



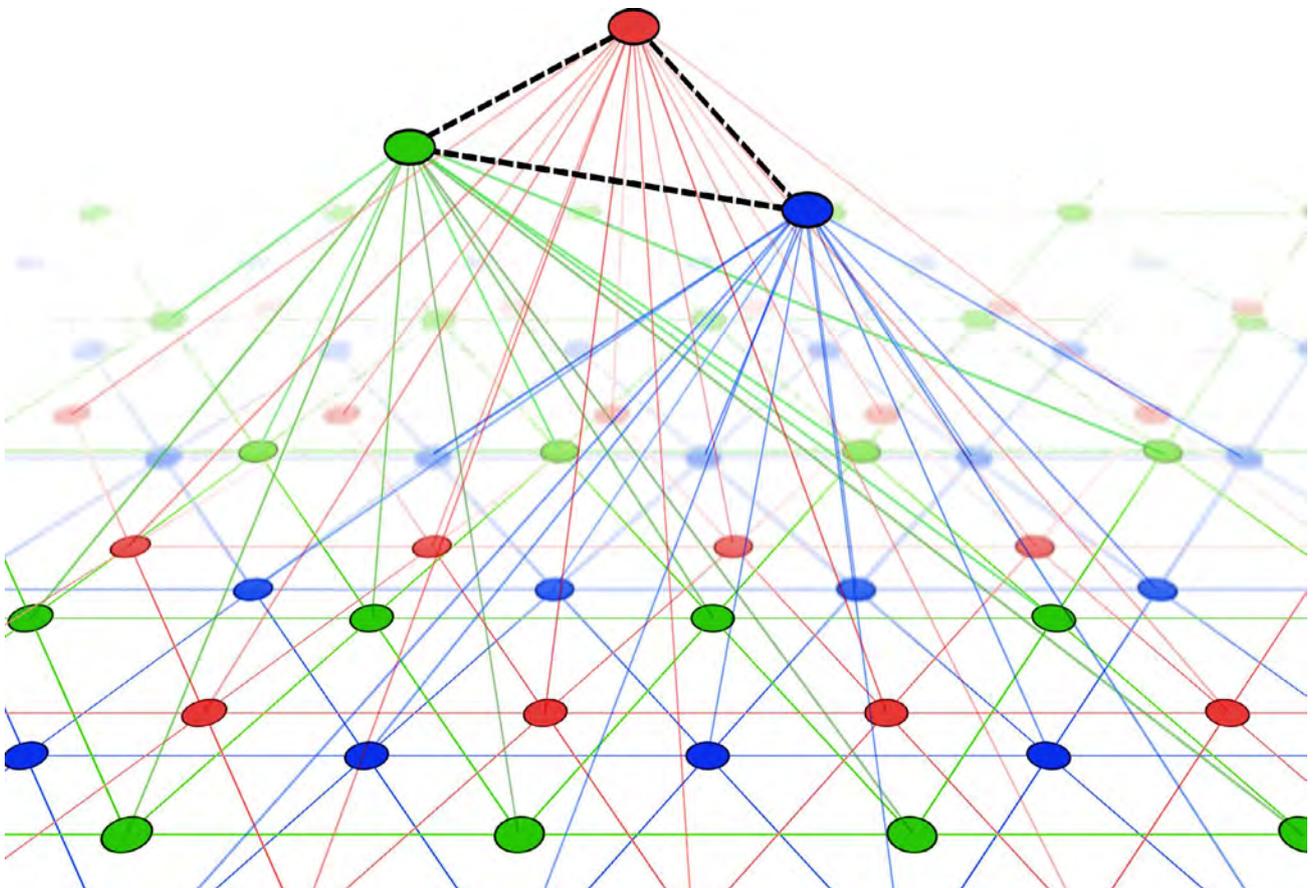
Organ der Gesellschaft für Informatik e.V.  
und mit ihr assoziierter Organisationen

Band 41 • Heft 6 • Dezember 2018



# Informatik Spektrum

E-Science



# APPLICATION ENGINEER (M/W)

## 3RD LEVEL SUPPORT UND WEITERENTWICKLUNG



d-fine ist ein führendes europäisches Beratungsunternehmen mit Standorten in Berlin, Düsseldorf, Frankfurt, London, München, Wien und Zürich. Mit über 800 hochqualifizierten Experten führen wir anspruchsvolle Projekte für die Finanzwelt, die Industrie, die Energiewirtschaft, den Gesundheitssektor oder die öffentliche Hand durch. Unsere mehr als 200 Kunden umfassen internationale Großkonzerne genauso wie Mittelstand und Start-Ups. Die Projektaufträge decken das gesamte Spektrum der Strategieberatung, der Fachberatung und der Technologieberatung ab. Ein starker Fokus liegt dabei auf Projekten mit analytischen und technologischen Herausforderungen. Als wachstumsstarkes Unternehmen mit hohem Qualitätsanspruch braucht unser Team am Standort Frankfurt Verstärkung. Der Einsatz erfolgt häufig auch auf Projekten direkt beim Kunden vor Ort (außerhalb des Rhein-Main Gebiets und teilweise im Ausland).

Wir suchen daher in Festanstellung und in Vollzeit einen **Application Engineer (m/w)**.

### Aufgaben

- Neu- und Weiterentwicklung sowie 3rd Level Support von Spezialanwendungen, die im Finanzumfeld in Deutschland und anderen Ländern eingesetzt werden

- Es handelt sich dabei um technisch und inhaltlich anspruchsvolle Eigenentwicklungen von d-fine unter Nutzung verschiedener Stacks wie ASP.NET oder moderner Big Data Frameworks
- Programmierarbeiten erfolgen in Teams mit fachlichen Experten – etwa für Machine Learning Algorithmen – von d-fine unter Nutzung agiler Entwicklungsmethoden
- Beratung von Kunden bei IT-Fragestellungen, Fokus moderner Architekturen

### Anforderungen

- Spaß an IT und hohe Lernbereitschaft zum weiteren Ausbau der eigenen technischen Kenntnisse
- Studium der (Wirtschafts-)Mathematik, Natur-/Ingenieurwissenschaften oder (Wirtschafts-)Informatik, bei passender praktischer Erfahrung auch andere Studienrichtung möglich
- Sehr guter Universitätsabschluss (Master, Diplom oder ggf. Bachelor) oder exzellenter FH-Abschluss (Master, Diplom)
- Praktische Erfahrung in objektorientierter Programmierung (Präferenz: C#, Java), gerne im Webumfeld und inkl. Versions- und Testmanagement
- Grundkenntnisse relationaler Datenbanken (z.B.: SQL Server oder Oracle) und Big Data Technologien (Hadoop, Spark, Kafka etc.) sind von Vorteil

- Sehr gute Deutsch- und gute Englischkenntnisse
- Sehr gute Kommunikationsfähigkeiten, Teamgeist und die Fähigkeit, komplexe Sachverhalte verständlich zu formulieren

### Wir bieten

- Intensive Einführungsschulungen in Ihr zukünftiges Aufgabenfeld
- Abwechslungsreiche und interessante Tätigkeit, angenehmes Betriebsklima und kurze Entscheidungswege
- Attraktive Sozialleistungen (30 Tage Jahresurlaub, Unfallversicherung, Zuschuss zu einem bundesweit vertretenen Fitness-Anbieter etc.)
- Individuelle Weiterbildungsmöglichkeiten, tolle Mitarbeitererevents und vieles mehr

Wenn Sie in einem Team hoch begabter und hoch motivierter Kollegen mitarbeiten wollen, große individuelle Freiräume, viel Eigenverantwortung sowie hervorragende Entwicklungsperspektiven suchen, freuen wir uns auf Sie.

### Sie sind interessiert?

Dann freuen wir uns auf Ihre Bewerbung unter Angabe der Kennziffer IS2018AE an: [careers@d-fine.de](mailto:careers@d-fine.de)

Bei Fragen können Sie sich gerne an Frau Marina Büttner (069 - 90 737 555) wenden.

**d-fine. analytisch. technologisch. quantitativ.**

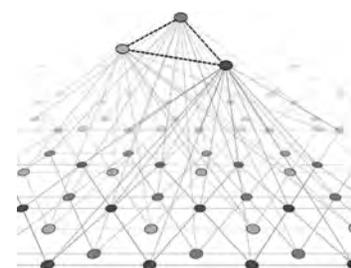
# Informatik Spektrum

Band 41 | Heft 6 | Dezember 2018

Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen

- EDITORIAL**  
T. Eymann, M. Goedicke, U. Lucke  
**357 E-Science als Herausforderung für die Informatik**
- HAUPTBEITRÄGE**  
M. von der Heyde, A. Hartmann, G. Auth, C. Erfurth  
**359 Zur disruptiven Digitalisierung von Hochschulforschung**
- R. Yahyapour  
**369 E-Science Infrastrukturen**
- L. Bernard, S. Mäs  
**376 e-science in den Geo- und Umweltwissenschaften**
- T. Kraft  
**385 Hybride Edition und analoges Erbe**
- O. Eriksson, E. Laure, E. Lindahl, D. Henningson, A. Ynnerman  
**398 e-Science in Scandinavia**
- W. Hazeleger, T. Bakker, R. van Nieuwpoort  
**405 eScience development and experiences in The Netherlands**
- N. Chue Hong  
**414 To achieve the goals of e-Science, we must change research culture globally**
- M. Katerbow, M. Royeck, A. Raabe  
**421 DFG-Förderung und der digitale Wandel in den Wissenschaften**
- Gesellschaft für Informatik e. V. (GI)  
**429 Informatik als Treiber für E-Science**
- V. Markl  
**433 Eine nationale Daten- und Analyseinfrastruktur als Grundlage digitaler Souveränität**
- AKTUELLES SCHLAGWORT**  
R. Fischer  
**440 Vektorarchitektur**
- FORUM**  
**445** Gewissensbits – wie würden Sie urteilen?/Deep Learning und Rechtsrisiken/  
Die Individualisierung des Nutzen/Rezension/Zum Titelbild
- 453 Mitteilungen der Gesellschaft für Informatik 254. Folge**  
Aus Vorstand und Präsidium/Personalien/Aus den GI-Gliederungen/  
Tagungsankündigungen/Aus den assoziierten Gesellschaften/  
Bundesweit Informatiknachwuchs fördern/GI-Veranstaltungskalender

451



Simulation von Gitterzellen

# Informatik Spektrum

Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen

Hauptaufgabe dieser Zeitschrift ist die Weiterbildung aller Informatiker durch Veröffentlichung aktueller, praktisch verwertbarer Informationen über technische und wissenschaftliche Fortschritte aus allen Bereichen der Informatik und ihrer Anwendungen. Dies soll erreicht werden durch Veröffentlichung von Übersichtsartikeln und einführenden Darstellungen sowie Berichten über Projekte und Fallstudien, die zukünftige Trends aufzeigen.

Es sollen damit unter anderem jene Leser angesprochen werden, die sich in neue Sachgebiete der Informatik einarbeiten, sich weiterbilden, sich einen Überblick verschaffen wollen, denen aber das Studium der Originalliteratur zu zeitraubend oder die Beschaffung solcher Veröffentlichungen nicht möglich ist. Damit kommt als Leser nicht nur der ausgebildete Informatikspezialist in Betracht, sondern vor allem der Praktiker, der aus seiner Tagesarbeit heraus Anschluss an die wissenschaftliche Entwicklung der Informatik sucht, aber auch der Studierende an einer Fachhochschule oder Universität, der sich Einblick in Aufgaben und Probleme der Praxis verschaffen möchte.

Durch Auswahl der Autoren und der Themen sowie durch Einflussnahme auf Inhalt und Darstellung – die Beiträge werden von mehreren Herausgebern referiert – soll erreicht werden, dass möglichst jeder Beitrag dem größten Teil der Leser verständlich und lesenswert erscheint. So soll diese Zeitschrift das gesamte Spektrum der Informatik umfassen, aber nicht in getrennte Sparten mit verschiedenen Leserkreisen zerfallen. Da die Informatik eine sich auch weiterhin stark entwickelnde anwendungsorientierte Wissenschaft ist, die ihre eigenen wissenschaftlichen und theoretischen Grundlagen zu einem großen Teil selbst entwickeln muss, will die Zeitschrift sich an den Problemen der Praxis orientieren, ohne die Aufgabe zu vergessen, ein solides wissenschaftliches Fundament zu erarbeiten. Zur Anwendungsorientierung gehört auch die Beschäftigung mit den Problemen der Auswirkung der Informatikwendungen auf den Einzelnen, den Staat und die Gesellschaft sowie mit Fragen der Informatik-Berufe einschließlich der Ausbildungsrichtlinien und der Bedarfsschätzungen.

## Urheberrecht

Mit der Annahme eines Beitrags überträgt der Autor Springer (bzw. dem Eigentümer der Zeitschrift, sofern Springer nicht selbst Eigentümer ist) das ausschließliche Recht zur Vervielfältigung durch Druck, Nachdruck und beliebige sonstige Verfahren das Recht zur Übersetzung für alle Sprachen und Länder.

Die Zeitschrift sowie alle in ihr enthaltenen einzelnen Beiträge und Abbildungen sind urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen schriftlichen Zustimmung des Eigentümers. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen. Jeder Autor, der Deutscher ist oder ständig in der Bundesrepublik Deutschland lebt oder Bürger Österreichs, der Schweiz oder eines Staates der Europäischen

Gemeinschaft ist, kann unter bestimmten Voraussetzungen an der Ausschüttung der Bibliotheks- und Fotokopiertantiemen teilnehmen. Nähere Einzelheiten können direkt von der Verwertungsgesellschaft WORD, Abteilung Wissenschaft, Goethestr. 49, 80336 München, eingeholt werden.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in dieser Zeitschrift berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürfen.

## Vertrieb, Abonnement, Versand

Papierausgabe: ISSN 0170-6012  
elektronische Ausgabe: ISSN 1432-122X  
Erscheinungsweise: zweimonatlich

Den Bezugspreis können Sie beim Customer Service erfragen: [customerservice@springer.com](mailto:customerservice@springer.com) Die Lieferung der Zeitschrift läuft weiter, wenn sie nicht bis zum 30.9. eines Jahres abbestellt wird. Mitglieder der Gesellschaft für Informatik und der Schweizer Informatiker Gesellschaft erhalten die Zeitschrift im Rahmen ihrer Mitgliedschaft.

Bestellungen oder Rückfragen nimmt jede Buchhandlung oder der Verlag entgegen. Springer, Kundenservice Zeitschriften, Tiergartenstr. 15, 69121 Heidelberg, Germany Tel. +49-6221-345-0, Fax: +49-6221-345-4229, e-mail: [customerservice@springer.com](mailto:customerservice@springer.com) Geschäftszeiten: Montag bis Freitag 8–20 h.

Bei Adressänderungen muss neben dem Titel der Zeitschrift die neue und die alte Adresse angegeben werden. Adressänderungen sollten mindestens 6 Wochen vor Gültigkeit gemeldet werden. **Hinweis gemäß § 4 Abs. 3 der Postdienst-Datenschutzverordnung:** Bei Anschriftenänderung des Beziehers kann die Deutsche Post AG dem Verlag die neue Anschrift auch dann mitteilen, wenn kein Nachsendeauftrag gestellt ist. Hiergegen kann der Bezieher innerhalb von 14 Tagen nach Erscheinen dieses Heftes bei unserer Abonnementsbetreuung widersprechen.

## Elektronische Version

[springerlink.com](http://springerlink.com)

## Hinweise für Autoren

<http://springer.com/journal/00287>

## Hauptausgeber

Prof. Dr. Dr. h. c. mult. Wilfried Brauer (1978–1998)  
Prof. Dr. Arndt Bode,  
Technische Universität München (seit 1999)

## Herausgeber

Prof. Dr. S. Albers, TU München  
Dr. B. Bartsch-Spörl, BSR Consulting, München  
Prof. A. Bernstein, Ph. D., Universität Zürich  
Prof. Dr. T. Braun, Universität Bern  
Prof. Dr. O. Deussen, Universität Konstanz  
Prof. Dr. G. Dueck, Neckargemünd  
Prof. Dr. G. Goos, KIT Karlsruhe  
Prof. O. Günther, Ph. D., Universität Potsdam  
Prof. Dr. W. Hesse, Universität Marburg

Dr. Agnes Koschmider, KIT Karlsruhe  
Dr.-Ing. C. Leng, Google  
Prof. Dr. P. Liggesmeyer,  
Fraunhofer IESE, Kaiserslautern  
Prof. Dr. T. Ludwig, Deutsches  
Klimarechenzentrum GmbH Hamburg  
Prof. Dr. F. Mattern, ETH Zürich  
Prof. Dr. Dr. h. c. H. C. Mayr,  
Universität Klagenfurt  
Prof. Dr. K.-R. Müller, TU Berlin  
Prof. Dr. W. Nagel, TU Dresden  
Prof. Dr. J. Nievergelt, ETH Zürich  
Prof. Dr. E. Portmann, Universität Fribourg  
Prof. Dr. F. Puppe, Universität Würzburg  
Prof. Dr. R.H. Reussner, Universität Karlsruhe  
Prof. Dr. S. Rinderle-Ma, Universität Wien  
Prof. Dr. O. Spaniol, RWTH Aachen  
Prof. Dr. D. Steinbauer, Wiesbaden  
Dr. D. Taubner, msg systems ag, München  
Prof. Dr. Herbert Weber, TU Berlin  
Prof. Dr. Katharina Zweig, TU Kaiserslautern

## Impressum

### Verlag:

Springer, Tiergartenstraße 17,  
69121 Heidelberg

### Redaktion:

Peter Pagel, Sybille Thelen  
Tel.: +49 611 787 8329  
e-mail: [Peter.Pagel@springer.com](mailto:Peter.Pagel@springer.com)

### Herstellung:

Philipp Kammerer,  
e-mail: [Philipp.Kammerer@springer.com](mailto:Philipp.Kammerer@springer.com)

### Redaktion GI-Mitteilungen:

Cornelia Winter  
Gesellschaft für Informatik e.V. (GI)  
Wissenschaftszentrum,  
Ahrstraße 45, D-53175 Bonn,  
Tel.: +49 228-302-145, Fax: +49 228-302-167,  
Internet: <http://www.gi.de>,  
e-mail: [gs@gi.de](mailto:gs@gi.de)

### Wissenschaftliche Kommunikation:

Anzeigen: Eva Hanenberg  
Abraham-Lincoln-Straße 46  
65189 Wiesbaden  
Tel.: +49 (0)611/78 78-226  
Fax: +49 (0)611/78 78-430  
[eva.hanenberg@springer.com](mailto:eva.hanenberg@springer.com)

### Satz:

le-tex publishing services GmbH, Leipzig

### Druck:

Printforce,  
The Netherlands

[springer.com](http://springer.com)

Eigentümer und Copyright  
© Springer-Verlag GmbH Deutschland,  
ein Teil von Springer Nature, 2019



**Torsten Eymann**  
Universität Bayreuth



**Michael Goedicke**  
Universität Duisburg-Essen



**Ulrike Lucke**  
Universität Potsdam

## E-Science als Herausforderung für die Informatik

Dieses Heft des Informatik Spektrum beschäftigt sich mit dem Thema E-Science. Von diesem Begriff gibt es vermutlich ebenso viele Schreibweisen wie Definitionen. E-Science beschreibt in unseren Augen in einer weit gefassten Definition die Nutzung von Informatikmethoden und Technologien für Zwecke der wissenschaftlichen Erkenntnisgewinnung. Dies schließt High-Performance Computing und Datenanalyseverfahren ebenso ein wie Forschungsdatenmanagement. In der Definition von Bohle, die wir im Positionspapier der GI nutzen, wird E-Science beschrieben als „die Anwendung von Informations- und Kommunikationstechnologie (ICT) in der modernen Wissenschaft und umfasst ein Spektrum von der Vorbereitung und Durchführung von Untersuchungen, über die Datenerfassung und Verbreitung von Ergebnissen bis hin zur Langzeitspeicherung und dem Zugang zu relevanten Materialien und Artefakten“. Ähnliche und andere Definitionen werden in den Einzelbeiträgen nochmals ausgeführt.

Man könnte also einwenden, dass es „nur“ um die Anwendung dieser Verfahren, z. B. in den Naturwissenschaften oder Geisteswissenschaften (Digital Humanities), ginge. Dieser Wahrnehmung folgend, waren in den letzten Jahren die wesentlichen Treiber von E-Science eher die Anwender, z. B. die Geowissenschaften oder die Physik, oder die Betreiberinnen und Betreiber von E-Research-Infrastrukturen, häufig die Bibliotheken von Hochschulen und Akademien.

In unseren Augen trifft das nicht die Essenz des Begriffs, denn E-Science beschreibt eine Vision des zukünftigen wissenschaftlichen Arbeitens und die mittel- bis langfristige Entwicklung der wissenschaftlichen Praxis dorthin. Wir möchten dieses Heft daher nutzen, den Informatikerinnen und Informatikern näherzubringen, warum es sich beim „E-“ in „E-Science“ nicht einfach um die Übertragung bestehender Anwendungen in ein digitales Umfeld handelt, sondern sich hinter dem Begriff neuartige und spannende Fragestellungen auch für die Informatik in Wissenschaft und Praxis verbergen. Die in diesem Heft vorgestellten Beiträge zeigen sowohl für die Wissenschaftslandschaft in Deutschland als auch im Vergleich für andere europäische Länder, welche Dynamik sich hinter dem Begriff verbirgt und warum es sinnvoll ist, sich gerade als Informatikerin oder Informatiker damit zu beschäftigen.

Den Anfang macht der Beitrag „Thesen zur disruptiven Digitalisierung von Hochschulforschung – Faktoren der Skalierung und ein Zukunftsszenario“, der in einer früheren Fassung auf der GI-Jahrestagung 2017 präsentiert wurde. Von der Heyde et al. binden damit das Thema E-Science in einen größeren Zusammenhang der Digitalisierung der Hochschulen allgemein ein.

Der Beitrag von Ramin Yahyapour zeigt den derzeitigen Stand der E-Science-Infrastruktur in Deutschland auf. Yahyapour ist Mitglied des Rates für Informationsinfrastrukturen (RfI) der Gemeinsamen Wissenschaftskonferenz, in der Wissenschafts- und Finanzminister aus Bund und Ländern über Leitlinien der Wissenschaftsförderung diskutieren. Der Rat hat sich bereits 2016 für die Gründung einer Nationalen Forschungsdateninfrastruktur (NFDI) ausgesprochen, die der Wissenschaft eine Grundversorgung an Speichermöglichkeiten und Dienstleistungen für die Nachnutzung der Daten anbieten soll. Auch für die europäische Wissenschaftslandschaft gibt es ähnliche Aktivitäten, wie Yahyapour darstellt.

Der Artikel von Bernard und Mäs über E-Science in den Geo- und Umweltwissenschaften zeigt exemplarisch, welche Anwendungsfragen an die Informatik und hier speziell die Geoinformatik herangetragen werden. Beispiels-

<sup>1</sup> Bohle S (2013) „What is E-science and How Should it Be Managed?“ Nature.com, Spektrum der Wissenschaft (Scientific American), [http://www.scilogs.com/scientific\\_and\\_medical\\_libraries/what-is-e-science-and-how-should-it-be-managed/](http://www.scilogs.com/scientific_and_medical_libraries/what-is-e-science-and-how-should-it-be-managed/).

weise wird gerade hier mit großen Datenmengen gearbeitet, insofern ergibt sich ein reiches Feld für alle Fragen rund um die Speicherung und Verarbeitung von Big Data. Die Unmöglichkeit, Ergebnisse dieser Verarbeitung manuell zu evaluieren, führt zu Fragen der Reproduzierbarkeit wissenschaftlicher Erkenntnisse.

Kraft stellt in seinem Beitrag „Hybride Edition und analoges Erbe“ den Stand der Technologie in den Digital Humanities vor. Am Beispiel der Digitalisierung einer Handschrift von Alexander von Humboldt wird nicht nur der technologische Ansatz beschrieben. Es wird auch deutlich, wie viel menschliches Wissen derzeit noch für die Generierung qualitativ hochwertiger Forschungsdaten benötigt wird und dass wir in der Digitalisierung an dieser Stelle noch einigen Weg vor uns haben.

