der Gauß-Allianz

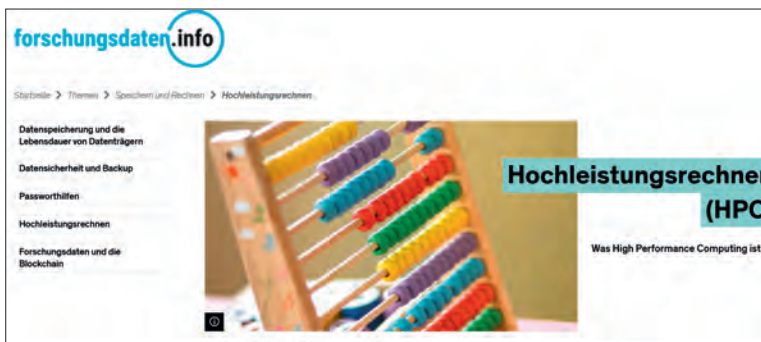


Abb. 2: Zum Hochleistungsrechnen auf forschungsdaten.info

➤ Forschungsfragen, Technik und Fragen unseres täglichen Lebens verlangen nach umfassenden und genauen Modellbildungen für komplexe Systeme, wie beim Wetter zum Beispiel, um Katastrophen vorherzusagen zu können oder zum Klimawandel forschen zu können. Der Zugang zu Hoch- und Höchstleistungsrechnern ist in der heutigen Forschung in allen Bereichen, von den Naturwissenschaften bis zu den Geisteswissenschaften, unverzichtbar. Aspekte sind Komplexität, Zeit und Speicher. Vor diesem Hintergrund ist das Thema im Forschungsdatenmanagement und für Bibliotheken als Partner in der Wissenschaft relevant.

„High-Performance-Computing (HPC) bezieht sich auf die Praxis, die Rechenleistung so zu bündeln, dass viel höhere Leistung als bei herkömmlichen Computern und Servern erzielt wird. HPC oder Supercomputing ist wie alltägliches Computing – nur leistungsfähiger. Es handelt sich dabei um eine Methode, große Datenmengen mit sehr hoher Geschwindigkeit zu verarbeiten, indem man mehrere Computer und Storage-Geräte als zusammenhängende Struktur verwendet. HPC oder Hochleistungsrechnen ermöglicht es, einige der weltweit größten Probleme in Wissenschaft, Technik und Wirtschaft zu untersuchen und Antworten darauf zu finden.“²

Die Plattform forschungsdaten.info hat zum Hochleistungsrechnen im Kontext des Forschungsdatenmanagement eine Seite mit weiterführenden Informationen.³ *„Da das Speichern von Forschungsdaten auf Speicherinfrastruktur immer wichtiger wird, sind diese Informationen auf der zentralen FDM-Informationenplattform nicht mehr fortzudenken.“⁴*

Am 21. April 2021 veranstaltete „forschungsdaten.info“ einen Vortrag mit Diskussion zum Thema Hoch-

¹ https://gauss-allianz.de/de/article/hpc_explained (Zugriff am 25.04.2022)

² <https://www.oracle.com/de/cloud/hpc/what-is-hpc/> (Zugriff am 25.04.2022)

³ <https://www.forschungsdaten.info/themen/speichern-und-rechnen/hochleistungsrechnen/> (Zugriff am 25.04.2022)

⁴ https://www.forschungsdaten.info/kalender-index/kalender-anzeige/2021/4/21/event/39629-forschungsdateninfo-live/tx_cal_phpicalendar/ (Zugriff am 25.04.2022)

leistungsrechnen (HPC) mit den Vortragenden Dr. Jennifer Buchmüller (KIT, SCC) und Brigitte Wellenkamp (KIM Universität Hohenheim), die beide Einblicke in HPC-Systeme, Infrastrukturen und Nutzungsszenarien gaben. Der Hochleistungsrechner HoreKa des KIT wurde vorgestellt.