Der Blick über die Landesgrenzen zeigt in drei Beiträgen, welche Aktivitäten in den Niederlanden, in Schweden und Großbritannien durchgeführt werden. Besonders in diesen Artikeln, bei denen Autorinnen und Autoren mit einem existierenden Erfahrungsschatz in E-Science zu Wort kommen, zeigt sich, dass die Beschränkung auf eine technologische Fragestellung zu kurz greift. Ohne den Einzelbeiträgen vorgreifen zu wollen, ist dies eine gute Entwicklung für die Informatik: es geht nicht nur um die Nutzung von Technologie. Gerade die Gestaltung der Benutzerschnittstelle, die Organisation von Software-Entwicklungsprozessen oder die Gewährleistung eines sicheren IT-Betriebs sind Fragestellungen, die von Informatikern durch ihre Ausbildung und Erfahrungen besser als von jeder anderen Fachrichtung beantwortet werden können. Die dort beschriebene Einsetzung einer „human e-cloud“ oder von „research engineers“ zeigt, dass die Zukunft der Wissenschaft im Zusammenspiel von hochqualifizierten Wissenschaftlerinnen und Wissenschaftlern mit einer komplexen technischen Infrastruktur liegt und dafür zentrale Support-Institutionen geschaffen werden sollten.

Im Beitrag von Hazeleger et al. wird dargestellt, wie die Niederlande von der Bereitstellung von eScience-Ressourcen durch das zentrale niederländische eScience-Center profitieren. Dabei ist der wesentliche Aspekt eben nicht die Technologie, sondern ein Stab von sehr gut ausgebildeten und multidisziplinären eScience Research Engineers sowie eine offene Wissensdatenbank zu Softwaretools für die Forschung und das Wissen, wie damit umzugehen ist.

Die schwedischen Aktivitäten werden im Artikel von Eriksson et al. vorgestellt. Auch Schweden nutzt ein dediziertes e-Science Research Center (SeRC), und auch dort ist nicht die Bereitstellung der Technologie der Engpass, sondern das Wissen über den effizienten und effektiven Umgang damit. Die schwedische Antwort ist ebenfalls die Bereitstellung von beratenden Wissenschaftlern als „human e-cloud“, also als Teil einer gedachten zentralen und unterstützenden Infrastruktur. Ausgehend von einer Kollaboration zwischen mehreren schwedischen Universitäten hat sich SeRC in der Zwischenzeit zu einem eigenständigen Forschungsinstitut und einem zentralen Player in der schwedischen Wissenschaftslandschaft entwickelt.

Schließlich beschreibt der Beitrag von Neil Chue Hong das Vorgehen in der britischen Wissenschaftslandschaft. Mit dem Software Sustainability Institute (SSI) wird eine eindeutige Linie zwischen der Bereitstellung forschungsunterstützender Software und der Durchführung von Wissenschaft gezogen. Besonders deutlich wird das am Beispiel der Reproduzierbarkeit wissenschaftlicher Ergebnisse, welche im Bereich des E-Science unbedingt auf die beweisbare Korrektheit von Code angewiesen ist – hier kann die Informatik nicht nur helfen, sondern ist sogar unverzichtbar.

Die derzeitigen Möglichkeiten und Anstrengungen der Deutschen Forschungsgemeinschaft (DFG) stellt anschließend der Beitrag von Katerbow dar. Dies mag insbesondere für Hochschulen, welche in das Thema E-Science verstärkt einsteigen möchten, nützlich sein, aber es zeigt auch sehr gut den gegenwärtigen Stand vor Einführung einer NFDI wie im Vortrag von Yahyapour.

Die Ausgabe wird durch das Positionspapier der Gesellschaft für Informatik zu „E-Science“ beschlossen. Unsere Motivation war es, die Rolle der Informatik in diesem Kontext zu stärken und in den wissenschaftspolitischen Diskurs einzubringen. Das Positionspapier stellt hierfür einen ersten Schritt dar, es sollte in der nächsten Zeit jedoch, zum Beispiel durch aktives Einbringen der deutschen Informatik-Community in den Aufbau der Nationalen Forschungsdateninfrastruktur, mit Leben gefüllt werden.

**Torsten Eymann**  
**Michael Goedicke**  
**Ulrike Lucke**

# Zur disruptiven Digitalisierung von Hochschulforschung

*Faktoren der Skalierung und ein Zukunftsszenario*

Markus von der Heyde  
Andreas Hartmann · Gunnar Auth  
Christian Erfurth

## Das Schattendasein der digitalen Forschung

In der bisherigen Diskussion von Chancen und Risiken der Digitalisierung für die Hochschulen lässt sich ein erkennbarer Schwerpunkt im Bereich Lehre und Studium ausmachen. Unter dem Oberbegriff Digitale Bildung werden hier u. a. aus didaktischer und/oder informationstechnischer Perspektive Diskussionen und Entwicklungen weitergeführt, deren Grundlagen in den Jahren zuvor unter Bezeichnungen wie E-Learning und E-Teaching erarbeitet wurden. Deutlich wird dieser Schwerpunkt in Publikationen, hochschulinternen und -übergreifenden Projekten und Arbeitsgruppen sowie Förderinitiativen von Bund, Ländern und wissenschaftsnahen Stiftungen (vgl. [9]).

Betrachtet man den zweiten Kernbereich von Hochschulen, die Forschung, fällt im Vergleich zur breiten Diskussion um die Digitale Bildung ein gewisses Schattendasein beim Ausloten und Erschließen von Digitalisierungspotenzialen auf. Aus institutioneller Sicht der Hochschulen sind es hier immer noch Themen wie Forschungsdatenmanagement, Forschungsinformationssysteme oder Langzeitarchivierung, welche die Agenda schon seit Jahren bestimmen, sich aber nur langsam weiterentwickeln. Im Bereich der Disziplinen können dagegen große Fortschritte verzeichnet werden, sei es bei den oft bemühten Digital Humanities, in der Medizin(informatik), Physik, Chemie, Biologie, Materialwissenschaften oder nahezu jeder anderen Disziplin. Wie schließlich einzelne Forschende in und mit ihrem Umfeld von der Digitalisierung betroffen sind, wie diese für die individuelle Forschungsarbeit genutzt und gestaltet werden kann

oder auch welche Bedrohungen und Risiken sich ergeben können, kann jeweils nur individuell beantwortet werden. Die Voraussetzung für das Finden solcher individueller Antworten ist jedoch ein grundlegendes Verständnis für Einflussfaktoren und Wirkmechanismen der Digitalisierung. Auf dieser Basis können Hochschulen bedarfsorientierte Serviceangebote und digitale Infrastrukturen schaffen, die für Forschende echten Nutzen bedeuten.

Mit den disruptiven Innovationen der Digitalisierung geht auch eine zerstörerische Wirkung einher, ganz im Sinne der schumpeterschen „schöpferische[n] Zerstörung“ [20]. In der Digitalisierung der Wirtschaft hat die Entwicklung innovativer Geschäftsmodelle immer wieder zu tiefgreifenden Strukturveränderungen ganzer Branchen geführt. Der Erfolg einzelner Unternehmen bedeutet Existenzbedrohung für unzählige andere, sei es im Einzelhandel durch Amazon und ebay, im Bereich Mobilität durch Uber oder im Hotelgewerbe durch AirBnB. An Hochschulen sind z. B. in den Wirtschaftswissenschaften Geschäftsmodelle Gegenstand der Forschung; aber umgekehrt ist der Kernprozess Forschung meist nur am Rande im Fokus von Geschäftsmodellen (beispielsweise

<https://doi.org/10.1007/s00287-018-01126-1>  
© Springer-Verlag Berlin Heidelberg 2018

Markus von der Heyde  
vdH-IT, Weimar  
E-Mail: info@vdh-it.de

Andreas Hartmann · Gunnar Auth  
Hochschule für Telekommunikation Leipzig  
E-Mail: {hartmann, auth}@hft-leipzig.de

Christian Erfurth  
Ernst-Abbe-Hochschule Jena  
E-Mail: Christian.Erfurth@eah-jena.de

## Zusammenfassung

Digitale Werkzeuge unterstützen den Forschungs- und Lehrprozess an Hochschulen fundamental. Ein tatsächlicher Paradigmenwechsel im Sinne der disruptiven Digitalisierung ist aber bisher – zumindest in Deutschland – noch nicht an Hochschulen bzw. in deren Kernprozessen Forschung und Lehre zu verzeichnen. Dieser Beitrag leitet aus Faktoren, die bei der disruptiven Digitalisierung von Unternehmen beobachtet werden, ein Zukunftsszenario ab, in dem letztlich die Steuerung der Kernprozesse von Hochschulen durch Geschäftsmodelle verschoben wird. Ein hypothetischer Rückblick aus diesem Szenario erlaubt uns trotz Unschärfe der Vorhersage Schlussfolgerungen zu ziehen, die für CIOs und Hochschulleitungen aktuell richtungsweisend sein können.

beim Wissens- und Technologietransfer). Umso mehr fällt auf, dass im Zuge der Digitalisierung vermehrt neue privatwirtschaftliche Akteure zutage treten, die auf Basis digitaler Technologien neuartige Geschäftsmodelle rund um Forschungsdaten, -prozesse und -ergebnisse entwickeln wie bspw. ResearchGate, Mendeley oder Academia.edu. Hier liegt die Vermutung nahe, dass aus dieser Entwicklung auch eine zerstörerische Wirkung für die Forschung an Hochschulen entstehen könnte. Allerdings ist ebenfalls erkennbar, dass daraus Vorteile und Verbesserungsmöglichkeiten erwachsen.

Vor diesem Hintergrund beschäftigt sich der Beitrag mit zwei zentralen Fragestellungen:

1. Welche Implikationen hat die Digitalisierung aktuell und zukünftig für die Forschung an Hochschulen?
2. Welche Empfehlungen lassen sich daraus für die Gestaltung der Digitalisierung von Hochschulen ableiten?

Zu Beginn zeichnen wir die Veränderung im Begriffsverständnis der Digitalisierung nach. Dann werden die von uns beobachteten Faktoren der Digitalisierung, die im Zusammenhang mit einer extremen Skalierung stehen, definiert und erläutert. Ihre potenzielle und notwendige Funktion

in der zerstörenden Wirkung wird an Beispielen gezeigt. Danach wird dieser Blickwinkel auf die Forschung an Hochschulen übertragen. Abschließend werden diese Ergebnisse genutzt, um ein zukünftiges Szenario zu gestalten. Es soll dazu anregen, nach den Implikationen für die Digitalisierung an Hochschulen zu fragen. Eine Zusammenfassung schließt diesen Beitrag mit Empfehlungen für die Hochschulleitungen ab.

## Der Begriff der Digitalisierung

In Erweiterung der ursprünglichen Bedeutung, der Umwandlung von analogen in digitale Signale, bezieht sich „Digitalisierung“ heute im Geschäftskontext auf den Einsatz informationstechnischer Innovationen. Durch diesen Einsatz verändern sich sowohl die angebotenen Produkte oder Dienstleistungen als auch die Kundeninteraktion und Geschäftsprozesse [25, 26]. Neue, fast vollständig im digitalen Raum ablaufende Geschäftsmodelle wurden in kurzer Zeit so erfolgreich, dass die Weiterführung von herkömmlichen, analogen Geschäften infrage gestellt wird. Der Wandel der Bedeutung zu etwas Bedrohlichem, Zerstörerischem wurde in Bezug auf Hochschulen durch [9] beschrieben. Dort entwickelten die Autoren aus dem Blickwinkel des traditionellen IT-Managements Thesen, um den CIOs von Hochschulen einen Ausgangspunkt für die Diskussion einer „Digitalen Agenda“ anzubieten. Um die polarisierenden Effekte hier klarer hervorzuheben, verstehen wir im Folgenden die „disruptive Digitalisierung“ als die bahnbrechende und qualitative Veränderung, welche digitale Abläufe auf innovative Weise vernetzt und damit eine zerstörende Wirkung auf etablierte Geschäftsmodelle entfaltet.

Die disruptive Digitalisierung von Geschäftsabläufen kann global beobachtet und in ihrer potenziellen Wechselwirkung mit entsprechenden Märkten beschrieben werden. Jedoch fehlt bisher eine strukturierte Analyse, welche Faktoren für den Anstoß und die Steuerung von Veränderungen notwendig sind und welche abhängige Variablen darstellen. Ist beispielsweise die globale Skalierung nur eine Konsequenz der Veränderung oder deren Ursache? Die zwei folgenden Beispiele zeigen qualitative Veränderungen, die letztlich auf einer quantitativen Skalierung von mehreren Zehnerpotenzen beruhen. Es handelt sich dabei nicht um den oft zitierten „Schmetterlingseffekt“, bei dem

## Abstract

Digital tools are essential for science as well as teaching and learning at the university level. However, an actual paradigmatic shift via disruptive digitization has not reached the higher education sector at its core, at least in Germany. This article develops a view on the essential factors which lead entrepreneurial businesses to success on a global scale. We transfer these drivers into the university scope, which results in a future business model for the modern research matchmaker. How digital platforms as brokers will change governance in the future enables us to draw conclusions about what CIOs and boards of directors should do today.

weitreichende Folgen durch relativ kleine Veränderungen in chaotischen Systemen hervorgerufen werden.

## Analogie von Optik und Informatik

Die Existenz von geschliffenen Quarzen kann in Ägypten bis ins 8. Jh. v. Chr. zurückverfolgt werden. Umstritten ist, ob die Nutzung eher als Brennglas oder als optische Linse erfolgte. Die moderne Optik mit der Erklärung optischer Eigenschaften von Linsen begann erst im 9. Jh. n. Chr., als das Sehen als Lichteinfall ins Auge verstanden wurde. Die alltägliche Verwendung zur Korrektur von Fehlsichtigkeit begann im ausgehenden 13. Jh. Die Mehrfachanwendung (Kaskadierung) von Linsen führte zur Erfindung des Mikroskops und des Fernrohrs im 16. bzw. 17. Jh. Die ursprüngliche Linse wurde zum Instrument der Forschung und eröffnete neue qualitative Räume, die bis dahin nicht wahrnehmbar, studierbar, nutzbar oder monetarisierbar waren. Die optische Vergrößerung war beobachtet, theoretisch beschrieben, graduell verfeinert und schließlich durch Mehrfachanwendung um einige Zehnerpotenzen verstärkt worden. Die Entwicklung verlief über Jahrhunderte, sodass Veränderungen über Generationen zu einer Veränderung von Arbeits- und Lebensgewohnheiten sowie weiteren Innovationen führte.

Dagegen entstand die Informatik in kurzer Zeit, ausgehend vom booleschen Binärsystem Mitte des 18. Jh.s, folgend dem theoretischen Verständnis von rechnenden Systemen durch Turing 1937

bis zur automatisierten Informationsverarbeitung ab 1957. Die inzwischen als klassisch angesehene Verwendung von Computern mit dem Schema von Eingabe, Verarbeitung und Ausgabe ist Grundlage der Automatisierung bis in die 1990er-Jahre. Die Kaskadierung von Verarbeitungsschritten erzeugte mit der weltweiten Vernetzung von verfügbaren Daten im Internet abermals eine Skalierung in der quantitativen Größenordnung mehrerer Zehnerpotenzen und schließlich eine gesamt-qualitative Veränderung.

Zusammenfassend leiten wir aus diesen Beispielen ab, dass die Skalierung eine notwendige Voraussetzung für die anschließend eintretenden qualitativen Veränderungen ist. Erst durch die extreme quantitative Skalierung kann am Ende eine qualitative Veränderung herbeigeführt werden. Daher richtet sich unser Interesse im Folgenden auf die Skalierung, geleitet von den Fragen: Was trägt maßgeblich zu extremer Skalierung bei? Welche Faktoren führen zu einer Skalierung? Ist die Vernetzung von Kommunikation eine notwendige Voraussetzung? Im folgenden Abschnitt beschreiben wir Faktoren, die in den letzten Jahren in der disruptiven Digitalisierung in vielen Beispielen immer wieder in Erscheinung getreten sind.

## Der Weg zur monopolistischen Skalierung: Faktoren der Digitalisierung

Bei vielen Beispielen der disruptiven Digitalisierung beobachten wir besonders deutlich die folgenden Faktoren, wenn durch eine globale Skalierung der digitalen Abläufe ein faktisches Monopol erreicht wird:

- **Vermittlerrolle („broker“):** Anbieter und Kunde finden sich per Internet über standardisierte Schnittstellen und den gemeinsamen (monopolisierten) Start der Suche. Wird die Suche in eine Plattform integriert, ist gleichzeitig eine Abgrenzung von Mitbewerbern zu beobachten. Oft bleibt die Vermittlung auch der einzige Teil der primär erbrachten Leistung. Weitere Leistungen müssen von Anderen, den einzelnen Teilnehmern auf der Plattform, erbracht werden [7].
- **Offenheit („openness“):** Die Teilnahme von Kunden und Anbietern einer Dienstleistung oder eines Angebots ist offen. Private Endkunden und Kleinanbieter akzeptieren händischen Pflegeaufwand ihrer Daten, wenn dadurch eine verbesserte Suche nach Dienstleistungen bzw. Absatz ihrer

Leistung ersichtlich ist. Im Zuge der Skalierung wird durch Schnittstellen großen Anbietern die Möglichkeit eröffnet, die nun aggregierten Märkte zu adressieren. Aus dem Blickwinkel der „Multisided Platforms“ [23] werden Offenheit und Dezentralität der verwendeten Technologien als Kerneigenschaften des „Digital Entrepreneurial Ecosystems“ gesehen.

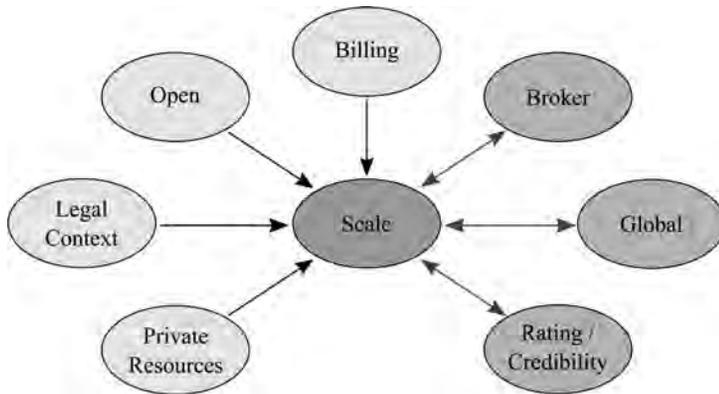
- **Bezahlen („billing“)**: Standardabläufe für einen einfachen Geldtransfer sind etabliert und binden alle häufig benutzten Wege ein. Bei kostenlosen Diensten kann ein verstecktes Geschäftsmodell vorhanden sein, dessen Bezahlvorgänge in kostenpflichtigen Mehrwertdiensten oder anderen Produkten zum Tragen kommen. Frederik Pferdt von Google erwähnt am Rande der Keynote der Campus Innovation 2017 [18], dass ein Bewerten und Kategorisieren von Bildern auf der kostenlosen Bild-Sharing-Plattform Picasa als Trainingsmaterial für KI-Algorithmen genutzt wurde.
- **Private Mittel („private resources“)**: Zur Wertschöpfung werden typischerweise Zeit, Dinge und/oder Services eingebracht, die eigentlich für den privaten Gebrauch bestimmt waren. Parallel oder sequenziell können sie weiterhin auch privat genutzt werden, sodass subjektiv keine Geschäftsinvestition vorliegt. Der damit niederschwellig erreichte Einstieg ermöglicht scheinbar jedem eine Teilnahme an der Wertschöpfung. Dieser Faktor wurde in [11] als ein kritischer Faktor beim Wachstum der Plattform Waze genannt.
- **Globales Angebot („global“)**: Der Service wird über das Internet angeboten, ist dadurch (fast) überall vorhanden und kann global genutzt werden. Durch die Offenheit gegenüber Anbietern und Abnehmern werden lokale Märkte zum Teilmarkt innerhalb der globalen Plattform.
- **Legaler Kontext („legal context“)**: Die Regeln (Vertrag) werden durch die Bindung an die Plattform normiert. Die Nutzung vereinfacht sich durch den hohen Standardisierungsgrad. Durch die geeignete Konstruktion eines Vertragswerks wird versucht, die Services direkt zwischen Nutzer und Anbieter zu regeln, obwohl sequenziell Anbieter und Abnehmer die Bindung mit der Plattform eingehen. Damit erlangen landesspezifische Regeln der Plattform Gültigkeit für die Vertragserfüllung zwischen Anbieter und Abnehmer. Dies kann u. a. Bezahlvorgänge, steuerliche Einordnung, Garantie,

Versicherung, Regressansprüche, Datenschutz und Verbraucherschutz betreffen.

- **Bewertungssystem („rating“/„credibility“)**: Die gegenseitige Bewertung (Rezension und Bewertung) ersetzt/unterstützt die Qualitätskontrolle. Sozialer Druck reguliert und reduziert den Aufwand für die Behandlung von Störungen zugunsten des Betreibers der Plattform. Eine wichtige Rolle spielt dabei die Transparenz, die Tapscott und Ticoll [24] als globale Kraft bei der Regulierung des neuen Marktes und aller relevanten Entscheidungen identifizierten.

Bisher konnten diese Faktoren nur empirisch durch Marktbeobachtung gelistet und beschrieben werden. Eine eingehende wissenschaftliche Evaluation steht aus. Mit einem zunächst eingeschränkten Blickwinkel der Gründerszene versuchen Sussan und Acs [23] einen Ordnungsrahmen zu formulieren, der ebenfalls drei der o. g. Punkte (Offenheit, die Brokerfunktion und die nutzerzentrierte Perspektive) beinhaltet. Eine alternative Suche nach Faktoren, welche die Skalierbarkeit von Geschäftsmodellen beeinflussen, haben Stampfl et al. [22] vorgenommen. Dort wurden die internen Faktoren Nutzung von Technologie, Struktur von Kosten und Einnahmen, Anpassungsfähigkeit an verschiedene Rechtsräume, Fähigkeit zur Nutzung von Netzwerkeffekten sowie die generelle Nutzerorientierung identifiziert. Diese Faktoren sind nicht homolog zu den von uns vorgeschlagenen externen Perspektiven, stehen aber über die Gestaltung des Geschäftsmodells in direkter Wechselwirkung. In diesem Papier konzentrieren wir uns nicht auf die internen Organisationsprinzipien der Start-ups, sondern beschreiben die Faktoren, die aus der Beobachtung (z. B. der Nutzerschnittstelle) von außen abzuleiten sind.

In weiteren Forschungsarbeiten wollen wir insbesondere die Vollständigkeit der Gesamtsicht sowie die jeweilige Notwendigkeit der Faktoren prüfen. Dabei soll auch hier die Wirkweise in Bezug auf die Skalierung und damit die Grundlage für eine quasimonopolistische Dominanz untersucht werden. Dass aus juristischer Perspektive globale digitale Plattformen schwer zu regulieren sind, diskutieren z. B. Evans [6] sowie Capobianco und Nyeso [3]. Unsere Hypothese der Wirkung auf und durch Skalierung ist in Abb. 1 dargestellt: Alle Faktoren begünstigen eine Skalierung z. B. der Plattform; die



**Abb. 1 Wirkung der Faktoren der Digitalisierung auf die Skalierung von Plattformen**

Skalierung wirkt ebenfalls fördernd auf die Faktoren „broker“, „global und rating/„credibility“, sodass eine selbstverstärkende und stabilisierende Wirkung entsteht.

In der Alltagserfahrung zeigen sich die Existenz der Faktoren und konkrete Ausprägungen z. B. in den folgenden Geschäftsfeldern und Plattformen: Mobilität (mit Uber, flinc, BlaBlaCar), Wohnen (mit AirBnB, CouchSurfing, Open Avenue) oder Risikokapital (mit Kickstarter, InvestX, CrowdMatrix, Crowd-Investment). Die erstgenannten Beispiele sind in ihrer Kategorie die aktuellen Kandidaten für eine globale Dominanz. Bisher sind zwar keine klaren Gewinner dieses Wettlaufs zu erkennen. In anderen Geschäftsfeldern (z. B. Google im Bereich der WWW-Suche) ist die gesellschaftliche akzeptierte Dominanz aber soweit fortgeschritten, dass Wörter wie „googeln“ bereits 2004 in den Duden [1] aufgenommen wurden.

Welche Rolle spielt dabei der von Tapscott und Williams durch „wikinomics“ beschriebene Wandel des „read-only“-Webs zum Web 2.0, bei dem „die Massen“ in partizipativen Ansätzen in die Wertschöpfung der Unternehmen eingebunden wurden [25]? Peters identifiziert in ihm eine der Quellen für die Kapitalisierung der Bildung (siehe Tab. 5 in [17]).

Darüber hinaus bleibt zu klären, ob Onlineplattformen die genannten Faktoren zusammenbinden und zu einem hinreichenden quantitativen Skalierungseffekt führen können. Bisherige Forschung fokussierte hier primär auf die intern bestimmbareren Faktoren, z. B. das Businessmodell oder die Optimierung der Plattform [26]. In zukünftiger Forschung ist zu klären, welche Faktoren unabhängige Variablen sind, also durch die Gesellschaft oder

Wirtschaft beeinflusst werden können, und welche Faktoren davon abhängig sind, also sekundär als Messgrößen beobachtet werden können.

Auch im wissenschaftlichen Bereich gibt es heute schon entsprechende Plattformen im Internet, z. B. ResearchGate, academia.edu, scienceopen.com, Mendeley, Peerus oder ORCID. Finden sich die o. g. Faktoren auf diesen Plattformen wieder? Zu welchem Ergebnis führt das und wie können wir frühzeitig die Veränderung verstehen? Was ist der richtige Zeitpunkt zur Partizipation und lohnt sich eine frühe Adaption? Es wird schnell offensichtlich, dass Hochschulen eine Antwort auf diese Fragen finden müssen.

### Beobachtungen in der Forschung

Dieser Abschnitt beschreibt Aufgaben in der universitären Forschung und digitalisierungsbedingte Veränderungen, die bereits heute erkennbar sind. Tabelle 1 zeigt eine Übersicht der Kernaufgaben und damit verbundene Aktivitäten im Forschungsprozess. Aktuell gibt es am Markt eine Vielzahl von digitalen Angeboten und Werkzeugen, die Teile aus dem Forschungsprozess unterstützen. Eine übergreifende Plattform gibt es bisher noch nicht.

Hinsichtlich dieser Abläufe gibt es eine Vielzahl von Beispielen der digitalen Unterstützung im klassischen Sinne der IT-Nutzung [15]. Das kann einzeln oder aufbauend im Sinne einer Kaskadierung zutreffen, wenn Szenarien nebenläufig oder parallel von mehreren Forschern ausgeführt werden bzw. Forschungsdaten eine mehrfache – meist ursprünglich nicht intendierte – Nutzung erfahren. Der Hypothese folgend entstehen erst dann disruptive Veränderungen in der Qualität, wenn



## Kernaufgaben und unterstützende Abläufe im Forschungsprozess

Kernaufgaben im Forschungsprozess	Unterstützende Abläufe	Vorhandene digitalisierte/disruptive Modelle für unterstützende Abläufe
<ul style="list-style-type: none"> <li>· Deskription (Beschreibung)</li> <li>· Explikation/Explanation (Erklärung)</li> <li>· Prognose (Vorhersage)</li> <li>· Konstruktion (Gestaltung)</li> <li>· Verifikation/Evaluation (Überprüfung)</li> </ul>	<ol style="list-style-type: none"> <li>1. Suchen/Finden</li> <li>2. Sammeln/Ordnen</li> <li>3. Dokumentieren</li> <li>4. Publizieren/Diskutieren/ Vernetzen (Community)</li> <li>5. Verwerten (Transfer) und Akquise (neue Projekte, Finanzierung)</li> <li>6. Management von Forschungsprojekten</li> </ol>	<ol style="list-style-type: none"> <li>1. Google Scholar, WorldWideScience, Web of Science, ORCID [16], IS/EN [12], scienceopen.com, academia.edu</li> <li>2. ResearchGate, Mendeley, Zotero, scienceopen.com</li> <li>3. GitHub, labguru</li> <li>4. Peerus, ResearchGate, Springer, Elsevier, figshare</li> <li>5. Google, Peerus, scienceopen.com</li> <li>6. Global Science Network, Science Connect</li> </ol>

mehrere Größenordnungen der quantitativen Skalierung überschritten werden. Drei aktuelle Beispiele zeigen hierzu, wie Mengeneffekte bei der Einbindung der Allgemeinheit (\*Open) entstanden sind, hinsichtlich der Bearbeitung, Finanzierung oder Ressourcenbereitstellung.

**Crowd-Working.** In Foldit [28] wird die Suche nach der richtigen Faltung von Molekülen durch viele Freiwillige spielerisch unterstützt. Die Algorithmen waren bisher nicht in der Lage, die tatsächliche Faltung, sprich: Geometrie eines Moleküls im 3D-Raum aus der Gensequenz so abzuleiten, dass chemisch wirksame Bindestellen z. B. in Enzymen korrekt gefunden wurden.

**Crowd-Funding.** Die Gen-Sequenzierung für spezielle Fälle seltener Krankheiten ist in der Grundlagenforschung oftmals schwierig zu finanzieren, da die Anzahl der Fälle aus Sicht der Pharmaindustrie keine Rendite im Vergleich zu anderen sehr häufigen Krankheiten und deren genetischer Disposition vermuten lassen. In dem Projekt Mass Genomics [13] wird daher über ein Crowd-Finanzierungsmodell versucht, eine Gemeinschaft für die genetische Arbeit an seltenen Krankheiten aufzubauen.

**Crowd-Ressourcen.** Die Suche nach der größten Mersenne-Primzahl [14] hat beachtliche Ressourcen (Faktor 2,5 größer als der schnellste Computer aus der TOP500-Liste) aus dem privaten Umfeld

in einem Projekt vereinigt. Die parallele Nutzung privater Ressourcen von rund 180.000 Nutzern war hier der Kernfaktor des Erfolgs, da der Einsatz von gängigen HPC-Ressourcen für diese Art der Grundlagenforschung oftmals gescheut wird. Durch die Einbindung von Enthusiasten mit mathematischem Interesse war es möglich, eine erhebliche Rechenleistung für die Suche zu koordinieren.

Für alle drei Projekte hat sich je eine Plattform inklusive relevanter Schnittstellen etabliert. Einige der Faktoren, die wir für die disruptive Digitalisierung der Wirtschaft genannt haben, sind bei diesen Plattformen gut sichtbar. Sie haben potenziell dazu beigetragen, dass eine quantitative Skalierung stattgefunden hat, obgleich es dafür bei keinem der Projekte vorher eine Garantie gab. Zwar sehen wir keinen Hinweis, dass in diesen Beispielen die genannte Veränderung des Forschungsprozesses begründet wäre. Nehmen wir jedoch an, die Faktoren würden hypothetisch weitergetrieben werden, bis die disruptive Digitalisierung sichtbar eingesetzt hat und die quasi-monopolistische Skalierung erreicht wurde. Wie sähe dieses Zukunftsmodell aus und welche Erkenntnisse können wir daraus ableiten?

### Ein Zukunftsmodell

Werden die Faktoren demnach als Treiber der Digitalisierung verwendet, lässt sich ein weitergehendes, disruptives Szenario der Forschung an Hochschulen entwickeln. Das Szenario dient bewusst als Untersuchungsobjekt, um im hypothetischen Rück-

blick die Veränderungen besser zu verstehen und Erkenntnisse daraus ableiten zu können.

### **Das Smartphone im Jahr 1998**

Betrachten wir eine Frage, die im Jahr 1998 hätte gestellt werden können: „Was wäre, wenn ich unterwegs mit meinem Handy Angebote auf Internetseiten nutzen könnte, indem ich einfach mit dem Finger auf das Display tippe?“. Rückblickend ist der Nutzen von Smartphones heute klar erkennbar. Der Kauf eines Fahrscheins im öffentlichen Nahverkehr oder das Bezahlen im Café hat sicher nicht die Entwicklung der Geräte ausgelöst. Dennoch ermöglicht die bloße Verfügbarkeit des Smartphones diese Funktionen.

Viele flexible, vernetzte Anwendungen sind heute kostengünstig zu realisieren, ohne dass dies bei der Konstruktion beachtet wurde. Welche Fragen hätten im Jahr 1998 dazu beigetragen, dennoch einige dieser Funktionen sicher vorauszusagen? Welche Fragen müssen wir uns im übertragenen Sinn stellen, um die Wirkung der disruptiven Digitalisierung in der universitären Forschung abschätzen zu können? Betrachten wir dazu ein fiktives Szenario in der Zukunft, welches bereits aktuell bekannte Modelle aus Tab. 1 verwendet.

### **Szenario eines Geschäftsmodells für universitäre Forschung im Jahr 2038**

Wie an den meisten Wochentagen fährt Prof. Gamchy heute zum Source-Campus der Universität, denn hier sind die Arbeitsbedingungen einfach besser als zu Hause. *Source* – das ist das Campus Center, das aus der Online Plattform *Share Our University Research* (SOUR) entstanden ist. Zunächst bot SOUR lediglich einen kostenlosen Service zur Suche, Dokumentation und Vernetzung von wissenschaftlichen Artikeln an. Auch Herr Gamchy meldete sich schnell an der offenen Plattform an, denn schließlich gab es weitere kostenlose Zusatzangebote. Wie viele Wissenschaftler begann er, referenzierte Literatur sowie eigene Artikel und Projektinformationen online zu sammeln. Im Gegensatz zu seiner früheren Uni-E-Mail konnte sich Herr Gamchy hier lebenslang mit anderen Forschern vernetzen und seine Forschung über eine eindeutige Identifikation kennzeichnen. Während damals einige Kollegen noch immer über den Sinn von Forschungs-Informationssystemen diskutierten, wuchs SOUR weiter und in wenigen Jahren waren dort nicht nur rund 80 % aller Wissenschaftler namhafter Universitäten weltweit

vertreten. Vielmehr hatten sich auch die Forschungsgruppen der großen Konzerne und öffentlicher Einrichtungen angeschlossen.

Heute werden täglich Tausende von Forschungsfragen auf der Plattform zur Lösung zwischen Industrie und Wissenschaftlern vereinbart. Der Wert der Forschungsleistung kann wenige Hundert bis mehrere Mio. EUR betragen, wobei eine unbürokratische Abrechnung im Dreiecksgeschäft den Wissenschaftlern enorm Zeit spart. Als dritter Partner behält SOUR rund 3–5 % mit jeweiligen Unter- und Obergrenzen ein. Nicht nur der Preis wird anhand zentraler Faktoren zwischen Auftragnehmer und Auftraggeber abgestimmt – auch ein öffentliches Bewertungssystem für Ethik und Qualität sorgt für Transparenz im Netzwerk. Die Plattform bindet beide Seiten juristisch so zusammen, dass der Rechtsraum für alle erträglich und in Bezug auf Datenschutz, ethische Rahmenfaktoren und steuerlichen Fragestellung meist vorteilhaft empfunden wird.

Später bildeten sich dann die SOUR Campus Center (Source) und übernahmen zunehmend das Forschungsgeschäft der Universitäten, sodass heute kaum eine Einrichtung ohne Kooperationsvertrag auskommt. In der Regel handelt es sich um einen eigenen Source-Campus, der auch durch öffentliche Finanzierungen gefördert wird. Selbst weite Teile der digitalen Lehrangebote werden über die Source realisiert. Die Ressourcen von Source sind zudem weltweit nutzbar, sodass Herr Gamchy bequem auf robotergesteuerte Labore in anderen Ländern zugreifen kann. Diese hochmoderne Ausstattung kann seine eigene Universität schon lange nicht mehr selbst finanzieren. Nichtsdestotrotz will Herr Gamchy kommendes Jahr ein paar Monate an seiner alten Fakultät arbeiten und die alte Bibliothek besuchen. Nur hier kann er sich zurücklehnen, konzentrieren und neue Projektideen entwickeln.

### **Implikationen für das strategische IT-Management an Hochschulen**

Das fiktive Szenario zeigt eine mögliche Zukunft. Gleichwohl gehören Universitäten und kirchliche Einrichtungen zu anscheinend selbststabilisierenden Organisationen, die Jahrhunderte mit nur wenigen inneren Veränderungen überdauern. Aber müssen sich Universitäten überhaupt ändern, damit das Szenario Wirklichkeit werden kann? Viele Elemente des Szenarios finden gerade außerhalb der Universität statt, weil sich die einzelnen Wis-

senschaftler disziplinar vernetzen und Plattformen außerhalb der traditionellen Institutionen für die Aufgaben des Alltags nutzen. Welche bremsende Rolle dabei die oft zitierten Bedenkenräger haben und ob diese in der Lage sind, die Entwicklung in andere oder sogar gezielte Richtungen zu lenken, bleibt Spekulation.

Heute sind Servicemanagement, Projektmanagement und Informationsmanagement an den Hochschulen als Werkzeuge für die Digitalisierung angekommen. Eine Strukturierung der Entscheidungsprozesse ist für die erfolgreiche Anwendung dieser Managementtechniken eines der stabilisierenden Elemente, welche von der DFG seit Jahren empfohlen wird [5]. Diese IT-Governance an Hochschulen ist dringend notwendig, wird aber oftmals durch „gewachsene“ Kulturen erschwert. Fasst man Hochschulen, die oft als lose gekoppelte Entscheidungssysteme [2] bezeichnet werden, allgemein als ein System von Systemen auf, ist die empfohlene Auflösung des Steuerungsparadox aus [19] anwendbar. Dort heißt es interessanterweise „For SoS, we contend that management is replaced by governance to cope with paradox.“

Unabhängig von der Absicht, Gewinn zu erzielen, befindet sich das Verständnis für Geschäftsziele, Geschäftsprozesse und Geschäftsobjekte an Hochschulen im Wandel. Erste Ansätze existieren, wie der Kerndatensatz Forschung [10] oder CAUDIT [4] zeigen. Für das strategische Management sind konkrete Informationen zu Prozessen und Geschäftsobjekten sowie deren Beziehungen untereinander von Bedeutung. Neben der Dokumentation der Prozesse müssen die zu realisierenden Applikationen bekannt sein, um effektiv Enterprise Architecture Management (EAM) zu betreiben. Mit Hilfe von Projekten und Informationen zu der Strategie und den Zielen der Hochschule kann der CIO die IT strategisch an der jeweiligen Hochschule ausrichten. In [9] sind diese klassischen Felder des IT-Managements dahingehend untersucht, ob sie zu einer Steuerung einer Hochschule unter der externen Wirkung der disruptiven Digitalisierung einen signifikanten Beitrag leisten könnten.

### **Was bedeutet dies für die Rolle eines Hochschul-CIOs?**

Unser Zukunftsmodell zeigt eine disruptive Veränderung des Geschäftsmodells Auftragsforschung an Hochschulen, die sich unserer Meinung nach

in ähnlicher Weise auch auf die heute öffentlich finanzierte (Grundlagen-)Forschung auswirken könnte. Da (private und öffentliche) Forschungsgelder einen erheblichen Teil der finanziellen Mittel der Hochschulen ausmachen, begreifen wir die damit befassten Prozesse als Umsetzung eines Geschäftsmodells. Aktuell wird (noch) kein Gewinn angestrebt, aber neben der Lehre ist die Forschung der klassische Kernprozess der Hochschulen. In Ländern mit Studiengebühren kann man auch die Lehre als Geschäftsmodell bezeichnen. Auch wenn diese Einordnung ungewohnt scheint, geht es letztlich wie in der Wirtschaft um Finanzierung und Einfluss. Für letzteres gibt es an den meisten Hochschulen noch zu wenig regulierende Prozesse im Sinne der IT-Governance [8].

Im Szenario haben treibende Faktoren des Marktes zu quantitativen Skaleneffekten geführt und letztlich eine qualitative Auswirkung auf die Organisationen ausgeübt. Es stellt sich die Frage, wie Hochschulen heute und in mittelbarer Zukunft (re-)agieren können und welche Rolle der Hochschul-CIO einnehmen kann bzw. sollte. Ginge es um Unternehmen und die Wechselwirkung mit den Märkten, rückten die Begriffe Agilität/Flexibilität, Schnelligkeit, Innovation/Kundennähe, Resilienz und Stabilität in den Vordergrund. Anders als Hochschulen können sich Unternehmen diesbezüglich mit den zuvor beschriebenen Handlungsfeldern wie z. B. EAM, IT-Governance und Strategie positionieren. Restrukturierungen der internen Organisation und Abläufe bis hin zur Änderung des Geschäftsmodells bereiten die Unternehmen auf die Herausforderungen der Digitalisierung vor – wenn auch ohne Garantie des Erfolgs. Die Forschungsperspektive betreffend wird der Markt jedoch im Wesentlichen durch die Community repräsentiert und daher sind des Weiteren die Kriterien Originalität, Relevanz, Objektivität und Überprüfbarkeit zu beachten. Somit leiten wir mehrere Erkenntnisse ab:

Zunächst sind und können CIOs nicht alleine für die Digitalisierung verantwortlich sein. Das System Hochschule ist eine Zusammensetzung aus Menschen, ihren Aufgaben und heute auch der Technologie, welche die Aufgaben unterstützt [21]. Die Herausforderungen betreffen damit gleichermaßen die Organisation und die Strategie der Hochschulen, sodass eine enge Abstimmung in den Hochschulleitungen und Digitalisierung als TOP-

Thema zwingend erscheint. Insbesondere muss die Digitalisierung im Sinne der Veränderung von Geschäftsmodellen unbedingt von der klassischen, unterstützenden IT differenziert werden, so wie es viele Unternehmen heute schon tun.

Weiterhin muss der CIO in der Community teilhaben und Entwicklungen genauso akribisch beobachten wie den allgemeinen Markt. Er muss erkennen und diskutieren, welche Innovationen und bereits vorhandene Plattformen das Potenzial für quantitative Skaleneffekte haben und somit eine disruptive Digitalisierung einleiten können. Hier genügt nicht der Blick auf die Softwarebranche, sondern vielmehr ist eine intensive Kommunikation mit den Wissenschaftlern gefragt. Sie sind Teil der Community und können am besten über Veränderungen und Plattformen berichten.

Zusätzlich müssen die Hochschulen mit Unterstützung ihrer CIOs einen Pfad identifizieren, bei dem sie in der Zukunft partizipieren, ohne dass dabei das eigene Geschäftsmodell unwiderruflich zerstört wird. Zweifelsohne wird das ein *gemeinsamer* Pfad sein müssen, denn individuelle Insellösungen haben bei aller Qualität keine Chance. CIOs müssen demnach mit vollständiger Kenntnis der eigenen Enterprise-Architektur (inkl. Forschungsinfrastruktur) die Hochschulen befähigen, auf neue Plattformen mit Agilität und Schnelligkeit reagieren zu können. Dabei muss gleichzeitig die eigene Integrität von Forschungsdaten und -abläufen gewahrt werden. Grundlage hierfür ist allerdings, dass CIOs in ihrer Tätigkeit durch die Organisation und Governance unterstützt und nicht behindert werden. So stehen Hochschulen zusätzlich vor der Herausforderung, ihre inneren Strukturen hinsichtlich der anstehenden Aufgaben anzupassen.

### Schlussfolgerung

Digitalisierung als gesellschaftliches Phänomen macht keine Ausnahmen. Die Abbildung, Vernetzung und fast unbegrenzte Skalierung von automatisierten Abläufen verändert die Gesellschaft. Es verändert daher, wie wir leben und arbeiten, wie wir denken und handeln. Wenn die Potenziale der Digitalisierung der Erfolg des Einen sind, müssen die Nachteile ebenfalls vorhanden sein. Der zerstörende Charakter der disruptiven Digitalisierung erreicht oft erst durch später sichtbare Folgen (z. B. unsere technologische Abhängigkeit) eine gesellschaftliche Dimension. Diese wird primär dann erreicht,

wenn die Nutzung von digitalen Technologien eine quasi-monopolistische Skalierung erreicht.

Zu beobachten ist, dass im Rückblick die erfolgreich digitalisierten und oft als Plattform etablierten Dienstleistungsangebote eine hohe Übereinstimmung in den beschriebenen Faktoren haben. Überträgt man die zu erzielende Wirkung auf die Forschung an Hochschulen, entsteht ein Zukunftsszenario. Teile daraus sind bereits heute als erste Ansätze erkennbar, insbesondere dort, wo Forschungsergebnisse gesammelt, vernetzt und weiterkommerzialisieren werden. Welche Effekte potenziell oder tatsächlich daraus für Hochschulen entstehen, ist Spekulation. Dennoch gibt es generell sinnvolle Handlungsoptionen für die aktuell verantwortlichen Präsidien und den CIO.

Der Einfluss besteht primär im Inneren der Organisation. Zurzeit werden oft Wirtschaftsbegriffe wie Agilität/Flexibilität, Schnelligkeit, Innovation/Kundennähe, Resilienz und Stabilität herangezogen, um dahinter stehende Probleme an Hochschulen anzusprechen. Stattdessen bedarf es einer expliziten, dokumentierten und wirkungsvollen Governance, nicht als Deckmantel einer entscheidungsverzögernden Bürokratie oder kompromissbetonten Gremienkultur, sondern als einziger Weg, in dem komplexen System Hochschule zu Transparenz und Kommunikation für alle IT betreffenden Entscheidungen zu kommen.

Die sich daraus entfaltende Wirkung steht im Zusammenhang mit dem Gesamtsystem der Gesellschaft und liegt damit außerhalb des direkten Einflusses von Einzelnen. Es erfordert das Zusammenwirken Vieler, um die positiven Entwicklungen so zu fördern, dass Freiheit von Forschung und Lehre nicht nur gesetzlich verankert, sondern tatsächlich vorhanden bleibt.

### Danksagung

Wir danken Herrn Prof. Dr. Peer Pasternack sowie den beiden anonymen Gutachtern für wertvolle Hinweise und hilfreiche Anregungen zur Fertigstellung dieses Beitrags.

### Literatur

1. Bibliographisches Institut GmbH (2004) Duden | googeln | Rechtschreibung, Bedeutung, Definition, Herkunft. <https://www.duden.de/rechtschreibung/googeln>, letzter Zugriff: 12.11.2018
2. Bode A (2002) Universität im Wandel: Die Rolle des CIO bei der Erneuerung der Prozesse. *Inform Consult Manag* 17:43–47
3. Capobianco A, Nyoso A (2017) Challenges for competition law enforcement and policy in the digital economy. *J Eur Compet Law Pract* 1–9, <https://doi.org/10.1093/jeclap/lpx082>

4. CAUDIT et al (2016) The Higher Education Reference Architecture, <http://www.cheita.org/publicationsresources/ea-model/>, last access: 12.11.2018
5. Deutsche Forschungsgemeinschaft (2016) Informationsverarbeitung an Hochschulen – Organisation, Dienste und Systeme: Empfehlungen der Kommission für IT-Infrastruktur für 2016–2020, Bonn
6. Evans DS (2017) The Emerging High-Court Jurisprudence on the Antitrust Analysis of Multisided Platforms. Social Science Research Network, Rochester, NY, <http://dx.doi.org/10.2139/ssrn.2948596>
7. Evans DS, Schmalensee R (2016) Matchmakers: The New Economics of Multisided Platforms. Harvard Business Review Press, Boston, MA. ISBN 978-1-63369-173-5
8. von der Heyde M, Breiter A (2015) Wer entscheidet? IT-Governance an Hochschulen. Lecture Notes in Informatics. Gesellschaft für Informatik, Bonn, S 651–662
9. von der Heyde M et al (2017) Hochschulentwicklung im Kontext der Digitalisierung – Bestandsaufnahme, Perspektiven, Thesen. In: Eibl M, Gaedke M (Hrsg) INFORMATIK 2017. Gesellschaft für Informatik, Bonn, S 1757–1772
10. Institut für Forschungsinformation und Qualitätssicherung (IFQ) et al (2015) Projekt „Spezifikation Kerndatensatz Forschung“ – Dokumentation der Ergebnisse, [http://www.kerndatensatz-forschung.de/version1/Ergebnisbericht\\_Projekt\\_KDSF.pdf](http://www.kerndatensatz-forschung.de/version1/Ergebnisbericht_Projekt_KDSF.pdf), letzter Zugriff: 12.11.2018
11. Ismail S (2014) Exponential Organizations: Why New Organizations Are Ten Times Better, Faster, And Cheaper Than Yours (And What To Do About It). Diversion Books, New York
12. Jähnichen S et al (2017) Individual perSonal data Auditable addrEss Number (ISÆEN), <https://gi.de/themen/beitrag/individual-personal-data-auditable-address-number-isaen/>, last access: 12.11.2018
13. Koboldt D (2012) Crowd-funded Exome Sequencing for Rare Genetic Diseases. <http://massgenomics.org/2012/07/crowd-funded-exome-sequencing-for-rare-genetic-diseases.html>, last access: 12.11.2018
14. Mersenne Research, Inc. Great Internet Mersenne Prime Search – PrimeNet, <https://www.mersenne.org/>, last access: 12.11.2018
15. Nentwich M, König R (2012) Cyberscience 2.0 – Research in The Age of Digital Social Networks. Campus, Frankfurt/Main
16. ORCID, Inc: ORCID: Our Mission, <https://orcid.org/about/what-is-orcid/mission>, last access: 12.11.2018
17. Peters MA (2009) Education, Creativity and the Economy of Passions: New Forms of Educational Capitalism. Thesis Eleven 96(1):40–63
18. Pferdt FG (2017) Zukunft jetzt – Kreatives Denken für Innovationen von morgen. Campus Innovation 2017, Hamburg
19. Sauser B et al (2009) System of systems management. In: Jamshidi M (eds) System of Systems – Innovations for The 21st Century. Wiley, Hoboken, pp 191–217
20. Schumpeter J (1997) Theorie der wirtschaftlichen Entwicklung: Eine Untersuchung über Unternehmerrgewinn, Kapital, Kredit, Zins und den Konjunkturzyklus. 9. Aufl. Duncker & Humblot, Berlin
21. SEBoK contributors: Socio-Technical features of systems of systems, [https://www.sebokwiki.org/wiki/Socio-Technical\\_Features\\_of\\_Systems\\_of\\_Systems](https://www.sebokwiki.org/wiki/Socio-Technical_Features_of_Systems_of_Systems), last access: 12.11.2018
22. Stampfl G et al (2013) An explorative model of business model scalability. International J Prod Dev 18(3–4):226–248, <http://dx.doi.org/10.1504/IJPD.2013.055014>
23. Sussan F, Acs ZJ (2017) The digital entrepreneurial ecosystem. Small Bus Econ 49(1):55–73, <http://dx.doi.org/10.1007/s11187-017-9867-5>
24. Tapscott D, Ticoll D (2003) The Naked Corporation: How The Age of Transparency Will Revolutionize Business. Simon and Schuster, New York
25. Tapscott D, Williams AD (2008) Wikinomics: How Mass Collaboration Changes Everything. Penguin, London
26. Urbach N, Ahlemann F (2016) Die Digitale Revolution – Wie technologische Trends die Business-Welt verändern. In: IT-Management im Zeitalter der Digitalisierung. Springer Gabler, Berlin, Heidelberg, S 1–20
27. Weill P, Woerner SL (2013) Optimizing your digital business model. MIT Sloan Manag Rev 54(3):71
28. Wikipedia (2015) Foldit, <https://de.wikipedia.org/wiki/Foldit>, letzter Zugriff: 12.11.2018

# E-Science Infrastrukturen

*in Deutschland und im internationalen Vergleich*

Ramin Yahyapour

## E-Science und die Wissenschaft

Über die Veränderung der Wissenschaft durch zunehmende Digitalisierung gibt es bereits seit einigen Jahren einen wissenschaftspolitischen und gesellschaftlichen Dialog. Ohne in eine kritische, teilweise ideologisch geführte Debatte einzusteigen, inwieweit diese Digitalisierung traditionelle Erkenntnisprozesse infrage stellt, lässt sich festhalten, dass es in zahlreichen Disziplinen Fragestellungen gibt, die ohne die rasante Entwicklung bei Rechnersystemen nicht erfasst und bearbeitet werden können. Ebenso lassen sich in immer größerem Umfang Daten erfassen, speichern und verarbeiten, wodurch sich neue Quellen für den Erkenntnisgewinn ergeben. Jim Grey hat in dem Buch „The Fourth Paradigm“ die Entwicklung der ersten drei Paradigmen der Forschung von 1) empirischer Beobachtung und Experimentieren zu 2) analytischen oder theoretischen Ansätzen und 3) rechen- und simulationsgestützter Forschung aufgezeigt [1]. Das vierte Paradigma ist die Gewinnung von neuen Erkenntnissen durch neue technische Möglichkeiten in der Zusammenführung, Verknüpfung und Auswertung von strukturierten und unstrukturierten Daten. Die Ansätze stehen nicht in Konkurrenz, sondern zeigen neue Möglichkeiten auf, die in den verschiedenen Fachdisziplinen unterschiedlich und häufig gleichzeitig zum Einsatz kommen.