„HoreKa gehörte bei seiner Inbetriebnahme Mitte 2021 zu den 15 leistungsfähigsten Rechnern Europas und belegte bei der Energieeffizienz sogar Platz 13 im internationalen Supercomputer-Ranking. Er kann eine Rechenleistung von ca. 17 PetaFLOPS erbringen – also etwa 17 Billionen Rechenoperationen in der Sekunde, was der Leistung von mehr als 150.000 Laptops entspricht.“⁵

Es geht beim HPC insgesamt um eine datenintensive, nachhaltige Forschung, um innovative Rechenzentren, spezialisierte Hardware-Technologien sowie komplexe und hochparallele Software. Das Angebot in Deutschland zum HPC ist in drei Ebenen strukturiert:

1. Gauß Centre for Supercomputing
2. Überregionale Rechenzentren und Nationales Hochleistungsrechnen
3. Leistungsfähiges Rechnen in ganz Deutschland.

Die Gauß-Allianz ist ein Netzwerk von 19 Mitgliedern und deren 21 führenden Rechenzentren im Bereich Hochleistungsrechnen in Deutschland. Der Infobrief der Gauß-Allianz⁶ informiert über wichtige Neuigkeiten innerhalb der deutschen HPC-Gemeinschaft (Entwicklungen in Deutschland, neue Systeme, Software oder Projekte). Die Gauß-Allianz⁷ informiert also generell zu HPC in Deutschland – Wer? Was? Wo?⁸ Numerische Simulationen aus Natur- und Ingenieurwissenschaften auf HPC-Systemen weisen eine lange Geschichte auf, ist HPC längst im Mainstream angekommen, so auch für die (Digitalen) Geisteswissenschaften. In diesem Verzeichnis finden Sie aktuelle sowie abgeschlossene Projekte aus dem nationalen Hoch- und Höchstleistungsrechnen: <https://gauss-allianz.de/de/project/all#projekte>

5 <https://www.scc.kit.edu/dienste/horeka.php> (Zugriff am 25.04.2022)

6 https://gauss-allianz.de/de/info_letter (Zugriff am 25.04.2022)

7 <https://gauss-allianz.de/de/hpc-ecosystem> (Zugriff am 25.04.2022)

8 https://gauss-allianz.de/de/article/hpc_explained (Zugriff am 25.04.2022)

9 https://www.bmbf.de/bmbf/de/forschung/digitale-wirtschaft-und-gesellschaft/supercomputing/supercomputing_node.html (Zugriff am 25.04.2022)

10 <https://researchoutreach.org/articles/continuous-flow-strategies-accelerated-materials-discovery/> (Zugriff am 25.04.2022)

11 <https://www.icds.psu.edu/events/data-science-workshop-open-science-research-reproducibility/> (Zugriff am 25.04.2022)

12 <https://www.eventbrite.co.nz/e/jupyter-tipsn-tricks-from-interactive-experiments-to-batch-jobs-and-more-registration-315871549307> (Zugriff am 25.04.2022)

13 <https://www.heibrids.berlin/> (Zugriff am 25.04.2022)

Das BMBF richtet mit dem Programm „Hoch- und Höchstleistungsrechnen für das digitale Zeitalter – Forschung und Investitionen zum High-Performance-Computing“ die Förderung und Finanzierung des High-Performance-Computing in Deutschland strategisch auf das kommende Jahrzehnt aus.

„Das BMBF fördert Ausbau, Betrieb und Vernetzung von Hoch- und Höchstleistungsrechnern, schafft die Voraussetzungen für nachhaltige und leistungsfähige Datentechnologien, -anwendungen und -infrastrukturen und leistet so einen Beitrag zur digitalen und technologischen Souveränität Deutschlands und der Europäischen Union.“⁹

In der Materialforschung wird HPC unterstützend eingesetzt:

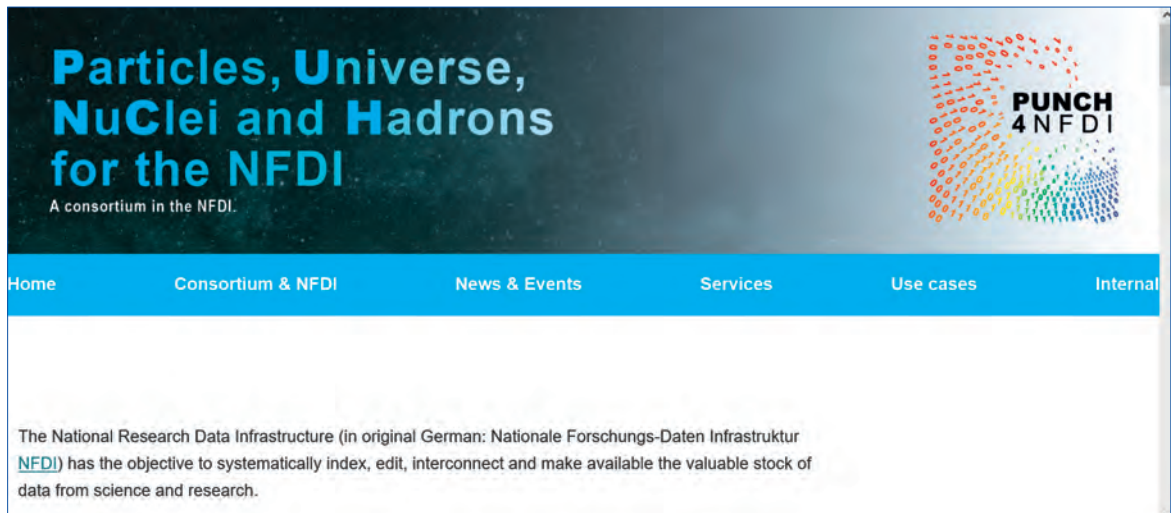
“Continuous flow strategies for accelerated materials discovery.”¹⁰ Im Kontext der Forschung, aber auch zum Forschungsdatenmanagement, in wissenschaftlichen Bibliotheken, werden inzwischen Veranstaltungen, wie Workshops,¹¹ Kurse und Coffee Lectures zu HPC und in dem Zusammenhang Jupyter Notebooks angeboten, um auf die Angebote aufmerksam zu machen.

“Are you using Jupyter to develop your code on NeSI’s HPC platform? Are you a fan of notebooks to quickly try ideas and would like to reuse the same code with more time, more memory, or a larger GPU? Or maybe you want to go beyond default kernels and run your code using a specific NVIDIA container?”¹²

Es gibt eine Kooperation hinsichtlich der Datenwissenschaften, an der Helmholtz-Graduiertenschule HEIBRiDS,¹³ die u.a. HPC beinhalten, wo 11 Institutionen in Berlin-Brandenburg in der Forschung im Bereich Genforschung bis zur Astronomie zusammenarbeiten.

„Die Datenwissenschaften werden für die Forschung von Jahr zu Jahr wichtiger,“ erläutert Uwe Ohler, Bioinformatiker am Max-Delbrück-Centrum für Molekulare

Abb. 3:
Einstiegsseite
zu den Web-
seiten von
PUNCH4NFDI



Medizin in Berlin (MDC) und einer der Initiatoren der School. So liefern moderne Bildgebungsverfahren Aufnahmen mit enormer Auflösung – entsprechend rasant wächst die Datenflut, die es zu speichern und zu verarbeiten gilt. Auch Datensicherheit und statistische Verfahren gewinnen konstant an Bedeutung.”

Das NFDI-Konsortium PUNCH4NFDI beschäftigt sich mit der Teilchen-, Astro-, Astroteilchen-, Hadronen- und Kernphysik mit dem Ziel die Anforderungen der beteiligten Bereiche hinsichtlich der Daten im Exabyte-Bereich zu identifizieren.

“The National Research Data Infrastructure (in original German: Nationale Forschungsdaten Infrastruktur NFDI) has the objective to systematically index, edit, interconnect and make available the valuable stock of data from science and research.”¹⁴

Die PUNCH4NFDI Task Area „Education, Training, Outreach, and Citizen Science“ befasst sich außerdem mit der Aus- und Weiterbildung von Experten und Studierenden zum Forschungsdatenmanagement sowie der Öffentlichkeitsarbeit:

“Data-intensive theoretical computations and simulations, often carried out at high-performance computer centres, complement this picture. All in all, a very diverse array of facilities of all sizes gather scientific data that are exploited by the PUNCH community. The data volumes produced by these facilities grow faster than the IT innovation cycle (Moore’s law). The necessary reduction of data (data loss, irreversibility challenge)

as well as the need for cross-experiment, cross-community, cross-science access are the main drivers behind the PUNCH4NFDI efforts.”