Begriffe wie E-Science oder E-Research sind aufgrund ihrer teilweise inflationären Verwendung und unklaren Definition nur eingeschränkt geeignet, um diese Art der Wissenschaft adäquat zu beschreiben. Es ist jedoch unstrittig, dass zahlreiche Disziplinen durch den effektiven Einsatz von digitalen Methoden für das Erfassen, Erstellen und die Kuratation

von Daten sowie die Analyse und Auswertung und die Vernetzung und Publikation von Daten profitieren. Es braucht zwei Hauptfaktoren, um dieses Potenzial nutzen zu können: Methodenwissen für eine rechner- und datengestützte Forschung sowie der Zugang zu geeigneten Infrastrukturen, die eine solche Forschung erlauben. Solche Infrastrukturen können spezielle Rechnersysteme sein, wie sie im wissenschaftlichen Hoch- und Höchstleistungsrechnen existieren; es können aber auch Datenrepositorien sein, in denen aktuelle wie historische Informationen verfügbar gemacht werden. Wenn wir den Zugang zur Infrastruktur näher betrachten, so bedarf es einer klareren Fassung, welche hier gemeint ist. Häufig finden sich hier Strukturen, die gänzlich unabhängig voneinander entwickelt wurden und betrieben werden.

## Hoch- und Höchstleistungsrechnen

Eine typische E-Science-Infrastruktur findet sich im Bereich Hoch- und Höchstleistungsrechner (High-Performance Computing – HPC). Hierunter versteht man Rechencluster oder Spezialcomputer, die für die Bearbeitung größerer Fragestellungen, z. B. bei Simulationen und Analysen, benötigt werden. Aufgrund der teilweise signifikanten Investitions- und Betriebskosten hat sich in Deutschland in Anlehnung an die Empfehlungen des Wissenschaftsrates aus 2012 [7] und 1995 [6] eine leistungsfähige na-

<https://doi.org/10.1007/s00287-018-01131-4>  
© Springer-Verlag Berlin Heidelberg 2018

Ramin Yahyapour  
Gesellschaft für wiss. Datenverarbeitung mbH Göttingen  
(GWDG)/ Georg-August-Universität Göttingen  
E-Mail: [ramin.yahyapour@gwdg.de](mailto:ramin.yahyapour@gwdg.de)

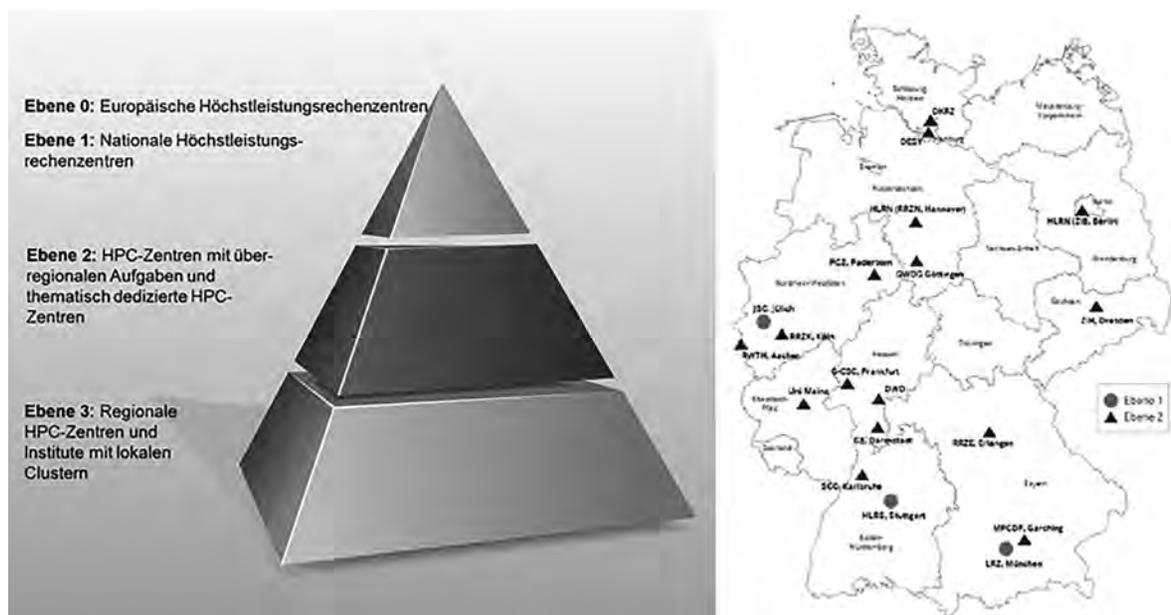


Abb. 1 Deutsche HPC-Versorgungspyramide. (Quelle: Gauß-Allianz e. V.)

tionale Infrastruktur herausgebildet. Dabei handelt es sich um ein hierarchisches Netzwerk, das über mehrere Ebenen je nach Größe und Komplexität strukturiert ist. Diese sogenannte „HPC-Pyramide“ ist dreistufig. Auf der obersten Ebene 1 befindet sich das „Gauß Centre for Supercomputing“, das aus dem Jülich Supercomputing Centre (JSC), dem Leibniz-Rechenzentrum (LRZ) in München und dem Höchstleistungsrechenzentrum Stuttgart (HLRS) besteht. Auf der Ebene 2 finden sich 15 HPC-Systeme an geografisch verteilten, meist universitären Zentren mit überregionalen Aufgaben und teilweise disziplinären Schwerpunkten. Auf der Ebene 3 befinden sich Rechnersysteme, meist an Universitäten, für den lokalen Bedarf. Diese Struktur hat sich bewährt, da die einzelnen Ebenen unterschiedliche Zugangsmechanismen und Schwerpunkte haben. Wissenschaftler und Wissenschaftlerinnen können an lokalen Zentren der Ebene 3 oder 2 mit geringem Aufwand an die Nutzung von HPC-Systemen herangeführt werden. Die höheren Versorgungsebenen erfordern meist eine Anpassung der Anwendungen und eine intensivere Auseinandersetzung mit den technischen Eigenschaften des jeweiligen existierenden Systems. Der Zugang unterliegt einem wissenschaftlich geführten Beantragungsprozess auf Rechenzeit. Dies ist bei Rechnern der Ebene 3 meist nicht notwendig, da diese für lo-

kale Nutzer unmittelbar offen sind und mit bzw. für die lokalen Nutzer beschafft wurden. Es gibt somit eine Durchlässigkeit von kleineren, lokalen Systemen zu größeren, überregionalen bzw. nationalen Zentren.

Im internationalen Vergleich besitzt Deutschland eine gute bis sehr gute Versorgung mit HPC-Rechnersystemen. Die TOP500-Liste zeigt zweimal im Jahr die jeweils 500 schnellsten bekannten Rechner der Welt [5]. Hier finden sich zahlreiche Rechner aus Deutschland, auch innerhalb der ersten 20. In der historischen Entwicklung ist erkennbar, dass Länder wie China ihre Investitionen in HPC kontinuierlich erhöht haben und die ersten Plätze belegen. In den USA ist die Anzahl an Höchstleistungsrechenzentren zwar leicht gesunken, aber ein stärkerer Fokus auf einzelne, dafür größere Systeme zu sehen. Deutschland schafft es regelmäßig, mit den Systemen des Gauß Centre for Supercomputing einen Platz unter den ersten zehn zu erringen.

Es gibt insgesamt einen wachsenden Bedarf an HPC-Kapazitäten in der Wissenschaft, da sukzessive mehr Anwender und neue Communities diese Systeme benötigen. Dies findet sich im Gegensatz zu einigen asiatischen Staaten jedoch nur eingeschränkt in gesteigerten Investitionen in Europa wieder. Es ist bemerkenswert, dass trotz diverser kommerzieller Cloud-Lösungen es aktuell kaum

externe Angebote gibt, die für die wissenschaftlichen HPC-Bedarfe in Art und Umfang eine echte Alternative darstellen. Dies findet sich auch so im internationalen Vergleich, sodass die wissenschaftlich genutzten HPC-Systeme prädominant an öffentlichen Einrichtungen betrieben werden.

Aufgrund der Empfehlung des Wissenschaftsrates zur „Finanzierung des Nationalen Hoch- und Höchstleistungsrechnens in Deutschland“ [8] wurde auf Ebene der Gemeinsamen Wissenschaftskonferenz (GWK) ein Prozess gestartet, um die Finanzierung insbesondere auf der Ebene 2 künftig neu zu gestalten. Die bisherige projektförmige Finanzierung wird den Anforderungen einer nachhaltigen Infrastrukturplanung bzgl. Personal und Baumaßnahmen nicht gerecht. Der Wissenschaftsrat empfiehlt dazu die Einrichtung von Nationalen Hoch- und Höchstleistungsrechenzentren (NHR) für die Ebene 1 und 2. Diese sollen künftig eine Planungssicherheit für mindestens zehn Jahre erhalten und im Rahmen eines Budgets, das neben Investitionen auch Betriebskosten und Beratungsaufwände abdecken soll, selbst entscheiden, wie die Mittel effizient einzusetzen sind. Eine vergleichbare Struktur findet sich bisher nicht in anderen Ländern und wird den Wissenschaftsstandort Deutschland mit einer logisch weiterentwickelten und langfristig zukunftsfähigen Versorgungsstruktur ausstatten.

Dabei ist zu beachten, dass Wissenschaftler auch weiterhin lokal auf der Ebene 3 Zugang zu ausreichend großen Systemen erhalten sollten, um dort die notwendigen Erfahrungen zu sammeln und an größere Infrastrukturen sukzessive herangeführt werden zu können. Es ist nicht von der Hand zu weisen, dass aufgrund von nationaler oder internationaler Sichtbarkeit politisch gern ein stärkerer Fokus auf einzelne Rechner mit besonders hoher Kapazität zu sehen ist. Wenn man wissenschaftliche Nutzer fragt, präferieren diese hingegen einen Zugang zu Systemen der Ebene 3 oder 2 aufgrund der einfacheren Zugänglichkeit und der direkten Anbindung an Berater. Eine ausgewogene Finanzierung und Ausstattung auf allen drei Ebenen erscheint daher weiterhin sinnvoll und wichtig.

### **Dateninfrastrukturen**

Neben Rechnern spielen wie zuvor beschrieben Daten eine wesentliche Rolle in einer Forschungsinfrastruktur. Die Landkarte hierzu ist wegen der starken Ausdifferenzierung der Datenarten

und Forschungscommunities jedoch wesentlich differenzierter. Bei klassischen Datenarten wie Texten, Bildern und Videos spielen Bibliotheken und Archive eine wichtige Rolle. Sie können auf eine lange Historie in der Kuration und Verfügbarmachung zurückblicken. Insbesondere in den Geisteswissenschaften sind sie zentrale Orte der wissenschaftlichen Arbeit.

Daneben gibt es jedoch zahlreiche Forschungsdatenzentren, die häufig disziplinär ausgerichtet sind und teilweise erst seit einigen Jahren existieren. Im DFG Projekt re3data.org wird ein webbasierter Überblick zu Forschungsdatenrepositorien geliefert, wonach in Deutschland alleine 293 Repositorien (Stand Juli 2017) verzeichnet sind. Die Finanzierung dieser einzelnen Strukturen ist sehr unterschiedlich gestaltet. Einige Einrichtungen sind trotz des Anspruchs auf Langzeitverfügbarkeit von Daten und Diensten nicht nachhaltig verankert und finanziert. Es stellt sich daher die Frage, wie viele Datenzentren notwendig sind, um die wissenschaftlichen Bedarfe abzudecken.

Der Rat für Informationsinfrastrukturen (RfII) wurde von der Gemeinsamen Wissenschaftskonferenz (GWK) im Jahr 2014 eingerichtet, um Empfehlungen zur Weiterentwicklung von Informationsinfrastrukturen auszusprechen. Die Aussagen sollen sich direkt an die Wissenschaft und über die GWK an Bund und Länder richten. Eine Aufgabe ist dabei unter anderem auch die Beobachtung von internationalen Entwicklungen und deren Berücksichtigung bei der Formulierung von deutschen Positionen. Die ersten Veröffentlichungen des Rates betreffen dabei Forschungsdaten im weiten Sinne. In der Empfehlung des RfII „Leistung aus Vielfalt“ [3] werden die Herausforderungen im deutschen Wissenschaftssystem im Umgang mit Forschungsdaten beschrieben und Empfehlungen für eine Nationale Forschungsdateninfrastruktur (NFDI) gegeben. Informationsinfrastrukturen werden dabei definiert als „technisch und organisatorisch vernetzte Dienste und Angebote zur Arbeit mit wissenschaftlich relevanten Daten, Informationen und Wissensbeständen.“ Sie unterstützen den effektiven Umgang mit Forschungsdaten und haben damit eine ermöglichende Rolle („enabling science“). Neben der Reproduzierbarkeit von Forschungsergebnissen spielt für die Wissenschaft meist eine verbesserte Nachnutzung und Vernetzung in Forschungsverbänden eine wichtige Rolle. Vergleichbar mit der

Diskussion zum Nationalen Hoch- und Höchstleistungsrechnen beschäftigt sich die GWK aktuell mit der Frage, wie auf Basis der Empfehlungen des Rates für Informationsinfrastrukturen die NFDI aufgebaut und weiterentwickelt werden kann. Für beide gesamtstaatlichen Aufgaben wollen Bund und Länder noch im Lauf des Jahres 2018 gemeinschaftliche Lösungen beschließen.

Zentrale Herausforderung ist die bedarfsorientierte Entwicklung dieser Forschungsdateninfrastrukturen und ihre nachhaltige Finanzierung. Einige existierende Forschungsdatenzentren haben aufgrund fehlender institutioneller Verankerung keine dauerhafte Finanzierungsperspektive. Durch projektformige Finanzierungen werden Strukturen zwar aufgebaut, es fehlt jedoch regelmäßig ein geordneter Übergang in eine nachhaltige Betriebsphase. Die üblichen Mechanismen der Fördermittelgeber sind auf nachhaltige Finanzierungen typischerweise nicht ausgelegt. Die NFDI soll hier Finanzierungsmechanismen eröffnen, die eine langfristige Planung ermöglichen.

Eine bedarfsorientierte Entwicklung erscheint zwar dringend notwendig, findet sich jedoch in bestehenden Strukturen nur eingeschränkt wieder. Häufig werden zentrale Angebote mit teilweise nur kleinen Nutzergruppen etabliert, während eine breite Akzeptanz nur schwer auszumachen ist. Eine geeignete Balance zwischen angebotsgetriebener und bedarfsorientierter Entwicklung erscheint dringend notwendig, um einen nachhaltigen, effizienten Betrieb mit wissenschaftsgetriebener Dienstentwicklung zu verbinden. Vergleichbar mit dem Höchstleistungsrechnen stellt sich die Frage, wie viele Datenzentren und Infrastruktureinrichtungen benötigt werden. Aufgrund der Digitalität der Informationen ist eine hohe Zentralisierung zwar technisch möglich und ökonomisch sicherlich vorteilhaft, trifft aber meist nicht die Bedürfnisse der Forschenden. Je weiter ein Dienstleistungszentrum für Daten oder Rechnersysteme vom Forschenden entfernt ist und eine möglichst breite Nutzergruppe zu bedienen hat, desto geringer ist naturgemäß der Einfluss des Forschenden auf die Entwicklung dieser Dienste und desto geringer die Akzeptanz. E-Science-orientierte Dienstangebote tun sich schwer, wenn die fachwissenschaftlichen Fragestellungen nicht Teil der Entwicklung sind. Die Forschenden sind hingegen nur eingeschränkt geeignet und interessiert, um selbst einen Servicebetrieb nachhaltig

zu betreiben. Auch wenn viele der derzeit existierenden wissenschaftlichen Datensammlungen ihre Entstehung dem Engagement einzelner Forschender verdanken, können sie an einem Lehrstuhl oder Institut oftmals aus eigener Kraft nur schlecht aufrechterhalten oder gar verstetigt werden. Wenn man berücksichtigt, dass der Großteil der Forschungsdaten beim Forschenden selbst entsteht und damit in den Hochschulen vorliegt, erfordert es ein sensibles Ausrüsten eines vernetzten Systems von durch Fachcommunities spezifizierten Anforderungen und individuellen Diensten, von wenigen definierten Dienstleistungszentren für den Betrieb dieser Dienste und von lokalen Unterstützungsstrukturen nahe beim Forschenden z. B. bei Schulung und Beratung.

### **Blick über den Zaun**

All diese Entwicklungen müssen im nationalen und internationalen Kontext gesehen werden. Immer häufiger findet Forschung nicht mehr allein am Schreibtisch eines einzelnen Wissenschaftlers, sondern in Verbänden statt. Einige Disziplinen sind ohne eine Vernetzung praktisch nicht mehr denkbar. Dies findet sich insbesondere bei Großexperimenten, wie wir es aus der Astrophysik, Teilchenphysik (siehe LIGO, LHC oder FAIR) oder in der Klimaforschung kennen. Großexperimente erfordern aus rein ökonomischen Gründen eine Zusammenarbeit von verschiedenen Forschungsgruppen über Landesgrenzen hinweg. Im Europäischen Raum gibt es zum Beispiel das Europäische Strategieforum für Forschungsinfrastrukturen (ESFRI), das, gegliedert in thematische Bereiche, eine Roadmap für gemeinsame Infrastrukturen etabliert hat. Deutschland ist zwar an zahlreichen ESFRI-Projekten beteiligt, jedoch ist die Abbildung in nationale Strukturen, die eine Anschlussfähigkeit herstellen, sehr heterogen und teilweise unübersichtlich organisiert. Es ist offensichtlich, dass nationale Forschungsinfrastrukturen nicht im luftleeren Raum existieren können, sondern einen Blick auf internationale Entwicklungen benötigen. Mindestens ist eine Abstimmung von fachspezifischen Standards notwendig, idealerweise eine Vernetzung von Diensten anzustreben.

Wenn man bzgl. nationaler Strukturen in das europäische Ausland blickt, sieht man, dass die Herausforderungen überall ähnlich gelagert sind. Der Rat für Informationsinfrastrukturen hat sich aufgrund der Gemeinsamkeiten der wissenschafts-

politischen Ausrichtung und den erkennbaren Parallelen bei der Entwicklung von nationalen Strategien insbesondere mit Kanada, den Niederlanden, dem Vereinigten Königreich und Australien beschäftigt [4]. Alle vier Länder haben bereits nationale Strukturen entwickelt, sowohl für Hochleistungsrechnen als auch für Dateninfrastrukturen. Es ist nicht überraschend, dass es Ländern wie den Niederlanden und Australien mit einer geringeren Anzahl von Stakeholdern leichter fiel, nationale Strukturen zu etablieren. Die stark durch Föderalismus geprägte Aufteilung in bundes- und landesspezifische Strategien führte in Deutschland zu einer heterogenen und diffusen Entwicklung von übergreifenden Strukturen. In den Niederlanden und im Vereinigten Königreich sind insbesondere die Infrastruktureinrichtungen SURF und Jisc im Bereich des Forschungsdatenmanagements zu nennen. Beide Einrichtungen betreiben diverse nationale Dienstleistungen und haben spezifische Angebote für Datenmanagement etabliert (z. B. das Digital Curation Centre von Jisc oder das Netherlands eScience Centre bei SURF). Auch in den genannten Ländern gibt es Friktionen durch die meist Bottom-up-geleiteten Entwicklungen der Wissenschaft und die eher Top-down-gesteuerte Verteilung von Infrastrukturressourcen. Alle genannten Länder haben aktuell einen vergleichbaren politischen Prozess, der über die jeweilige Gesamtkonstruktion kritisch reflektiert und eine Neubewertung anstellt. So versucht Australien, diverse existierende Strukturen in einer Australian Research Data Cloud zusammenzufassen. Es wird deutlich, dass die genannten Länder in einigen Bereichen Deutschland zwar voraus sind, jedoch weiterhin mit grundlegenden Problemen der adäquaten Finanzierung, geeigneter Evaluationsmechanismen und bedarfsorientierter Weiterentwicklung beschäftigt sind. Allen Strukturen gemein ist die Herausforderung, die Dienste näher an die einzelnen Forschenden heranzuführen und diese geeignet einzubinden. Auch dort finden sich Defizite bei Akzeptanz und Bewusstseinsbildung der Forschenden im Umgang mit solchen Strukturen.

Europa bemüht sich seit zwei Jahren in der inhaltlichen Ausgestaltung der European Open Science Cloud, wobei auch hier ein eher Top-down- und infrastrukturgesteuerter Prozess erkennbar ist, der ein Netzwerk von nationalen Strukturen erfordert. Es bleibt abzuwarten, ob dies letztlich

von Erfolg gekrönt sein wird und ob der Nutzen wirklich in der Wissenschaft ankommt bzw. dort akzeptiert wird. Bisher waren die zentral angestoßenen Aktivitäten der Europäischen Kommission nur eingeschränkt erfolgreich, wenn die Strukturen nicht fokussiert auf definierte Forschungsdisziplinen waren. Die Erfahrungen aus den ESFRI-Projekten zeigen auch, dass diese Strukturen langfristig angelegt sein müssen und teilweise erst nach zehn Jahren echte Erfolge erkennbar sind. In Zeiten von politisch definierten Fünf-Jahres-Programmen erfordert dies Weitblick und einen langen Atem.

### **Die Rolle der Wirtschaft**

Sowohl in der deutschen Forschungslandschaft als auch im internationalen Vergleich ist die Rolle der privaten Wirtschaft im Bereich der Infrastrukturen wenig ausgeprägt. Das ist überraschend, wenn man bedenkt, dass im privaten Sektor enorme Datenmengen gesammelt werden (z. B. Sensordaten, soziale Netzwerke, Bewegungsdaten mobiler Geräte oder medizinisch relevante Informationen aus dem Consumermarkt). Ebenso ist allgemein bekannt, welche enormen Rechnerkapazitäten in kommerziellen Cloudstrukturen verfügbar geworden sind. Daher wäre es naheliegend, eine Vernetzung auch bei Forschungsinfrastrukturen intensiver zu betrachten. Im Bereich der Publikationen haben Verlage wissenschaftlich relevante Bestände, wobei der eskalierte Kampf um faire und wissenschaftsadäquate Lizenzmodelle den Grundkonflikt aufzeigt. Es besteht größeres Misstrauen in der Wissenschaft, inwieweit eine nachhaltige und verlässliche Basis der Zusammenarbeit mit den kommerziellen Interessen der Firmen vereinbar ist. Die Datenhoheit bei meist öffentlich geförderter Forschung im Bereich der Wissenschaft sowie eine langfristige Handlungsautonomie führen bisher zu separaten Strukturen, die nur in einzelnen Forschungsprojekten zueinander finden. Ob dies dauerhaft aufrecht zu erhalten sein wird, bleibt abzuwarten.

### **Beratung und Ausbildung**

Allen Strukturen ist gemein, dass diese nur so gut genutzt werden können, wie die Forschenden das notwendige Wissen im Umgang mit den Methoden und Ressourcen haben. Die Ausprägung einer Datenkultur in der Wissenschaft steht noch am Anfang

und erfordert intensive Beschäftigung in den Fachgesellschaften und auch an den Hochschulen. Es besteht die Notwendigkeit, Ausbildungs- und Trainingsprogramme (Studiengänge) zu entwickeln, was bereits an zahlreichen Hochschulstandorten erkannt wurde. Hierzu gehören Ausbildungsmodule im Umgang mit Daten im Bereich der Kuration, Langzeitarchivierung und Publikation, aber auch neue Lehre im Bereich der Datenauswertung und des Umgangs mit den zugehörigen Infrastrukturressourcen, wozu auch Großgeräte wie Hochleistungsrechner und Datenanalyseplattformen zählen. An einigen deutschen Universitäten hat dies zu eigenen Studiengängen geführt, die im weitesten Sinne mit Data Science beschrieben werden können und eine Kombination aus Informatik, Statistik, Mathematik und anwendungsbezogenen Komponenten sind. Dabei sind sicherlich die Anforderungen unterschiedlich, je nachdem, ob man Anwender bzw. Nutzer dieser Methoden und Dienste ist oder ob eine innovative Entwicklung neuer Methoden erwartet wird. Die bestehenden Entwicklungen an den Hochschulen zu solchen neuen Studienprogrammen zeigen die funktionierende Eigensteuerung als Reaktion auf neue gesellschaftlich und wissenschaftlich relevante Anforderungen. Es ist bereits absehbar, dass der Bedarf an diesen Kompetenzen enorm ist und für viele Jahre ein Engpass in der Entwicklung sein wird, geeignetes Personal zu finden. Hier steht die Wissenschaft in Konkurrenz mit den Anforderungen aus der Wirtschaft, die ähnliche Kompetenzprofile sucht. Die bekannten Defizite in Entlohnung und Entwicklungsperspektiven werden dabei ein erfolgskritischer Faktor für die Konkurrenzfähigkeit der deutschen Forschungslandschaft sein. Ebenso wird es entscheidend sein, den Wandel zur daten- und rechnergetriebenen Wissenschaft möglichst schnell und adäquat für sich umzusetzen. Hier kann man optimistisch sein, dass die Selbststeuerung der Wissenschaft solche Entwicklungen geeignet aufnimmt. Bisher sind neue Forschungsmethoden, die Vorteile im wissenschaftlichen Erkenntnisprozess versprechen oder gezeigt haben, recht schnell aufgenommen und umgesetzt worden. Ähnlich wird es auch diesmal sein, wenn einzelne Forschungsdisziplinen die weitergehenden Möglichkeiten im rechner- und datengestützten Umgang für sich entdecken. Die Frage wird sein, wer diese Entwicklungen frühzeitig umsetzt und damit einen Vorteil in der eigenen Forschungsprogrammierung erzielt.

## **Ausblick zu E-Science-Infrastrukturen**

Wie dargestellt, befindet sich die Ausgestaltung von forschungsunterstützenden Infrastrukturen aktuell in der Diskussion und wahrscheinlich an einem Wendepunkt in der künftigen Förderstruktur. Dabei werden Anstrengungen unternommen, um einige und teilweise langjährig bekannte Defizite zu adressieren. Dabei gab es in Deutschland eigentlich nie ein Erkenntnisproblem, wo Handlungsbedarfe in der Infrastrukturentwicklung bestehen. Dies zeigen die diversen Publikationen, insbesondere von Wissenschaftsrat, Hochschulrektorenkonferenz oder Gemeinsamer Wissenschaftskonferenz. Stattdessen gab es vielmehr ein Umsetzungsproblem, um in der föderalen Struktur mit diversen Fördermittelgebern neue Modelle zu finden, die eine Vernetzung über Landes- und Disziplinergrenzen in nationalen und internationalen Strukturen ermöglichen.

Auch wenn nun in den kommenden Jahren hoffentlich mit neuen Mechanismen zu rechnen ist, handelt es sich weiterhin um einen dynamischen und stetig weiter zu entwickelnden Prozess. Die aktuellen Anstrengungen laufen auf die Entwicklung von nationalen Vernetzungsstrukturen hinaus, in denen sich Fachcommunities zusammenfinden und selbst organisieren können, und dabei mit geeigneten Infrastrukturanbietern zusammenarbeiten. Dabei wird häufig noch wenig betrachtet, dass Forschungsdateninfrastrukturen nicht unabhängig von anderen Infrastrukturen zu sehen sind. Alle Aussagen treffen in ähnlicher Form, eventuell mit weniger Komplexität aufgrund einer geringeren Anzahl an Zentren, auf Hoch- und Höchstleistungsrechnerstrukturen zu; ebenso gibt es Entwicklungen zu vernetzten Strukturen für Zentren, bei denen Forschungsgroßgeräte effizient gemeinsam genutzt werden sollen, wie dies z. B. für Hochdurchsatzsequenzierung (NGS), Elektronenmikroskopie, Bioinformatik etc. gilt [2]. Auch wenn Daten prinzipiell zwar transferiert werden können, zeigt sich im Alltag, dass große Datenmengen zunehmend immobil sind und eine Verarbeitung eher bei den Daten stattfindet, als die Daten zur Bearbeitung zu transportieren. Am Ende wird man daher feststellen, dass diese unterschiedlichen Infrastrukturen logisch und bedarfsorientiert zusammenwachsen müssen, sodass Datengewinnung, -analyse, -speicherung und -archivierung einfach und bruchfrei möglich sind. In technischer Analogie würde man hier Overlay-Netzwerke als sinnvolles Ergebnis vermuten, in

denen Fachcommunities die für sich notwendigen Dienste und Strukturen definieren, wobei sie auf die bestehendem Infrastrukturzentren zurückgreifen, die diese Anforderungen am besten erfüllen und dabei eine horizontale Vernetzung und Synergien über Fachdisziplinen erreichen. Dass Fachcommunities immer wieder völlig neue Zentren oder Dienste entwickeln, wird sich ökonomisch nicht vertreten lassen. Durch einen Wettbewerb der bestehenden Infrastruktureinrichtungen als potenzielle Partner für den Betrieb von Diensten lassen sich marktorientierte Mechanismen etablieren, die verhindern, dass am Bedarf vorbei Angebote geschaffen werden. Bei der Etablierung oder Auswahl von Gerätezentren wäre daher als Kriterium zu prüfen, welche Strukturen und Vernetzungen diese für Datenverarbeitung und -archivierung bereits mitbringen bzw. in Kooperationen nachnutzen können. Für Bibliotheken, Archive und Rechenzentren bedeutet dies enge Kooperationen mit den Forschenden. Für Hochschulen bedeutet dies eine individuelle Entwicklung einer Infrastrukturstrategie, bei der lokale Beratungsangebote mit eigenen und nicht

lokalen, übergreifenden Infrastrukturlösungen zusammengreifen. Hier bieten sich Felder für neue Kooperationsmodelle, um im Rahmen von begrenzten finanziellen Ressourcen den eigenen Forschern eine effektive und effiziente Forschungsumgebung zu liefern.

## Literatur

1. Hey T, Tansley S, Tolle K (eds) The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research 2010, <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/>
2. Leopoldina (2014) Zukunftsreport Wissenschaft zu Herausforderungen der Omics-Technologien für Deutschlands Infrastrukturen in Forschung und Lehre. ISBN: 978-3-8047-3283-4
3. Rfll-Empfehlung (2016) Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland. Juni 2016, urn:nbn:de:101:1-201606229098
4. Rfll-Fachbericht (2017) Entwicklung von Forschungsdateninfrastrukturen im internationalen Vergleich. Juli 2017
5. TOP500 Supercomputer Sites (2017) 49th TOP500 at ISC 2017 in Frankfurt, Juni 2017, <https://www.top500.org/lists/>
6. Wissenschaftsrat (1995) Empfehlung zur Versorgung von Wissenschaft und Forschung mit Höchstleistungsrechenkapazität. In: Wissenschaftsrat: Empfehlungen zur Ausstattung der Wissenschaft mit moderner Rechner- und Kommunikationstechnologie. Wissenschaftsrat, Köln, S 51–70
7. Wissenschaftsrat (2012) Strategische Weiterentwicklung des Hoch- und Höchstleistungsrechnens in Deutschland. Positionspapier, Berlin, Drs. 1838-12
8. Wissenschaftsrat (2015) Empfehlungen zur Finanzierung des Nationalen Hoch- und Höchstleistungsrechnens in Deutschland. Stuttgart, Drs. 4488-15, April 2015, <https://www.wissenschaftsrat.de/download/archiv/4488-15.pdf>

# *e-science in den Geo- und Umweltwissenschaften*

*Ansätze und Anforderungen*

Lars Bernard · Stephan Mäs

## **Digitale Geo- und Umweltwissenschaften**

Die Geo- und Umweltwissenschaften zählen sicherlich zu den Disziplinen, in denen die Digitalisierung nicht nur Einzug gehalten hat, sondern bereits in weiten Teilen etabliert ist. Forschungsarbeiten in fast allen Bereichen und über den gesamten Forschungszyklus erfolgen in großem Umfang digital bzw. digital unterstützt. Die satelliten- und flugzeuggestützte Fernerkundung der Erdoberfläche ist beispielsweise als Beobachtungsverfahren bereits sehr früh vollständig digitalisiert erfolgt und hat die weitere Digitalisierung stimuliert. Heutige Satellitenmissionen liefern Forschern digitale Daten in erheblichem Umfang: Allein das erst seit 2014 operationelle und noch im Aufbau befindliche Europäische Satellitenprogramm Copernicus<sup>1</sup> hat Wissenschaftlern Ende 2017 monatlich rund 300 Terrabytes neue Erdbeobachtungsdaten kostenfrei zur Verfügung gestellt und verzeichnet monatliche Downloads im Umfang vom 2,7 Petabytes<sup>2</sup> – bei stark ansteigendem Trend der Datenbereitstellungen und -zugriffe. Auch an der Erdoberfläche (*in situ*) werden Umweltbeobachtungen heute in großem Maße direkt digital erfasst und vielfach über Sensornetze den Forschern in (nahe) Echtzeit zur Verfügung gestellt. Neue Applikationen die für Umweltbeobachtungen die Zeitreihendaten unterschiedlicher Sensornetzbetreiber und für unterschiedliche Parameter zusammenführen (Abb. 1), zeigen einerseits eindrucksvoll die Breite und Umfänge der verfügbaren Sensordaten und skizzieren andererseits das künftige Anwendungspotenzial für den integrierten Zugriff auf unterschiedliche Umweltsensordaten.

<sup>1</sup> <http://www.copernicus.eu/>.

<sup>2</sup> <https://scihub.copernicus.eu/reportsandstats/>.

Seit mehr als 30 Jahren werden Geoinformationssysteme (GIS) als DV-Instrumentarium der Geo- und Umweltwissenschaften eingesetzt, um raum- und zeitvariante Daten zu analysieren. Methoden zur Analyse und Simulation von beispielsweise atmosphärischen, hydrologischen oder sozioökonomischen Prozessen und deren Interaktionen basieren auf verschiedenen Simulationsstrategien (statistisch, numerisch, agentenbasiert, etc.), sind teilweise sehr rechenintensiv oder/und datenintensiv [19, 23]. Räumliche Analysen werden durch Funktionalitäten heutiger GIS umfassend unterstützt (z. B. geometrisch-topologische Analysen, Netzwerkanalysen, morphologische Analysen, geostatistische Verfahren, multikriterielle Bewertungen). Komplexere raum-zeitliche Simulationsanwendungen (z. B. numerische Strömungsmodelle, agentenbasierte Verfahren zur Populationsentwicklung) sind typischerweise lose für die Vorbereitung der Eingangsdaten bzw. Analyse der Ergebnisdaten an GIS gekoppelt.

Auch wenn sich bereits zahlreiche Ansätze für die Etablierung eines geregelten und nachhaltigen Forschungsdatenmanagements einerseits und andererseits für ein – im Sinne von e-science – kollaboratives Arbeiten in den Geo- und Umweltwissenschaften finden, existieren noch viele Herausforderungen. Entlang eines idealisierten und

---

<https://doi.org/10.1007/s00287-018-01138-x>  
© Springer-Verlag Berlin Heidelberg 2018

---

Lars Bernard · Stephan Mäs  
Technische Universität Dresden,  
Fakultät Umweltwissenschaften, Professur für Geoinformatik,  
Helmholtzstraße 10, 01069 Dresden  
E-Mail: [lars.bernard@tu-dresden.de](mailto:lars.bernard@tu-dresden.de)



Abb. 1 Beispiel einer Applikation, die einen integrierten und performanten Zugriff auf die Monitoring-Zeitreihen von weltweit rund 400.000 unterschiedlichen Umweltbeobachtungssensoren erlaubt (<https://opensensorweb.de>)

## Zusammenfassung

Die Geo- und Umweltwissenschaften arbeiten heute in großem Maße und entlang des gesamten Forschungsprozesses digital: Von der Beobachtung, Erfassung und kontinuierlichen Messung der Umweltphänomene mit unterschiedlicher Sensorik über die Analyse dieser Beobachtungsdaten und Erstellung von Modellen und Simulationen zum Verständnis oder der Vorhersage von Umweltprozessen, der Synthese von Daten verschiedener Quellen zur Beschreibung und Bewertung von Umweltveränderungen bis hin zur Publikation der Ergebnisse als wissenschaftliche Beiträge, Empfehlungen oder interaktive Visualisierungen. Allerdings gibt es viele Brüche in den Arbeitsprozessen und vielfache Hemmnisse für ein kollaboratives, auch interdisziplinäres wissenschaftliches Arbeiten im Sinne von e-science. Hier skizziert der Beitrag aktuelle Entwicklungsstände, derzeitige Forschungsansätze und verbleibende Herausforderungen für Informatik und Geoinformatik.

vereinfachten Forschungszyklus (*Datenakquise und -integration – Analyse und Modellierung – Ergebnispublikation*) versucht dieser Beitrag schlaglichtartig für die Aspekte Datenaustausch sowie Austausch

von Analyse- und Modellierungsmethoden, einige aktuelle Ansätze und Forschungsfragen für die (Geo-)Informatik in diesem Kontext darzustellen.

## Austausch, Exploration und Integration von Geo- und Umweltdaten

Anbieter (und Nutzer) von Geo- und Umweltdaten sind, neben der Wissenschaft selbst, in hohem Maße die Verwaltung und kommerzielle Anbieter, aber auch freiwillige und bürgerwissenschaftliche Initiativen [2]. Hier haben speziell die Aarhus-Konvention der UNEC<sup>3</sup> zur öffentlichen Bereitstellung von Umweltinformation und die INSPIRE EU-Direktive<sup>4</sup> zum Aufbau einer Europäischen Informationsinfrastruktur für raumbezogene Umweltinformationen dazu geführt, dass heute zahlreiche und vielfältige behördliche Umweltdaten online und immer häufiger frei verfügbar sind. Auch die Zahl der aus Forschungskampagnen verfügbaren Daten steigt beständig: Das Verzeichnis der Forschungsdatenrepositorien re3data.org listet für das Schlagwort *environment* mehr als 200 Repositorien auf, von denen wiederum allein das Repository PANGAEA<sup>5</sup> aktuell mehr als 370.000 Datensätze aus mehr als 12 Mio. Beobachtungen enthält. Gleichzeitig finden sich an vielen Stellen Initiativen, Anstöße und

<sup>3</sup> <http://www.unece.org/env/pp/welcome.html>.

<sup>4</sup> <https://inspire.ec.europa.eu/>.

<sup>5</sup> <https://www.pangaea.de/>.

## Abstract

Today's work in geosciences and environmental sciences is to a great extent digital, all along the research cycle. Starting from observations and continuous measurements of environmental phenomena using various sensor techniques, this holds along the subsequent data analysis and development of models and simulation runs to understand and forecast environmental processes, over the synthesis of data from various sources to describe and assess environmental changes towards the publication of the results as scientific papers, recommendations or interactive visualizations. However, there are several discontinuities in the related working processes as well as obstacles towards collaborative, also interdisciplinary, scientific work as envisioned by e-science. In this context, the paper discusses the status, related research approaches and remaining challenges for Computer Science and Geoinformatics.

*best practices* – nicht nur für das gesamte Wissenschaftssystem [17, 29], sondern auch konkret für die Geo- und Umweltwissenschaften [3, 11] – zur weiteren Entwicklung des Forschungsdatenmanagements und zur Etablierung des Austauschs und der Nachnutzung von Forschungsdaten.

## Datenexploration

Dank hinreichend konsolidierter und etablierter Metadatenstandards sind bereitgestellte amtliche Daten und die in den Repositorien verfügbaren Forschungsdaten in der Regel derart beschrieben, dass eine erfolgreiche Datenrecherche auch über mehrere, verteilte Kataloge möglich ist. Aktuelle Arbeiten zeigen, wie diese Metadatenstandards darüber hinaus genutzt werden können, um für wissenschaftliche Datenrecherchen nicht nur die raumzeitliche und thematische Abdeckung, sondern auch die Entstehungsgeschichte der Daten (*provenance*) zu beschreiben und auf den Daten basierende relevante wissenschaftliche Beiträge zu referenzieren (Abb. 2).

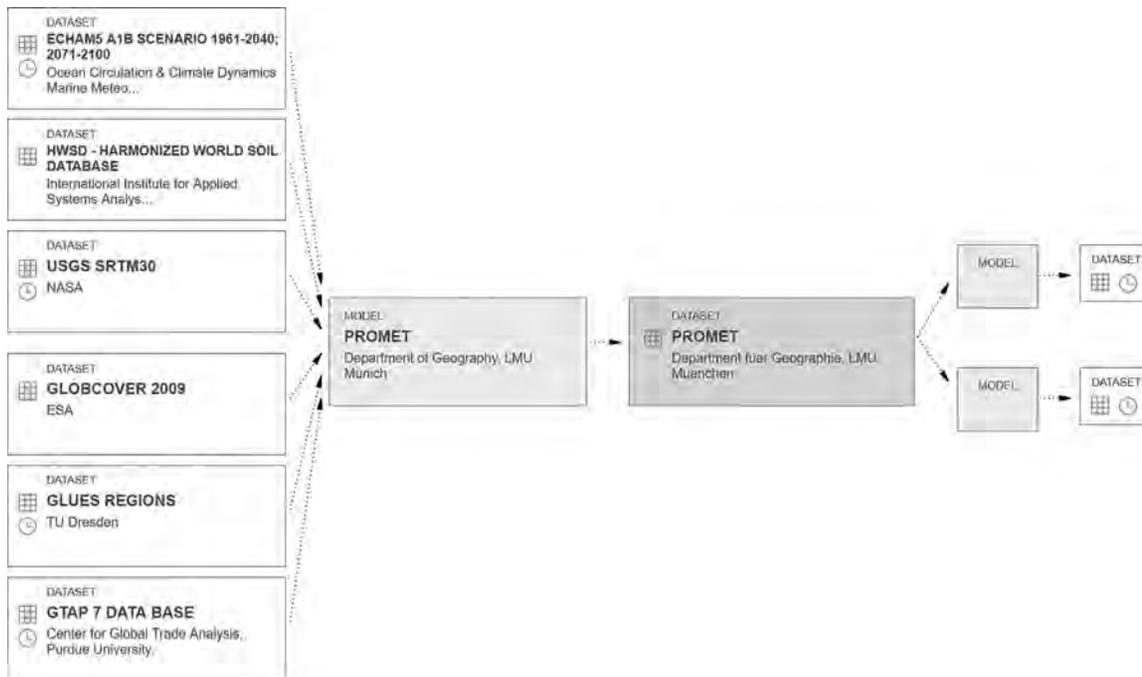
Allerdings finden sich für diese von Wissenschaftlern als sehr nützlich eingestuft und positiv aufgenommenen Anwendungen in der Regel in den aktuellen Metadatenbeständen nicht ausreichend

Inhalte. Die Eignung bzw. die Limitierungen der recherchierten Daten bezüglich der eigenen Anwendungszwecke können daher selten befriedigend beurteilt werden. Das Problem scheint hier weniger das Fehlen von Metadatenstandards – derer gibt es sicher ausreichend viele, die für die verschiedenen Anwendungszwecke einsetzbar sind. Es fehlt hingegen einerseits an vielen Stellen sicher noch ein echter Kulturwandel in Form von Erkenntnis und Willen, Daten zu publizieren und auch umfassende und hilfreiche Metainformationen bereitzustellen. Andererseits sind die manuellen Annotationsaufwände zur Erstellung von Metadaten noch deutlich zu hoch und es mangelt in den datenführenden Systemen an *smarten* Ansätzen zur weitgehend automatisierten Ableitung und Nachführung der benötigten Metadaten.

Hier setzen Arbeiten an, die darauf zielen das Forschungsdatenmanagement von der Erfassung und Simulation über die Forschungsdatenpublikation, Datenqualifikation bis zur Bereitstellung von qualifizierten Daten möglichst umfassend und bruchfrei zu unterstützen. Erste prototypische Lösungen [5] bieten Plattformen für den internen Datenaustausch in verteilten Forschungsteams, die aus einer Modellierung der Analyse- und Datenflüsse entsprechende Datenstrukturen automatisiert ableiten, die Integration externer Umweltdatendienste erlauben und Werkzeuge zur Qualifizierung der Forschungsergebnisse sowie für die Publikation als standardisierte Webdienste anbieten. Eine solche durchgängige Unterstützung des Forschungsdatenmanagements erlaubt es, Metadaten weitgehend automatisiert aus den verschiedenen Annotationen und durchgeführten Prozessierungen zu ermitteln und zu aktualisieren. Die Erstellung der Metadaten und Qualifizierung der Daten zu publizierbaren Ergebnissen ist so nicht mehr von der Datenbearbeitung entkoppelt und nicht länger leidige – gern vergessene – Pflichtaufgabe am Ende eines Forschungsprojekts, sondern ein von Beginn an mitlaufender Prozess.