NFDI4Cat¹⁵ wäre als weiteres NFDI-Konsortium zu nennen, das sich auf den Bereich der Katalyse im Zusammenhang mit HPC spezialisiert. Hier ist das High Performance Computing Center Stuttgart (HLRS)¹⁶ mit involviert und aktiv tätig. Interessante und wichtige Beispiele zur Orientierung zum FDM und HPC innerhalb des NFDI4Ing-Konsortiums¹⁷ beinhaltet der Beitrag „Forschungsdatenmanagement für Daten aus High Performance Measurement and Computing an der TU München – Best-Practice Beispiele und Anwendungen“.¹⁸ Der Workshop „Forschungsdatenmanagement für Daten aus High Performance Computing (Tier 1) – Best-Practice-Beispiele und Anwendungen“ (Lehrstuhl für Aerodynamik und Strömungsmechanik, TUM) fand am 06. April 2022 zu folgenden Inhalten statt:

- Forschungsdatenmanagement: Motivation, Konzepte, Beispiele
- Nationale Forschungsdateninfrastruktur NFDI und NFDI4Ing
- Ontologien und Repositorien
- Besonderheiten für HPC-Daten und Lösungsansätze
- Tools für (große) Forschungsdaten und Best-Practice von HLRS, JSC, LRZ

Der genannte Workshop richtete sich an Anwender an den (nationalen) Tier 1 Rechenzentren, Forschern

¹⁴ <https://www.punch4nfdi.de/> (Zugriff am 25.04.2022)

¹⁵ <https://nfdi4cat.org/> (Zugriff am 25.04.2022)

¹⁶ <https://www.uni-stuttgart.de/en/university/news/showcase/hlrs/> (Zugriff am 25.04.2022)

¹⁷ <https://nfdi4ing.de/> (Zugriff am 25.04.2022)

¹⁸ Benjamin Farnbacher, Nils Hoppe, Katja Kessler, Stephan Peinkofer, Stephan Hachinger, & Christian Stemmer. (2021, July 22). Forschungsdatenmanagement für Daten aus High Performance Measurement and Computing an der TU München – Best-Practice Beispiele und Anwendungen. Zenodo. <https://doi.org/10.5281/zenodo.5139408> (Zugriff am 25.04.2022)