## Datenintegration

Die in den verfügbaren Daten beschriebenen Phänomene reichen über zahlreiche räumliche, zeitliche und thematische Skalen. Qualitäten, Strukturen und Semantiken der erhobenen, genutzten und abgeleiteten Daten differieren ebenso wie die Nutzungsbedingungen. Die für die Geo- und Um-



## LINEAGE

### PROCESS DESCRIPTION

RATIONALE not set  
 DESCRIPTION Simulation of potential yield on representative sampling units in a Glues Region for 16 Glues-Plants

### PROCESS PARAMS

TIME OF EXECUTION 2012-01-01T00:00:00  
 SOFTWARE REFERENCE not defined  
 PROCESSOR OF DATA Department of Geography, LMU Munich

## PUBLICATION

### 1. PUBLICATION

TITLE OF PUBLICATION  
 Mausler, W., Bach, H. (2009):  
 PROMET - Large scale  
 distributed hydrological  
 modelling to study the impact  
 of climate change on the  
 water flows of mountain  
 watersheds, Journal of  
 Hydrology.  
 PUBLISHING DATE  
 2009-07-22  
 CITATION  
<http://www.sciencedirect.com/science/article/pii/S0022169409004478>

**Abb. 2 Beispiel der Darstellung der Entstehungsgeschichte und Verwendung von Umweltdaten unter Nutzung standardisierter Metadaten [11]**

weltforschung betrachteten Daten können im Sinne von *volume* und *variety* sicherlich als *Big Data* gelten, in der Forschung beispielsweise zu Naturkatastrophen geht es oft auch um *velocity* im Kontext der Datenerfassung, -übertragung und -analyse. Zur Schaffung von Interoperabilität und zur Harmonisierung der Datenmodelle für Geo- und Umweltdaten finden sich gesetzliche Rahmenbedingung, internationale Normen und zahlreiche De-facto-Standards, die sich in unterschiedlicher

Weise und verschieden weitreichend durchgesetzt haben [2].

Aber auch wenn neben den existierenden Regelungen und Standards die Verortung in Raum und Zeit zunächst einmal einen Referenzrahmen liefern, stellt die Zusammenführung von Daten unterschiedlicher Quellen immer noch erhebliche Herausforderungen. In Forschungsanwendungen werden solche Datenfusionen beispielsweise notwendig, um räumlich oder zeitlich benachbarte

Daten zu einem Datensatz mit einer möglichst vollständigen Abdeckung zusammenzuführen, die Datensätze für ein Untersuchungsgebiet thematisch anzureichern, Daten zu aktualisieren und um Veränderungen, Unterschiede oder mögliche Fehler zu erkennen. Die Aufwendungen für die notwendigen Datenkuratierungen, Schematransformationen und Datenfusionierung sind heute meist erheblich. Die Prozesse dazu sind oft nicht oder nur begrenzt automatisiert. Diese Problematik adressieren aktuelle Forschungsarbeiten, die zeigen, wie unter Verwendung von automatisiert abgeleiteten geometrischen, topologischen und semantischen Ähnlichkeitsmaßen Vorschläge für Fusionspfade zwischen zwei oder mehreren Datensätzen abgeleitet werden können. Solche Fusionspfade können dann beispielsweise in Linked-Data-Strukturen abgelegt und so für weitere Abfragen zugänglich gemacht werden [25, 26]. Die explizierte Bereitstellung der Ähnlichkeitsmaße und Fusionspfade schafft insbesondere die Möglichkeit, ad-hoc und transparent Daten für eigene Anwendungszwecke zu integrieren. Weitere aktuelle Forschungsarbeiten zeigen, wie weit aus Fernerkundungsdaten automatisiert Informationen zu Objekten der Erdoberfläche abgeleitet werden können [8]. Die bis dato wenig genutzten Verfahren des maschinellen Lernens finden in Anbetracht der heute verfügbaren großen Menge geeigneter *Lern-daten* zur Objekterkennung in Fernerkundungsdaten eine immer maßgeblichere Rolle.

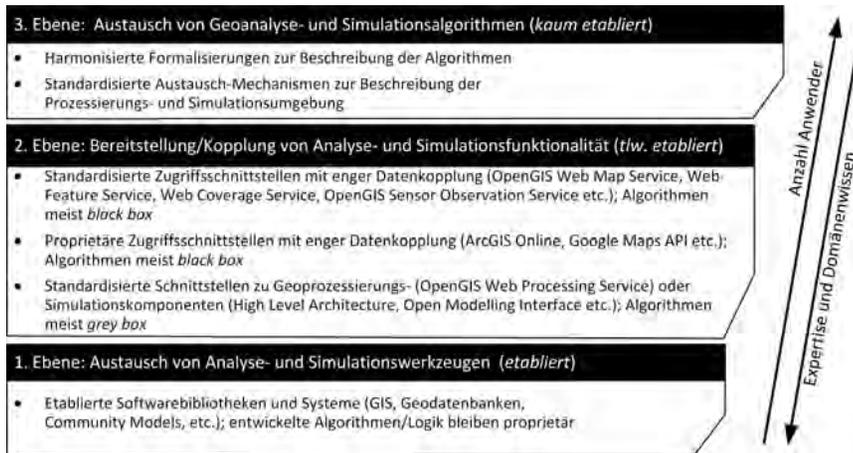
Künftige Forschungsarbeiten sollten hier auf eine weitere Zusammenführung der unterschiedlichen Verfahren zielen. So gilt es beispielsweise, die verschiedenen Datenfusionsansätze derart zu kombinieren, dass Attribute eines Geoobjekts automatisiert und etwa unter Einbindung neuer Sensordatenströme ergänzt, aktualisiert bzw. auf Veränderungen untersucht werden können. Darüber hinaus sollten solche Verfahren auch (automatisiert) immer Unsicherheitsmaße und Information zur Datengeschichte und -genese liefern können.

### **Kollaborative raumzeitliche Analysen und Umweltmodellierungen**

Wang et al. [24] haben *CyberGIS* als eine Bezeichnung für die aktuell im Kontext von Cloud-Umgebungen und Big-Data-Anwendungen entstehende neue Generation von GIS eingeführt, die neben die etablierten desktop- und serverbasierten GIS sowie zu den Programmierschnittstellen für

webbasierte Geoinformationsangebote treten sollen. Unter Nutzung von Cyberinfrastructures [15] als Kombination von Hochleistungsrechnern, gemeinsamen großen Datenrepositorien, Visualisierungsumgebungen, hoch performanten Netzumgebungen sollen Domänenexperten und Entwicklern solche *CyberGIS* als geeignete Umgebungen bereitgestellt werden, um raum-zeitliche Analysen auch für sehr große – bis dato nicht effizient nutzbare – Datenmengen bzw. Berechnungsumfänge durchführen zu können. Beispiele sind hier Analysen auf den immer umfangreicheren Fernerkundungsdaten oder die zusammenhängende Berechnung von hydrologischen Parametern in hoher zeitlicher und räumlicher Auflösung für große Untersuchungsgebiete und in (nahe) Echtzeit für zeitkritische Hochwasserwarnungen. Insbesondere erlauben *CyberGIS*, nicht nur komplexere und umfangreichere Analyseberechnungen durchzuführen, sondern dafür auch unter Einsatz etwa von Monte-Carlo-Ansätzen in nahe Echtzeit Unsicherheitsmaße abzuleiten, für die bisherige Systeme nicht ausreichend Rechenleistung aufbieten können [7]. Gerade für Vorhersagesysteme und entscheidungsunterstützende Systeme ist die Bereitstellung solcher Unsicherheitsmaße eine sehr wichtige und wertvolle Erweiterung. Um Hochleistungsrechner und die vernetzten Rechenressourcen effizient einzusetzen, sollten die Anwendungen parallelisierbar sein. Teilweise lassen sich Algorithmen bereits durch eine (räumliche) Zerlegung der Eingangsdaten einfach auf verteilte Ressourcen aufteilen – etwa bei der Anwendung lokaler wirkender Operatoren. Für komplexere dynamische Simulationen stellen sich jedoch umfangreichere und zum Teil nicht triviale Anforderungen zur Parallelisierbarkeit der Algorithmen.

Zur weiteren Entwicklung von *CyberGIS* stellen sich viele Forschungsfragen. Diese betreffen generelle Themen wie beispielsweise Cybersecurity mit dem Ziel der Sicherstellung von Informations- und Systemsicherheit sowie Lizenz- und Datenschutz in den verteilten Anwendungen. Dabei gibt es in Anwendungsbereichen wie der Umweltmedizin und der Auswertung von Bewegungsprofilen oder sozioökonomischen Daten aktuell häufig einen Zielkonflikt zwischen dem Schutz personenbezogener Daten und effizienten und umfassenden wissenschaftlichen Analysen, der sich derzeit nur selten zufriedenstellen lösen lässt. Eine zukünftige Formalisierung von (ge-



**Abb. 3**  
**Interoperabilitätsebenen für den Austausch von Geoanalyse- und Umweltsimulationsalgorithmen.**  
(Verändert nach [2])

setzlichen) Datenschutzbestimmungen derart, dass diese als abrufbare und prüfbare Regeln für Analysesoftware bereitstehen, kann hier ein Weg sein, um eine automatisierte Durchsetzung eines Datenschutzes zu ermöglichen und die Anwendungseffizienz nicht über Gebühr einzuschränken.

Die Schaffung von Interoperabilität für verteilte GIS allgemein und im speziellen Fall CyberGIS ist ein zentraler Aspekt zahlreicher Ansätze und Forschungsarbeiten in der Geoinformatik. Ziel ist dabei, Grundlagen zu schaffen, die es erlauben, Simulations- und Geoprozessierungsalgorithmen auszutauschen, vielfältig für eigene Arbeitsabläufe und über idealerweise Domänengrenzen hinweg zu kombinieren und systemunabhängig und transparent zu beschreiben. Grob lassen sich hier nach Grad der erreichten Interoperabilität und Grad der Etablierung drei Ebenen unterscheiden (Abb. 3).

Die erste Ebene umfasst den traditionell und auch heute bei Weitem überwiegenden Austausch von Analyse- und Simulationswerkzeugen über die gemeinsame Entwicklung und Nutzung von meist als Open Source zur Verfügung gestellten Softwarebibliotheken (z. B. ArcGIS, GRASS GIS, PostGIS, QGIS) oder Community Models (z. B. Community Earth System Model<sup>6</sup>, OpenGeoSys<sup>7</sup>). Allerdings beschränkt sich die Nutzung – und der Austausch – dieser Systeme meist auf Nutzer mit einem vertieften Fachwissen und weit überdurchschnittlicher Softwareentwicklungskompetenz. Diese Systeme werden in der Regel nicht fachübergreifend genutzt und die

verwendeten Algorithmen sind zwar idealerweise in Dokumentationen beschrieben, aber dennoch proprietär.

Eine zweite Ebene machen Ansätze zur Bereitstellung und Kopplung von Analyse- und Simulationsfunktionalität aus. Webbasierte Dienst-schnittstellen erleichtern die Nutzung, auch über Domänengrenzen hinweg und helfen, größere Anwendergruppen zu erreichen. Zahlreiche etablierte Angebote stellen Umweltdaten und darauf aufsetzende und eng an die Daten gekoppelte Visualisierungs- und Abfragefunktionen über standardisierte Schnittstellen (z. B. OpenGIS Web Map Service, Web Feature Service, OpenGIS Web Coverage Service; vgl. [2]) oder proprietäre Schnittstellen zur Verfügung (z. B. Google Maps API, ArcGIS Online). Die Nutzung der Analysefunktionalität für eigene Daten ist jedoch nur begrenzt oder überhaupt nicht möglich. Über diese Schnittstellen werden auch Ergebnisse aus (vorab ausgeführten) Simulationsläufen verfügbar gemacht (z. B. des International Panel on Climate Change, IPCC<sup>8</sup>) oder ein robustes und in der Regel vorparametrisiertes Modell erlaubt eine Ad-hoc-Ausführung, etwa zur Abschätzung der Luftverschmutzung unter Nutzung aktueller Wetter- und Luftschadstoffdaten [20, 28]. Weniger etabliert, aber dennoch in einigen Ansätzen zu finden, sind Nutzungen standardisierter Schnittstellen, um eine Kopplung unterschiedlicher Geoprozessierungsfunktionalitäten und/oder Simulationskomponenten auch in loser Kopplung an die eingehenden Daten zu ermöglichen. Der sehr generische Schnittstellen-

<sup>6</sup> <http://www.cesm.ucar.edu/>.

<sup>7</sup> <http://www.opengeosys.org/>.

<sup>8</sup> <http://www.ipcc-data.org/>.

standard des *OpenGIS Web Processing Service* erlaubt es, Geoprozessierungsfunktionalität – z. B. zur raum-zeitlichen Verschneidung von Umweltdaten oder zur räumlichen Interpolation von punkthaft erhobenen Messwerten – als Webdienst bereitzustellen [2]. Die Funktionalität wird dabei im Sinne einer *grey box* gekapselt, die Spezifika des Algorithmus werden also freitextlich, nicht jedoch formal beschrieben [12]. Für die Kopplung von Umweltsimulationskomponenten existieren etwa mit der *High Level Architecture* [9] oder dem *OGC Open Modelling Interface*<sup>9</sup> Ansätze zur Harmonisierung und Standardisierung der Schnittstellen. Für die Integration von Simulationskomponenten, speziell über Grenzen der Teildisziplinen hinweg, konnte sich jedoch bisher kein Standard etablieren. Die Entwicklungsaufwände für integrierte Umweltmodelle (z. B. zur gekoppelten Simulation von Szenarien zur sozio-ökonomischen Entwicklung und Effekte auf Atmosphäre, Ozean und Vegetation) die idealerweise auch Rückkopplungen berücksichtigen, bleiben daher sehr hoch und die heutigen integrierten Modelle sind eher monolithisch und damit Erweiterungen recht aufwendig [23].

Die dritte Interoperabilitätsebene zielt auf den einfachen Austausch von Analyse- und Simulationsalgorithmen zwischen Systemen bzw. Umweltforschern unterschiedlicher Disziplinen. Hier finden sich viele verschiedene Ansätze auf ganz unterschiedlichen Formalisierungs- bzw. Abstraktionsebenen, die sich zwischen teilweise generischen skriptbasierten Ansätzen [10] bis hin zu Formalisierungen einer Geoprozessierungsalgebra [6, 18] bewegen. Sie haben jedoch selten einen praxisrelevanten Reifegrad erreicht und finden kaum Anwendung in umweltwissenschaftlichen Projekten. Einzig Teilkonzepte, wie die Egenhofer-Operatoren [4] oder die von Tomlin entwickelte Map Algebra [22] haben sich etabliert und sind vielfach in heutigen Datenbanksystemen und GIS zu finden. Hier bleibt es Aufgabe der Forschung, zu prüfen, wie weit realisierbar die Vision von Beschreibungsmöglichkeiten für Algorithmen zur raum-zeitlichen Analyse und Modellierung ist, um zu ermöglichen, dass (a) diese Algorithmen im Sinne eines *moving code* problemlos zwischen verschiedenen Anwendungsumgebungen ausgetauscht werden können [13], (b) daraus komplexere Ana-

lysen leicht in ausführbare Workflows orchestriert werden können [6], (c) Wissenschaftlern zu ihren Analysefragestellungen passende Algorithmen und Operatoren vorgeschlagen werden [21, 27] und (d) durchgeführte Analyseschritte transparent und für andere Forscher nachvollziehbar gemacht werden können. Alternativ dazu stehen pragmatische Ansätze zur Schaffung von Datenmediatoren, die unter Nutzung gemeinsamer Ontologien und von webbasierten Diensteschnittstellen eine Integration von unterschiedlichen Simulationsmodellen zwar im Wesentlichen auf den Datenaustausch beschränken, jedoch eine effizient nutzbare, und mit geringerem Aufwand realisierbare Lösung skizzieren [1].

## Herausforderungen und Fazit

Mit dem Ziel des *reproducible research* sind immer häufiger und in vielen Disziplinen wissenschaftliche Publikation von Forschungsergebnissen an die Publikation der zugehörigen Forschungsdaten und/oder der entwickelten Algorithmen und Software gekoppelt. Für die in den Umweltwissenschaften meist als Open Source erfolgenden Softwarepublikationen haben sich verschiedene Plattformen (SourceForge, GitHub etc.) etabliert. Auch für die Datenpublikationen gibt es etablierte Repositorien (z. B. PANGAEA), in denen Datensätze durch einen Digital Object Identifier (DOI) eindeutig referenzierbar sowie durch standardisierte Metadatenangaben und pragmatische Vorgaben zu Datenstrukturen recherchierbar gemacht werden. Es fehlt derzeit jedoch an einheitlichen Vorstellungen bzw. Absprachen zu Aspekten wie Qualitätssicherung, Versionierungen, Lizenzierung oder langzeitlicher Archivierung von Forschungsergebnissen der Umweltwissenschaften als Daten und Software.

Es finden sich generelle Anforderungen von Fördermittelgebern zum Forschungsdatenmanagement und verschiedene auch fachspezifische Empfehlungspapiere und Ansätze [3, 14], die hier jedoch jeweils nur Teilaspekte adressieren. Hier sind weitere gemeinsame Arbeiten von Fachgesellschaften in den Umweltwissenschaften, Forschungseinrichtungen und Fördermittelgebern notwendig. Der Rat für Informationsinfrastrukturen (RfII<sup>10</sup>) skizziert mit der Nationalen Forschungsdateninfrastruktur (NFDI) einen Rahmen für diese Bemühungen [16].

<sup>9</sup> <http://www.openmi.org/>.

<sup>10</sup> <http://www.rfii.de>.

Bereits in 2008 wurde die Schwerpunktinitiative *Digitale Information*<sup>11</sup> der Allianz der deutschen Wissenschaftsorganisationen zur Koordination und Bündelung von Aktivitäten zur Digitalisierung in der Wissenschaft begonnen. Diese Initiative hat ebenfalls eine Reihe grundlegender Empfehlungen und Leitlinien, beispielsweise zum Forschungsdatenmanagement und dem Umgang mit Forschungssoftware publiziert, die wissenschaftlichen Fachgesellschaften und Einrichtungen als Referenz dienen können.

Eine der großen Herausforderungen liegt sicherlich in der Etablierung von allgemein akzeptierten Formen der Qualifizierung und Qualitätsbewertung für zu publizierende Forschungsdaten und Softwarewerkzeuge. Vorstellbar sind hier beispielsweise die Formulierung unterschiedlicher Qualitätslevel, Nutzung von Selbstverpflichtungsverfahren oder Evaluierungsverfahren (*peer reviews*). Weiterhin denkbar sind geeignete moderierte Verfahren zur kontinuierlichen Qualitätsverbesserung und Versionierung der Publikationen. Daran anknüpfen kann ein Monitoring der Nachnutzung von Daten und Werkzeugen, welches dann auch als Grundlage für Einschätzungen zur Relevanz der publizierten Daten und Software, bis hin zur Etablierung von entsprechenden Nutzungs- und Zitierungsindizes dienen kann.

Weitere große Herausforderungen, etwa die Sicherstellung der langfristigen Archivierung der als relevant erachteten Ergebnisse oder die Etablierung und Durchsetzung von Lizenz- und Urheberschutzrechten, erscheinen weniger fachspezifisch und sollten als generelle Querschnittsaufgabe für das künftige Forschungsdatenmanagement adressiert werden.

Um *e-science* in den Umweltwissenschaften vollumfänglich zu etablieren, sind noch einige Anstrengungen notwendig. Neben den in diesem Beitrag angeführten Forschungsaufgaben für die (Geo-)Informatik gehören dazu insbesondere auch die weitere Förderung und Etablierung einer offenen Daten- und Softwarekultur, die auf Kollaboration, Austausch und Transparenz unter Nutzung der entstehenden Plattformen und Systeme zielt.

Wesentliches Element hierfür ist sicherlich die Anpassung der Aus- und Weiterbildung von

Umweltwissenschaftlern. Schwerpunkte sollten dabei nicht allein der Kompetenzerwerb für das Forschungsdatenmanagement sowie zum fachkundigen Umgang mit den Analyse-, Modellierungs- und Kollaborationswerkzeugen sein, die sie sich bereits – allerdings nur vereinzelt – in derzeitigen Curricula und Promotionsprogrammen finden. Sondern es sollte auch eine Sensibilisierung für Qualitätsanforderungen und für rechtliche Aspekte bei der Arbeit mit und Publikation von Forschungsdaten erreicht werden, die heute nicht oder nur sehr rudimentär in Qualifizierungsprogrammen zu finden ist.

## Literatur

1. Belete GF, Voinov A, Morales J (2017) Designing the distributed model integration framework – DMIF. *Environ Modell Softw* 94:112–126, <https://doi.org/10.1016/j.envsoft.2017.04.003>
2. Bernard L, Brauner J, Mäs S, Wiemann S (2016) Geodateninfrastrukturen. In: Freuden W, Rummel R (eds) *Handbuch der Geodäsie: 5 Bde.* Springer, Berlin Heidelberg, S 1–32
3. Bertelmann R, Gebauer P, Hasler T, Kirchner I, Peters-Kottig W, Razum M, Recker A, Ulbricht D, van Gasselt S (2014) Einstieg ins Forschungsdatenmanagement in den Geowissenschaften. Helmholtz-Zentrum Potsdam – Deutsches GeoForschungszentrum GFZ, Potsdam, <https://doi.org/10.2312/lis.14.01>
4. Egenhofer M (1994) Spatial SQL: A query and presentation language. *IEEE Trans Knowl Data Eng* 6(1):86–95
5. Henzen D, Müller M, Jirka S, Senner I, Kaeseberg T, Zhang J, Bernard L, Krebs P (2016) A scientific data management infrastructure for environmental monitoring and modelling. Paper presented at the 8th International Congress on Environmental Modelling and Software, Toulouse, France
6. Hofer B, Mäs S, Brauner J, Bernard L (2017) Towards a knowledge base to support geoprocessing workflow development. *Intern J Geogr Inf Sci* 31(4):694–716, <https://doi.org/10.1080/13658816.2016.1227441>
7. Hu H, Lin T, Wang S, Rodriguez LF (2017) A cyberGIS approach to uncertainty and sensitivity analysis in biomass supply chain optimization. *Applied Energy* 203:26–40, <https://doi.org/10.1016/j.apenergy.2017.03.107>
8. Hussain M, Chen D, Cheng A, Wei H, Stanley D (2013) Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS J Photogr Remote Sens* 80:91–106, <https://doi.org/10.1016/j.isprsjprs.2013.03.006>
9. IEEE (2011) 1516 IEEE standard for modeling and simulation (M&S) high level architecture (HLA) framework and rules. Institute of Electrical and Electronics Engineers, New York
10. Knox S, Meier P, Yoon J, Harou JJ (2018) A python framework for multi-agent simulation of networked resource systems. *Environ Modell Softw* 103:16–28, <https://doi.org/10.1016/j.envsoft.2018.01.019>
11. Mäs S, Henzen C, Müller M, Bernard L (2014) GLUES GDI – Eine Geodateninfrastruktur für wissenschaftliche Umweltdaten. *gis.SCIENCE* 2014(4):129–137
12. Müller M (2015) Hierarchical profiling of geoprocessing services. *Comput Geosci* 82:68–77, <https://doi.org/10.1016/j.cageo.2015.05.017>
13. Müller M, Bernard L, Brauner J (2010) Moving code in spatial data infrastructures – web service based deployment of geoprocessing algorithms. *Trans GIS* 14(5):101–118, doi:10.1111/j.1467-9671.2010.01205.x
14. Nüst D, Konkol M, Pebesma E, Kray C, Schutzzeichel M, Przybytzin H, Lorenz J (2017) Opening the publication process with executable research compendia. *D-Lib Mag* 23(1/2), <https://doi.org/10.1045/january2017-nuest>
15. NSF (2007) Cyberinfrastructure vision for 21st century discovery [online]. US National Science Foundation Cyberinfrastructure Council, [https://www.nsf.gov/pubs/2007/nsf0728/nsf0728\\_1.pdf](https://www.nsf.gov/pubs/2007/nsf0728/nsf0728_1.pdf), last access: 29.9.2018
16. Rat für Informationsinfrastrukturen (2016) Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland. Rat für Informationsstrukturen, Göttingen
17. Rat für Informationsinfrastrukturen (2017) Entwicklung von Forschungsdateninfrastrukturen im internationalen Vergleich: Bericht und Anregungen Rat für Informationsinfrastrukturen. urn:nbn:de:101:1-201711084652

<sup>11</sup> <https://www.allianzinitiative.de/>.

18. Scheider S, Gräler B, Pebesma E, Stasch C (2016) Modeling spatiotemporal information generation. *Intern J Geogr Inf Sci* 30(10):1980–2008, <https://doi.org/10.1080/13658816.2016.1151520>
19. de Smith MG, Longley P (2015) Geospatial analysis. <http://www.spatialanalysisonline.com>, last access: 19.12.2018
20. Stasch C, Foerster T, Autermann C, Pebesma E (2012) Spatio-temporal aggregation of European air quality observations in the Sensor Web. *Comput Geosci* 47:111–118, <https://doi.org/10.1016/j.cageo.2011.11.008>
21. Stasch C, Scheider S, Pebesma E, Kuhn W (2014) Meaningful spatial prediction and aggregation. *Environ Modell Softw* 51:149–165, <https://doi.org/10.1016/j.envsoft.2013.09.006>
22. Tomlin D (1990) *Geographic information systems and cartographic modelling*. Prentice-Hall, Englewood Cliffs
23. Voinov A, Shugart HH (2013) “Integronsters”, integral and integrated modeling. *Environ Modell Softw* 39:149–158, <http://dx.doi.org/10.1016/j.envsoft.2012.05.014>
24. Wang S, Anselin L, Bhaduri B, Crosby C, Goodchild MF, Liu Y, Nyerges TL (2013) CyberGIS software: a synthetic review and integration roadmap. *Int J Geogr Inf Sci* 27(11):2122–2145, <https://doi.org/10.1080/13658816.2013.776049>
25. Wiemann S (2017) Formalization and web-based implementation of spatial data fusion. *Comp Geosci* 99:107–115, <https://doi.org/10.1016/j.cageo.2016.10.014>
26. Wiemann S, Bernard L (2016) Spatial data fusion in spatial data infrastructures using Linked Data. *Intern J Geogr Inf Sci* 30(4):613–636, [doi:10.1080/13658816.2015.1084420](https://doi.org/10.1080/13658816.2015.1084420)
27. Wiemann S, Karrasch P, Bernard L (2018) Ad-hoc combination and analysis of heterogeneous and distributed spatial data for environmental monitoring – design and prototype of a web-based solution. *Intern J Digit Earth* 11(1):79–94, <https://doi.org/10.1080/17538947.2017.1326984>
28. Wiemann SR, Karrasch P, Brauner J, Pech K, Bernard L (2012) Classification-driven air pollution mapping as for environment and health analysis. Paper presented at the iEMSs Sixth Biennial Meeting: International Congress on Environmental Modelling and Software (iEMSs 2012). International Environmental Modelling and Software Society
29. Wissenschaftsrat (2012) Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020, <http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf>, letzter Zugriff: 19.12.2018

# Hybride Edition und analoges Erbe

*Editionsphilologie und Alexander von Humboldt-Forschung  
in der digitalen Sattelzeit*

Tobias Kraft

Wer sich eine Weile mit ihm beschäftigt, bekommt den Eindruck: Für Alexander von Humboldt begeistert sich fast jeder. Das Interesse an dem preußischen Kosmopoliten und weltreisenden Natur- und Kulturforscher aus Berlin war schon zu Lebzeiten ein internationales Phänomen, Humboldt war als Wissenschaftler eine Berühmtheit des öffentlichen Lebens. Seit einigen Jahren erlebt das Interesse an Humboldt erneut eine Konjunktur und das nicht nur wegen der weltweit erfolgreichen Biographie von Andrea Wulf [10]. Es ist das intellektuelle Erbe von Humboldts Wissenschafts- und Weltverständnis, das in den Bann zieht. Zu den wohl berühmtesten Zeugnissen dieses Erbes gehören Humboldts Amerikanische Reisetagebücher. Die Staatsbibliothek zu Berlin (SBB) konnte 2013 im Auftrag der Stiftung Preußischer Kulturbesitz und mit Unterstützung eines Konsortiums aus öffentlichen und privaten Geldgebern die neun wertvollen Tagebuchbände für die öffentliche Hand erwerben. Das Bundesministerium für Bildung und Forschung (BMBF) hatte den mitfinanzierten Ankauf unter eine Bedingung gestellt: Die neun in Leder gebundenen Manuskriptkonvolute sollten zusammen mit dem umfangreichen handschriftlichen Nachlass, der bereits in der SBB aufbewahrt wurde, vollständig digitalisiert und wissenschaftlich untersucht werden. Berlin übernahm die Restaurierung und digitale Erschließung über den Handschriftenkatalog *Kalliope* und die *Digitalisierten Sammlungen* der SBB, an der Universität Potsdam startete 2014 ein großes Forschungsprojekt zur „Genealogie, Chronologie und Epistemologie“ der Tagebücher, das 2017 seinen Abschluss fand.

## Humboldts Handschrift wird digital

Die digitale Erschließung des kulturellen Artefakts „Humboldts Amerikanisches Reisetagebuch“ ist damit fertiggestellt. Die Nachlassdigitalisate stehen der Welt unter einer offenen Lizenz (CC BY-NC-SA 3.0) zur Verfügung, die Metadaten der Objekte sind verzeichnet und liegen in einer guten Erschließungstiefe vor. Doch das ist aus naheliegenden Gründen bloß der erste Schritt in der Generationenaufgabe, Humboldts handschriftlichen Nachlass digital nutzbar zu machen. Der zweite Schritt, die Erschließung der Handschrift selbst – der Essays, Reflexionen und Reiseschilderungen, der Notizen, Tabellen und Messprotokolle, schließlich der Skizzen und Zeichnungen – hat gerade erst begonnen. 2015 hat unter dieser Vorgabe das Editions- und Forschungsvorhaben „Alexander von Humboldt auf Reisen – Wissenschaft aus der Bewegung“ (AvH-R) an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) seine Arbeit aufgenommen. Seitdem entsteht in den Büros und der Forschungsbibliothek am Gendarmenmarkt die auf 18 Jahre projektierte hybride *edition humboldt*, die den handschriftlichen Gelehrtennachlass unter dem Gesichtspunkt seiner großen hemisphärischen Forschungsreisen veröffentlichten wird. Hybrid meint: Es entstehen sowohl eine gedruckte Fassung (*edition humboldt print*) in insgesamt 16 Bänden als auch eine Online-Edition

<https://doi.org/10.1007/s00287-018-01130-5>  
© Springer-Verlag Berlin Heidelberg 2018

Tobias Kraft  
Berlin-Brandenburgische Akademie der Wissenschaften,  
Jägerstraße 22/23, 10117 Berlin  
E-Mail: kraft@bbaw.de

(*edition humboldt digital*) der transkribierten und wissenschaftlich kommentierten Manuskriptbestände. Neben dem Konvolut aus Tagebüchern der Amerikanischen (1799–1804) und der Russisch-Sibirischen Reise (1829) werden – zum größten Teil erstmals – Nachlassbestände zu verschiedenen Themenkreisen erschlossen. Dazu zählen Arbeitsdokumente, Fahnenkorrekturen, Manuskriptbögen und Briefkorrespondenzen zu verschiedenen Erkenntnisinteressen aus dem weiten Horizont Humboldt'scher Wissenschaft: Pflanzengeographie und Biowissenschaften, Forschungsreisen und wissenschaftliche Sammlungen, Geographie und Geologie, Astronomie und Kartographie, sowie Anthropologie und Kulturgeschichte.

## **Digitale Modellierung und Erfassung**

Das Ziel, diese Erschließung in das Ergebnis einer Hybrid-Edition zu gießen, könnte im Angesicht gegenwärtiger Entwicklungen wie das Signum einer bald vorübergehenden Zeit erscheinen. Braucht man wirklich noch Bücher für die Edition solcher Texte? Ist die Entscheidung, hybrid zu publizieren, nicht bloß der Tatsache geschuldet, dass die Leserschaft aus wissenschaftlicher Community, interessierten Laien und den Humboldt-Fans des Bildungsbürgertums noch nicht ausreichend an digitale Lektüren gewöhnt ist? Wohl eher nicht: Zahlreiche Studien deuten darauf hin, dass das gedruckte Buch weiterhin die Bildschirmlektüre schlägt, wenn es um das konzentrierte Lesen geht, sowohl in Geschwindigkeit als auch in der Qualität des Lektüreprozesses [5]. Geisteswissenschaftliche Grundlagenforschung wie jene der BBAW steht zugleich in einer Tradition akademischer Praktiken, die auf Haltbarkeit gestellt ist: Akademienvorhaben sind Projekte mit ein, zwei oder drei Jahrzehnten Laufzeit, die Werke erstellen, die den Quellenstandard für Generationen liefern sollen. Weiterhin schlägt das Buch auf dieser Strecke den elektronischen Datensatz und seine Verarbeitungstechnologien um Längen, vor allem bei den Faktoren Haltbarkeit, Verfügbarkeit und öffentliche Infrastruktur. Wir wissen (ohne darüber weiter nachzudenken), dass Bücher unter angemessener Lagerung und konservatorischer Pflege Jahrhunderte halten und entsprechend lange benutzbar bleiben können. Es gibt komplexe Infrastrukturen und ein ausdifferenziertes System spezialisier-

ter Berufsstände, um diesen Erhaltungsstand sicherzustellen.

Wer allerdings die Entwicklungen der letzten zehn Jahre im Bereich von Digital Humanities und Open Science näher beobachtet, der weiß auch, dass eben diese Garantien der Halt- und Verfügbarkeit auch für digitale Forschungsdaten über einen Prozess zunehmender Standardisierung weltweit in immer höherem Maße zugesichert werden können, gerade durch den Ausbau entsprechender Infrastrukturen und den Aufbau darauf spezialisierter Berufsstände. Im Zuge dieser Entwicklung entstehen in den Geisteswissenschaften jene Kulturtechniken, die die Voraussetzungen dafür schaffen, das kulturelle Erbe und die Kulturerbeforschung zu digitalisieren. Doch diese Entwicklung ist ein langer Prozess und wer auf dem Weg dahin Entscheidungen für Einzelprojekte treffen muss, ist nicht selten mit unsicheren Prognosen konfrontiert. Die Laufzeit der Langzeitvorhaben an Akademien erweist sich hier als großer strategischer Vorteil. Nur Projekte mit einer solchen Dauer können sich erlauben, über den Tellerrand der unmittelbaren Aufgabe hinauszudenken und zukünftige Entwicklungen in den Blick zu nehmen. Anders formuliert: Ein Vorhaben, das 2017 an einer hybriden Edition historischer Handschriften arbeitet, muss offen bleiben für die Frage, welche Wünsche und Erwartungen man wohl in drei, fünf, zehn oder fünfzehn Jahren an eine solche Edition stellen wird.

Der Status quo ist dabei äußerst signifikant und deutet tatsächlich eine Zeitenwende an. Bislang, das erscheint offensichtlich, galt für Editionen dieser Art das Primat des Buchdrucks: Nur hier konnte aus Sicht der Philologien die Edition von Grundagentexten gelingen. Und natürlich werden auch weiterhin mit gutem Grund wissenschaftliche Editionen als Bücher produziert. Dennoch ist der Vorrang des Buches längst infrage gestellt. Noch schärfer: Die Beweislast scheint umgekehrt worden zu sein, wenn die DFG schreibt:

*Nicht für jedes Editionsprojekt wird eine digitale Veröffentlichungsform erwartet – in jedem Fall jedoch eine Sicherung und Bereitstellung der Textdaten in digitaler Form [...].*

*Für viele Editionsprojekte ist im Hinblick auf die Benutzbarkeit und die weiteren Forschungsarbeiten eine allein digitale Veröffentlichung ausreichend und bietet die größten Vorteile. [...]*

Die Veröffentlichung als Druckausgabe kann sowohl in Ergänzung zu einer digitalen Veröffentlichung geschehen (so genannte Hybridedition) als auch die alleinige Publikationsform sein. In beiden Fällen ist im Editions-konzept zu begründen, inwiefern die Printausgabe der Zielsetzung der Edition entspricht und welche Vorteile sie bietet [1].

Wer also heute Edition sagt und dabei im Sinne eines Hybridprodukts (auch) ein Buch meint, scheint auf eine Art Zwischenstufe aus analoger und digitaler Arbeits- und Produktionsmethode zu setzen. Tatsächlich aber ist die Arbeit an einem Vorhaben wie der *edition humboldt* von Beginn an digital. Die Modellierung, Erfassung und Repräsentation von Text sind Teile eines rein digitalen Arbeitsprozesses geworden, der auch einen konzeptionellen Wandel impliziert. Editionsrichtlinien orientieren sich heute in Umfang, Flexibilität und Formalisierung am maschinenlesbaren Rahmen eines Kodierungsschemas und nicht mehr allein an den Entscheidungen einer Reihenherausgeberin. Orts- und Personenregister sind keine Zeichenabfolge von Namen und Zahlen in einem Satzspiegel für den Buchdruck sondern werden zu normdateigestützten, relationalen Datenbanken. Bedeutungsvoller Text wird zu einem semantisch kodierten Textdatensatz.

Egal, in welche Form sich dieser komplexe Datensatz später gießt, ob als XML, HTML oder PDF-Vorlage für die Druckerei, er wird als Code digital produziert und vom digitalen Produktions- und Repräsentationsverfahren her konzipiert. Der informationelle Kern der Edition sind also die auf spezifische Weise kodierten XML-Textdaten, die den beiden Dekodierungsweisen einer Hybrid-Edition – dem PDF-Drucksatz und dem HTML-Code der Webseite – zugrunde liegen.

### Digitale und „analoge“ Daten

Das Beispiel der *edition humboldt digital* ([edition-humboldt.de](http://edition-humboldt.de)) zeigt, wie die verschiedenen Elemente dieses Textdatensatzes im Zusammenspiel mit X- und Datenbanktechnologien das zeitgemäße Modell einer kritischen digitalen Textedition (*digital scholarly edition*) ergeben können. Die digitale Strategie der Edition setzt sich zusammen aus zwei Komponenten: die möglichst standardkonforme und nachnutzungsorientierte Produktion

verlässlicher Daten auf der einen und die integrative Kuratierung vorhandener *analoger Daten* in die Architektur der digitalen Edition auf der anderen Seite.

Zunächst zur Produktion: Um heute auf verlässlicher Grundlage nachnutzbare Daten im Bereich digitaler Editionen zu erstellen, kann die Fachcommunity digitaler Editionsphilologen auf ein ganzes Set etablierter Standards zurückgreifen.

### TEI-XML-Schema

Das Korpus der *edition humboldt digital* wird auf Grundlage des TEI-XML P5-Standards kodiert. Das gilt nicht nur für das gesamte Markup der Texttranskription und -kommentierung, sondern ebenso für die Erstellung der Personen-, Orts- und Institutionenregister sowie für die Ereignisentitäten der Alexander-von-Humboldt-Chronologie [3]. Die Orientierung am XML-Kodierungsstandard der Text Encoding Initiative (TEI) hat sich in funktionalem Umfang und Verbreitung nicht nur seit einigen Jahren zunehmend gegenüber alternativen Kodierungssprachen oder -systematiken in der Anwendung durchgesetzt, sondern wurde 2013 von der DFG auch als *Must-have* zeitgenössischer digitaler Editionsprojekte festgeschrieben [2].

Allein eine Festlegung auf TEI-XML löst aber weder alle formalen Auszeichnungsprobleme des Editors noch gibt es auf jede funktionale Anforderung einer Edition eine eindeutige Antwort. Dies zeigt ein Blick auf die Richtlinien der TEI. Wer sich die vollständigen Richtlinien auf Englisch unter <http://tei-c.org> als PDF-Version lädt, bekommt ein Handbuch im Umfang von 1887 Seiten geliefert [9]. Da die philologisch präzise Auszeichnung von Textmerkmalen aber nicht funktionieren kann wie eine Reise nach Rom, steht jedes Projekt vor der Entscheidung, aus der Masse an verfügbaren Optionen ein spezifisches Tag Set auszuwählen, das den editorischen Standard des Einzelprojekts regelt. Auch hier zeichnen sich zunehmend Standardisierungsprozesse ab, an denen auch die *edition humboldt digital* partizipiert. So wurde das *edition-humboldt*-Tag-Set in enger Abstimmung mit dem Basisformat des Deutschen Textarchivs (DTABf) entwickelt, was selbst wiederum jüngst von der DFG als *Best-Practice*-Beispiel empfohlen wurde und damit als quasi-normierter TEI-XML-Standard zumindest im deutschspra-



## Georg Forster

1754–1794



Naturwissenschaftler, Schriftsteller und Weltreisender; ging 1766 mit seinem Vater nach London; 1772–1775 Teilnahme an der zweiten Weltumseglung Cooks; 1777 Veröffentlichung seiner Reisebeschreibung „A voyage round the world“; 1778–1784 Lehrer für Naturgeschichte am Collegium Carolinum in Kassel; 1784 Annahme eines Rufes nach Vilnius (Wilna) als Professor für Naturwissenschaften; 1785 Heirat mit Therese Heyne in Göttingen; 1790 Reise mit Humboldt durch das Rheinland, Brüssel, Den Haag, Amsterdam, London, Paris und anschließend Veröffentlichung der „Ansichten vom Niederrhein“; 1792–1793 führende Beteiligung an der Revolution in Mainz, 25. März 1795 Abreise nach Paris als Deputierter des Mainzer Konvents

Alternative Namen bzw. Schreibungen: Georg Johann Adam Forster

### Erwähnt in Briefen

Datum	Korrespondent	Ort
28.01.1791–20.02.1791	An Samuel Thomas von Soemmerring	Hamburg
06.02.1794	An Samuel Thomas von Soemmerring	Bayreuth
21.02.1801	An Karl Ludwig Willdenow	Havanna
13.01.1849	Von Carl Sigismund Kunth	Berlin

### Autor folgender Titel im Literaturregister

De plantis esulentis insularum oceanii australis commentatio botanica

### Erwähnungen in der Chronologie

24. September 1789  
21. März 1790  
Juni 1790

### Nachweise in gedruckten Editionen

#### Jugendbriefe

Humboldt, Alexander von (1973): *Die Jugendbriefe Alexander von Humboldts 1787–1799. Herausgegeben und erläutert von Ilse Jahn und Fritz G. Lunge.* Berlin: Akademie-Verlag (Beiträge zur Alexander-von-Humboldt-Forschung, 2).

Nr. 28, Nr. 10, Nr. 26, Nr. 30, Nr. 32, Nr. 33, Nr. 34, Nr. 38, Nr. 39, Nr. 43, Nr. 46, Nr. 48, Nr. 52, Nr. 54, Nr. 55, Nr. 56, Nr. 57, Nr. 58, Nr. 59, Nr. 61, Nr. 62, Nr. 65, Nr. 72, Nr. 84, Nr. 88, Nr. 92, Nr. 93, Nr. 96, Nr. 103, Nr. 164, Nr. 297, Nr. 350, Nr. 392, Nr. 411

#### Briefwechsel mit dem preußischen Kultusministerium

#### Briefe der Reise in die amerikanischen Tropen

#### Briefe Humboldt-USA

#### Humboldt-Bots-Reymond Briefwechsel

#### Tagebuch Reise mit van Geuns 1789

#### Humboldt-Spiker Briefwechsel

#### Humboldt-Costa Briefwechsel

#### Humboldt-Ritter Briefwechsel

#### Briefe zur Steingutfertigung in Rheinsberg 1792

### Links zu externen Websites

- Nachweise, edierter Briefe zu dieser Person in *correspondenz*
- Hidden Rooms
- Deutsches Textarchiv
- Allgemeine Deutsche Biographie (ADB)
- Neue Deutsche Biographie (NDB)
- Mitglieder der Königlich-Preussischen Akademie der Wissenschaften
- Wikisource
- Wikipedia

### Beziehungen zu Personen

- Vater: Johann Reinhold Forster
- Ehefrau: Therese Heyne

Abb. 2 Personenregister, Eintrag „Forster, Georg“ (<https://edition-humboldt.de/H0000177>, letzter Zugriff: 9.12.2018)

GeoNames. Das Orts- und Institutionenregister der *edition humboldt digital* umfasst derzeit rund 1.800 redaktionell überprüfte Orts- und über 100 Institutioneneinträge, denen GeoNames-IDs zugewiesen oder die als Ersteinträge in die GeoNames-Datenbank eingepflegt werden konnten. Das Beispiel der Berliner Sternwarte, die zudem zu Humboldts Lebzeiten (u. a. aufgrund seines Einflusses)

den Standort wechselte, zeigt den differenzierten Einsatz lokal spezifischer GeoNames-Normdateien besonders nachdrücklich. Die Einträge zur Sternwarte (Institutionen) und zu Berlin (Orte) machen darüber hinaus deutlich, wie die Verknüpfung aus Personen-, Orts- und Institutionenregister in der *edition humboldt digital* konzeptionell umgesetzt wird (Abb. 3 und 4).

edition humboldt digital    Reisetagebücher    Themen    Briefe    Chronologie    Register    🔍

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

## Berliner Sternwarte

Berlin

Die Berliner Sternwarte war eine astronomische Forschungseinrichtung, die in Verbindung mit der Kurfürstlich-Brandenburgischen Societät der Wissenschaften gegründet und unter der nachfolgenden Königlich-Preussischen Societät beziehungsweise der Königlich-Preussischen Akademie der Wissenschaften als Königliche Sternwarte zu Berlin betrieben wurde.

1711–1835  
Berlin, Neustadt, Letzte Straße No. 64



**Nachweise in gedruckten Editionen**

**Briefwechsel mit dem preussischen Kultusministerium**

Humboldt, Alexander von (1985): Vier Jahrzehnte Wissenschaftsförderung. Briefe an das preussische Kultusministerium 1818-1858. Herausgegeben von Kurt-R. Biermann. Berlin: Akademie Verlag (Beiträge zur Alexander-von-Humboldt-Forschung, 14).

Nr. 18

**Briefe der Reise in die amerikanischen Tropen**

**Humboldt-Bois-Reymond Briefwechsel**

**Humboldt-Encke Briefwechsel**

1855–1913  
Berlin, Lindenstraße, später Enckeplatz 5A



**Nachweise in gedruckten Editionen**

**Briefwechsel mit dem preussischen Kultusministerium**

Humboldt, Alexander von (1985): Vier Jahrzehnte Wissenschaftsförderung. Briefe an das preussische Kultusministerium 1818-1858. Herausgegeben von Kurt-R. Biermann. Berlin: Akademie Verlag (Beiträge zur Alexander-von-Humboldt-Forschung, 14).

Nr. 24, Nr. 94

**Briefe Humboldt-USA**

**Humboldt-Encke Briefwechsel**

**Erwähnungen im Personenregister**

- Bernoulli, Johann
- Bode, Johann Elert
- Encke, Johann Franz
- Galle, Johann Gottfried
- Schivel, Karl Friedrich

**Abb. 3** Einrichtungsregister, Eintrag „Berliner Sternwarte“ (<https://edition-humboldt.de/H0000531>, letzter Zugriff: 9.12.2018)

## Brückentechnologien in der digitalen Sattelzeit

Wie aber gelingt es dem editionsphilologisch geschulten Forschenden – etwa einer Germanistin, einem Romanisten oder einer Wissenschaftshistorikerin – diese (auch) formal nicht trivialen Textdaten zu produzieren?

Die Editoren und Projektmitarbeiter der *edition humboldt digital* arbeiten zu diesem Zweck mit der digitalen Arbeitsumgebung *ediarum*. Dabei handelt es sich um eine von der TELOTA-Initiative der BBAW entwickelte Datenarchitektur- und Softwarelösung, die es auch dem in Markup ungeübten Wissenschaftler erlaubt, Transkriptionen von Manuskripten in

## Berlin

Schreibort		
Datum	Korrespondent	Ort
17.07.1827	An Carl Sigismund Kunth	Berlin
nach 01.01.1828– vor 31.12.1828	An Franz Julius Ferdinand Meyen	Berlin
01.09.1828	An Carl Sigismund Kunth	Berlin
05.05.1829	An Karl Christian Gmelin	Berlin
11.05.1829	An Franz Julius Ferdinand Meyen	Berlin
01.08.1829– 22.05.1850	An Marie-Josèphe Kunth	Berlin
nach 01.01.1830– vor 31.05.1830	An Franz Julius Ferdinand Meyen	Berlin
nach 01.01.1830– vor 31.05.1830	An Franz Julius Ferdinand Meyen	Berlin
nach 01.01.1830– vor 31.05.1830	An Franz Julius Ferdinand Meyen	Berlin
nach 01.01.1830– vor 27.09.1830	An Franz Julius Ferdinand Meyen	Berlin
14.05.1830	An Franz Julius Ferdinand Meyen	Berlin
nach 01.10.1830– vor 22.05.1850	An Carl Sigismund Kunth	Berlin
nach 23.04.1832– vor 06.05.1859	An Franz Julius Ferdinand Meyen	Berlin
nach 16.05.1832– vor 31.12.1852	An Carl Sigismund Kunth	Berlin
26.10.1832	An Carl Sigismund Kunth	Berlin
30.10.1832	An Carl Sigismund Kunth	Berlin
nach 01.01.1833– vor 31.12.1833	An Franz Julius Ferdinand Meyen	Berlin
26.05.1833	An Carl Sigismund Kunth	Berlin
nach 01.07.1833– vor 02.08.1833	An Franz Julius Ferdinand Meyen	Berlin
08.07.1833	An Franz Julius Ferdinand Meyen	Berlin
08.07.1833	An Carl Sigismund Kunth	Berlin
09.09.1799	An Franz Julius Ferdinand Meyen	Berlin



Einrichtungen an diesem Ort

- Ägyptisches Museum Berlin
- Allgemeine Kriegsschule Berlin
- Altes Museum (Berlin)
- Berliner Gewerbeschule
- Berliner Sternwarte
- Botanischer Garten (Berlin)
- Gewerbeinstitut Berlin
- Kartographisches Institut
- Königliche Bibliothek (Berlin)
- Mineralogisches Museum der Berliner Universität
- Neues Museum (Berlin)
- Preussische Akademie der Künste
- Preussische Akademie der Wissenschaften
- Preussisches Finanzministerium
- Preussisches Kultusministerium
- Seehandlungsgesellschaft
- Tierarzneischule (Berlin)
- Universität zu Berlin (Alma Mater Berolinensis)
- Zoologischer Garten (Berlin)
- Zoologisches Museum (Berlin)

Abb. 4 Ortsregister, Eintrag „Berlin“ (Ausschnitt; <https://edition-humboldt.de/H0005221>, letzter Zugriff: 9.12.2018)

TEI-konformem XML in einer benutzerfreundlichen Oberfläche zu bearbeiten, mit einem Text- und Sachapparat zu versehen und sich online als HTML-Output anzeigen zu lassen. Die in ediarum zu diesem Zweck eingesetzten X-Technologien und Softwarekomponenten sind u. a.: oXygen XML Author, eXistdb, XQuery und XSLT.