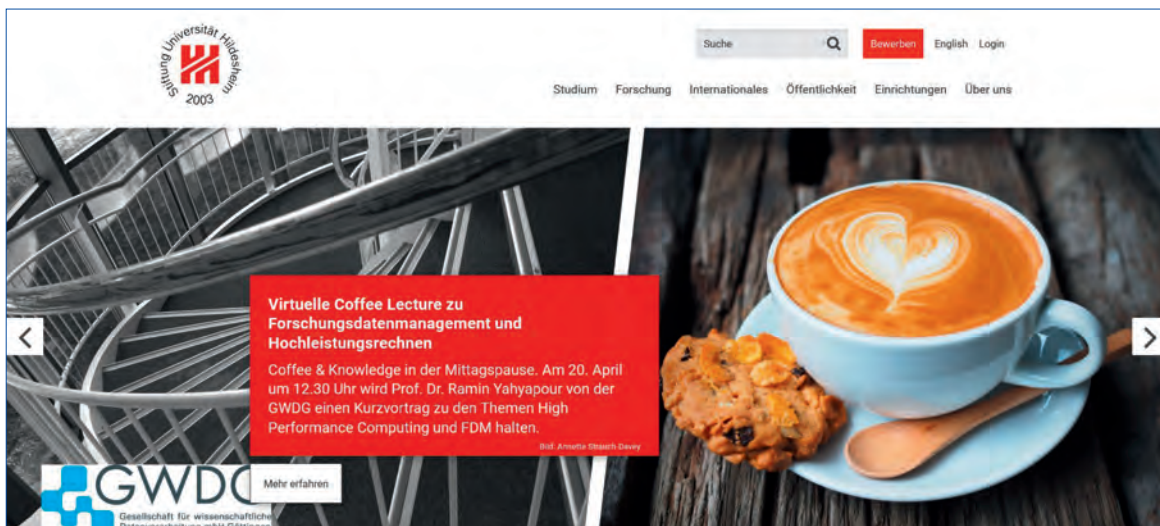


Abb.4: Ankündigung der Coffee Lecture zum High Performance Computing auf der Webseite der Universität Hildesheim

de mit großen, immobilen Datensätzen, Personen im Forschungsdatenmanagement und alle weiteren Interessierten (Zielgruppen). Am 20. April 2022 hielt Prof. Dr. Ing Ramin Yahyapour von der GWDC in Göttingen im Rahmen einer Coffee Lecture zum FDM der UB Hildesheim einen Vortrag zum Thema HPC und Forschungsdatenmanagement, bei dem sich 42 Teilnehmende im BBB-Raum zusammengefunden hatten. Prof. Yahyapour erklärte, dass die GWDC eine gemeinsame gemeinnützige Einrichtung von Universität Göttingen und der Max-Planck-Gesellschaft mit folgenden, diversen überregionalen Aufgaben in folgenden Bereichen (Selektion) ist:

- Nationales Hochleistungsrechenzentrum
- Höchstleistungsrechner (HLRN) – Betreiber für die 7 norddt. Bundesländer
- Nationales DLR HPC Zentrum
- Datenzentrum in 3 NFDI Konsortien
- Host für DARIAH EU, Dt. National Bibliothek, GFBio,
- OER OpenEduHub
- Cloud Betreiber

Fazit aus der Coffee Lecture:

„HPC ist Bestandteil der Forschungsinfrastruktur zusammen mit Forschungsdatenmanagement zu verstehen. HPC erzeugt Daten, unterstützt aber auch die spätere Nutzung jenseits der reinen Sicherung und Bereitstellung von Daten.“ (Prof. Dr. Ing. Ramin Yahyapour in der Coffee Lecture der UB Hildesheim am 20. April 2022).

Die GWDC hat ca. 150.000 Nutzer, drei Rechenzentren, 12.000 Server und mehr. Es gibt eine kritische Infrastruktur für sensible Daten. Motivation für HPC aus der Sicht von Prof. Ramin Yahyapour ist, dass vie-

le wissenschaftliche Probleme durch Modellierung und Simulation behandelt werden. Dieses erlaubt die Analyse von Varianten und Hypothesen mit geringem Aufwand bzgl. Laborexperimenten. Eine Skalierung auf größere Rechnersysteme wird benötigt. „Heute ist die Nutzung von HPC zunehmend Mainstream“, so Yahyapour bei der Coffee Lecture! Außerdem: „Forschungsinfrastruktur ist nicht nur digitale Infrastruktur!“ Die Teilnehmenden der Coffee Lecture erfuhren mehr zum Zugang zu HPC-Ressourcen in Deutschland, zur Einrichtung von Nationalen Hochleistungsrechnern (NHR) in Deutschland. Der norddeutsche Supercomputer HLRN-IV Emmy wurde erklärt¹⁹ und dass HPC oft mit FDM gekoppelt wird (Open Access, Open Data, Open Science). Forschungsdaten lassen sich übertragen, kopieren, spiegeln, verändern, was maßgeblich vom Datenvolumen abhängt. Lokale Strukturen sind wichtig wie Beratungen, leichtgewichtiger Zugang und Bedarfsermittlungen. |



Annette Strauch-Davey
 Forschungsdatenmanagement (FDM)
 Research Data Management (RDM)
 Stiftung UNIVERSITÄT Hildesheim
 Universitätsbibliothek(UB)
 straucha@uni-hildesheim.de

¹⁹ <https://www.hlrn.de/supercomputer/hlrn-iv-system/> (Zugriff am 25.4.2022)