Die Textarbeit in ediarum ist komfortabel insofern sie dem Bearbeiter erspart, „hart“ zu kodieren, aber sie verlangt dem Word-geschulten Auge eine folgenreiche Adaption ab. Im Unterschied zur Suggestion des Word-Prozessors, das Druckerpapier buchstäblich bei der Verfertigung des Textes zu beschreiben, setzt ediarum auf ein „what you see is

what you mean“-Interface. Dieser semantische (und nicht typografische) Ansatz ist eine für Textwissenschaftler zugleich anregende, aber auch irritierende Annäherung an den kodierten „Rohtext“, da sie eine Umstellung des analytischen Sehens, also auch des analytischen Lesens verlangt. Gezeigt wird in der „author-Ansicht“ des über ediarum angepassten oXygen XML-Editors eben nicht das, was die Webseite oder der Drucksatz später ausgibt, sondern das, was die Kodierung an Bedeutung im transkribierten Text selbst an zusätzlicher Information anlegt (Abb. 5). Text der Transkription und Metatext der Kodierung werden visuell ineinander geblendet, bleiben dabei strukturell voneinander getrennt und



Abb. 5 Bildschirmausschnitt des oXygen-Editors in der digitalen Arbeitsumgebung ediarum

können so separat bearbeitet werden. Schon in dieser *internen* Ansicht entsteht im Laufe der Arbeit eine Art zweiter Lektürekompentz, die sich in der Folge auch durchaus auf die Fähigkeit des ediarum-trainierten Bearbeiters überträgt, XML direkt zu lesen und zu verstehen.

Medientypologisch könnte man also davon sprechen, dass eine Brückentechnologie des digitalen (ediarum) eine andere Brückentechnologie des typographischen Zeitalters (Word) ersetzt. Für den Fall der Editionsphilologien markiert dieser Stabwechsel von einer zu einer anderen (digitalen) Technologie nicht weniger als den Scheitelpunkt jener Zeit, die wir in Anlehnung an Reinhart Kosellecks berühmtes Postulat auch als „digitale Sattelzeit“ bezeichnen können. Diese Sattelzeit markiert einen Übergang zwischen einer mehrheitlich analog zu einer mehrheitlich digital organisierten Gesellschafts- und Kulturordnung, die mit der Einführung der Personal Computer an den Arbeitsplätzen und in den Eigenheimen der westlichen Welt Mitte der achtziger Jahre ihren Anfang nahm und seitdem in unterschiedlichen Phasen alle Bereiche des gesellschaftlichen, politischen, wirtschaftlichen und kulturellen Lebens zu einem fundamentalen Wandel geführt hat.

Brückentechnologien sind der technologisch-mediale Ausdruck dieses Wandels. Die quasi-haptische Word-Simulation einer bereits infolge des Schreibens bedruckten Papierseite gab dem Editionsphilologen in der Vergangenheit zumindest das Gefühl, stets an einem zukünftigen Buch zu arbeiten, das vermeintlich bereits bei der Arbeit auf dem Bildschirm als Druckseitenfolge fertig gestellt wurde. Die typografischen, satztechnischen, format- und rezeptionsspezifischen Begrenzungen des Mediums Buch hingegen spielten dabei nur selten eine Rolle. Entweder sie wurden durch das editionsphilologische Projekt bis an die Grenzen des im Buchdruck möglichen getrieben oder es herrschte das Diktum einer alternativlosen Welt, in der das Buch nun einmal das einzige Medium für die philologische Textkonstitution und Sachkommentierung eines anderen Textes war. Die Vorstellung aber, dass ein Buch notwendigerweise das richtige oder gleich das beste Medium zur analytischen Aufbereitung eines anderen Buches (oder zumindest Textes) sei, ist in hohem Maße zeitgebunden. Sie gehört zur mentalitätsgeschichtlichen Spätphase einer Zeit, die in der Kulturtradition des Buchdrucks kaum in der Lage war, über die Handlungsoptionen einer Textproduktion als Buch hinauszudenken. Pa-

trick Sahle hat hierfür in seiner dreibändigen Studie zu *Digitalen Editionsformen* den Begriff des „typographischen Erbes“ geprägt [9]. Eine Online-Edition zu entwickeln heißt also nicht weniger als sich dieses analogen Erbes bewusst zu sein, um sich von seinen nur vermeintlichen Zwängen an entscheidenden Punkten lösen zu können.

Als verantwortlicher Herausgeber und Bearbeiter einer Edition kann man die sich daran anschließenden Entwicklungsfragen vom Ziel her angehen: Was will ich mit der Edition erreichen? Für wen mache ich sie? Man kann die Frage aber auch aus der Perspektive der Datenmodellierung und der Nachnutzung formulieren: Was muss ich tun, damit ich den Schritt der Konvertierung eines analogen Erbes in digitale Daten als Mehrwert für die Forschung nutzbar machen kann? Wie spreche ich die potenziellen Leser auch als Nutzer von Textdaten an, ohne vorher zu wissen oder wissen zu müssen, was mit diesen Daten geschieht? Wie also entwickle ich eine Edition, die nicht schon vorher einschränkt, an wen sie sich mit welchem Zweck und Ziel richtet?

Die Antwort auf diese Frage dürfte Informatiker nicht überraschen. Sie lautet in Kürze: offene Standards, gute Dokumentation, einfach zu nutzende Schnittstellen und möglichst offene Lizenzen.

### TEI-XML, Lizenzierung und OAI-PMH

Die *edition humboldt digital* bildet da keine Ausnahme. Natürlich gibt es auch hier einen konkreten Erwartungshorizont, den Editoren und die Alexander-von-Humboldt-Forschung im engeren Sinne teilen, wenn sie sich über den Anspruch an Qualität, Umfang und Funktionalität der Edition austauschen. Wer aber heute eine digitale (oder hybride) Edition erstellt, muss über diesen Erwartungshorizont hinausdenken und ein klares Bewusstsein dafür entwickeln, an der Produktion eines Datensatzes zu arbeiten, der potenziell weit über die Lektürebedürfnisse eines literatur-, kultur- oder wissenschaftshistorischen Publikums hinausgeht.

Daher werden die Texte und Datensätze der Edition über eine Schnittstelle [7] als TEI-XML bereitgestellt und können von dort „geharvestet“ werden. Projektinterne IDs für Personen, Orte und Institutionen werden durch Norm-IDs (GND, GeoNames etc.) ersetzt. Falls kein Eintrag in einer Normdatei existiert, wird die Entität über eine offizielle Projekt-URI referenziert. Zusammen mit der freien Lizenzierung der TEI-XML-Daten unter der

Creative-Commons-Lizenz CC BY-SA 4.0 wird so eine umfassende Nachnutzung der Ressourcen in anderen Kontexten und Disziplinen ermöglicht.

Um die Vernetzung und die Auffindbarkeit der Texte der digitalen Edition zu erhöhen, werden alle Metadaten auch über das „Protocoll for Metadata Harvesting“ der Open Archive Initiative (OAI-PMH) bereitgestellt. Dadurch ist der gesamte Datenbestand der *edition humboldt digital* nach einmaliger Anmeldung bei wissenschaftlichen Suchmaschinen, etwa der Bielefeld Academic Search Engine (BASE), nachgewiesen und auffindbar.

Für die AvH-Forschung bietet die nativ digital erstellte Edition ein sehr kompakt kuratiertes Netzwerk an relevanten Forschungsdaten und stellt Humboldt-Recherchen weltweit auf eine neue Grundlage. Zum anderen aber – und hier entstehen die eigentlichen Mehrwerte – wird die Humboldt-Forschung der (nahen) Zukunft auf immer mehr digital erschlossene und vergleichbar hergestellte Quellen und Textdatensätze zurückgreifen können. Die intelligente Verbindung dieser Text- und Datenkorpora schließlich erlaubt der Humboldt-Forschung neue Zugänge zu ihrem Material. Zwei Beispiele aus der aktuellen Version der *edition humboldt digital* sollen diesen Ansatz verdeutlichen:

### Die Alexander von Humboldt-Chronologie

Die Alexander-von-Humboldt-Chronologie wurde erstmals 1968 als *Alexander von Humboldt. Chronologische Übersicht über wichtige Daten seines Lebens* im Berliner Akademie-Verlag veröffentlicht. 1983 kam es zu einer zweiten, vermehrten und berichtigten Auflage, die 2001 als Online-HTML-Ressource erneuert und bis 2015 vom Leiter der Alexander-von-Humboldt-Forschungsstelle Ingo Schwarz (BBAW) betreut und erweitert wurde. Für die Integration dieser wichtigen und seit rund 50 Jahren analog und digital kuratierten Daten in die *edition humboldt digital* wurden die chronologischen Einträge zu Humboldts Leben als Einzelereignisse in über 1400 TEI-XML-Dokumente konvertiert. Seitdem arbeitet das Projektteam an der sukzessiven Indexierung der Chronologie in den Datenbestand der Personen-, Orts- und Institutionenregister. Wie bei der Erstellung der Registerdaten aus gedruckten Registertexten jahrzehntelanger Forschungstätigkeit ist auch die Konvertierung der Chronologie ein Beispiel für eine zunehmend digi-

edition humboldt digital    Reisetagebücher    Themen    Briefe    Chronologie    Register    ?    🔍

## Alexander von Humboldt-Chronologie – 1805

◉ Weitere Informationen zu dieser Chronologie

27 Treffer

Treffer im Deutschen Textarchiv

1805: Antrittsrede gehalten bei seiner Einführung in die Königlich-Preussische Akademie der Wissenschaften von Alexander von Humboldt.  
Humboldt, Alexander von; Antrittsrede gehalten bei seiner Einführung in die Königlich-Preussische Akademie der Wissenschaften. [Berlin], [1805].

Treffer im Deutschen Textarchiv

1805: Versuche über den Zitterrochen  
Humboldt, Alexander von; Versuche über den Zitterrochen. In: Neues allgemeines Journal der Chemie, Bd. 6, H. 2 (1805), S. 166-172.

07. Januar 1805  
H. liest im „Institut“ „Discours préliminaire: Essai sur la géographie des plantes“.

21. Januar 1805  
H. und Gay-Lussac lesen im „Institut“ über „Expériences sur les moyens eudiométriques, et sur la proportion des principes constituans de l'atmosphère.“

Alexander von Humboldt an Karl Ludwig Willdenow. [Paris], 1. Februar 1805  
So gross auch der Wirrwarr meiner eingenen Geschäfte ist, so werde ich doch Zeit finden, Deine Aufträge zu besorgen. Zu Dupetit-Thouard, der ein gar hölzerner Mensch ist, gehe ich heute selbst. Es ist unendlich schade, dass Dein guter Genius Dich nicht dieses Jahr statt nach Triest nach Pa [...]

Treffer auf correspSearch

Alexander von Humboldt an König Friedrich Wilhelm III. von Preußen. Paris, 10. February 1805  
Schwarz, Ingo (2015): „etwas hervorzubringen, was meines Königs und meines Vaterlandes werth sein kann“ – Briefe von Alexander von Humboldt an Friedrich Wilhelm III., 1805 In: HIN - Humboldt im Netz. Internationale Zeitschrift für Humboldt-Studien (Rotterdam - Berlin) XVI, 31, S. 3-10. Online verfügbar unter: <http://bib.dnlg.org/10.18443/218>

11. Februar 1805  
H. liest im „Institut“ über „deux nouveaux genres de poissons de l'ordre des Apodes“ und über eine „espèce de silure, vomie quelquefois par les volcans dans les environs de Quito“.

19. Februar 1805  
H. wird ordentliches Mitglied der Akademie der Wissenschaften zu Berlin.

Suche

Jahr	1766	1767	1769	1775
	1777	1778	1779	1781
	1783	1784	1786	1787
	1788	1789	1790	1791
	1792	1793	1794	1795
	1796	1797	1798	1799
	1800	1801	1802	1803
	1804	1805	1806	1807
	1808	1809	1810	1811
	1812	1813	1814	1815
	1816	1817	1818	1819
	1820	1821	1822	1823
	1824	1825	1826	1827
	1828	1829	1830	1831
	1832	1833	1834	1835
	1836	1837	1838	1839
	1840	1841	1842	1843
	1844	1845	1846	1847
	1848	1849	1850	1851
	1852	1853	1854	1855
	1856	1857	1858	1859

Suchoptionen

- Briefe aus dieser Edition
- Schriften von A. v. Humboldt im Deutschen Textarchiv

Anzeigooptionen

Hinweis: Sowohl correspSearch als auch die Digitalisate-Bibliographie stellen nur Metadaten bereit. Diese Anzeigooptionen können daher nur verwendet werden, wenn oben kein Suchbegriff angegeben wurde.

- Briefe von und an Humboldt aus correspSearch
- Schriften von A. v. Humboldt in der Digitalisate Bibliographie (nur ohne Suchbegriff)

Absenden

Die Erstellung der Datenbestände der Chronologie ist ein fortlaufender Prozess, Umfang und Genauigkeit der Daten wachsen mit dem Vorschreiten des Vorhabens. Hinweise, Ergänzungen, Berichtigungen werden dankbar entgegengenommen. Bitte schreiben Sie an [damm-dvnr@bibw.de](mailto:damm-dvnr@bibw.de)

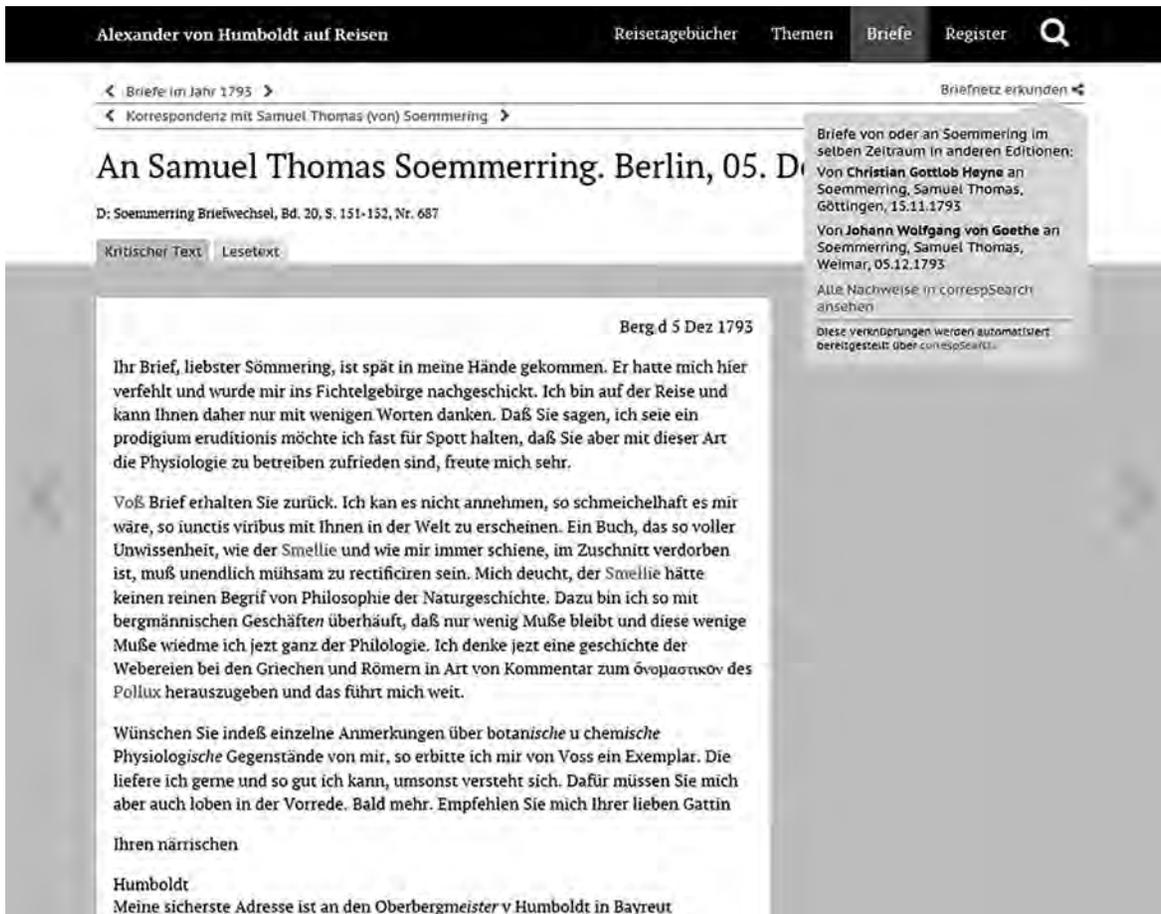
**Abb. 6 Anzeige der Chronologie im Jahr 1805 mit chronologisch sortierten Quellen aus der Edition, der correspSearch-Datenbank und den Volltexten des Deutschen Textarchivs**

tale Kuratierung vormals nur im Druck vorliegender Forschungsdaten.

### correspSearch und DTA

Der Volltext der Chronologie ist durch die XML-Konvertierung und Integration in das Framework der Edition erstmals vollständig durchsuchbar. Die

Anzeige der Chronologie kann darüber hinaus optional erweitert werden um verschiedene Text- und Datenressourcen, die aus dem Forschungsdatensatz der Chronologie eine vernetzte Architektur sehr spezifischer Informationen für die Alexander-von-Humboldt-Forschung machen, wie sie bisher in dieser Genauigkeit nicht möglich war.



**Abb. 7** *correspSearch*-Option „Briefnetz erkunden“ am Beispiel des Soemmerring-Briefwechsels, <https://edition-humboldt.de/H0002655>, letzter Zugriff: 9.12.2018

Das Prinzip dahinter ist sehr einfach: Eine Chronologie sind Ereignisse organisiert an der Kette ihrer zeitlichen Abfolge. Das erlaubt es, zeitlich bestimmbare Dokumente wie Briefe oder publizierte Texte in die Ansicht der Chronologie zu integrieren (Abb. 6). Dies ist derzeit möglich für

- edierte Dokumente der *edition humboldt digital* (derzeit über 400 Briefe und Dokumente),
- Humboldt-Korrespondenzen aus der Briefdatenbank *correspSearch* (derzeit rund 4000 verzeichnete Briefe aus 43 Publikationen) sowie
- Volltexte aus dem Korpus von Humboldts unselbständig erschienenen Schriften aus dem Angebot des Deutschen Textarchivs (derzeit 166 Einzeltexte).

### Correspondence Metadata Interchange Format (CMIF)

Der Webservice *correspSearch* ist, wie bereits am Beispiel der Chronologie gesehen, eng mit der *edition humboldt digital* verknüpft. Grundlage für diese Datenverknüpfung ist das gemeinsam genutzte TEI-XML-Datenformat für Briefmetadaten CMIF. Mit der Option „Briefnetz erkunden“ sind die kontextrelevanten Metadaten der Briefdatenbank *correspSearch* so direkt in die Einzelansicht der Briefe eingebunden (Abb. 7).

Das Beispiel zeigt einen Brief von Humboldt an Samuel Thomas Soemmerring vom 05. Dezember 1793. Über die Schaltfläche „Briefnetz erkunden“ werden dynamisch auf Grundlage der *correspSearch*-Datenbanken „Umkreis-Korrespondenzen“ angezeigt: in diesem Fall Briefe von Heyne und Goethe an Soemmerring, geschrieben im selben Zeit-

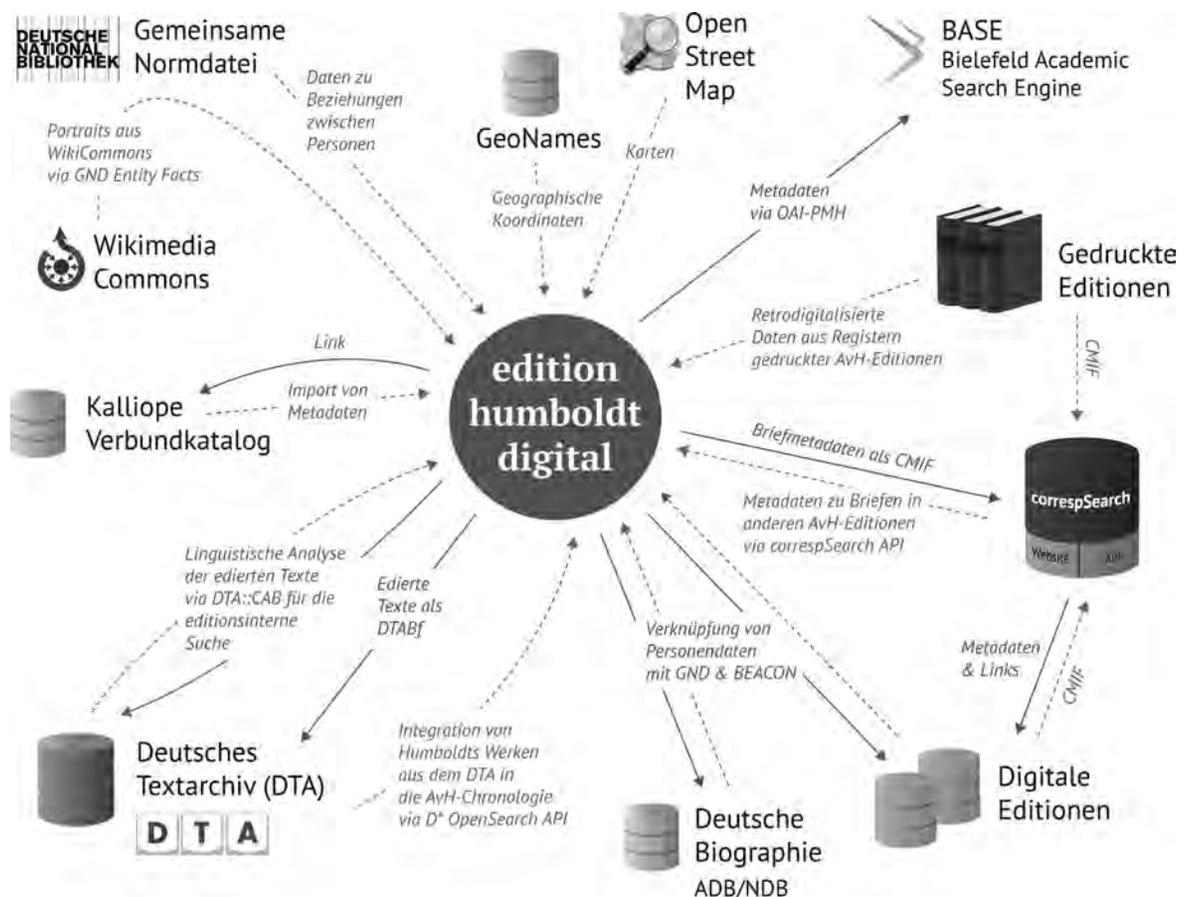


Abb. 8 Schematische Darstellung der vernetzten Datenarchitektur der edition humboldt digital

raum wie Humboldts Brief an Soemmerring. Der Alexander-von-Humboldt-Forschung werden auf diese Weise wertvolle Zusatzquellen erschlossen, die Humboldts komplexes Korrespondentennetzwerk erstmals in dieser Form mit Korrespondenzen seiner Zeitgenossen vernetzen und für übergreifende Forschungsfragen bearbeitbar machen.

### Heute und in Zukunft: Vernetzte Daten und Webservices

Die vorgestellten Beispiele und Nutzungsszenarien zeigen: Für die *edition humboldt digital* wurden an mehreren Stellen Daten aus Vorgänger- oder Drittprojekten neu aufbereitet oder nachgenutzt bzw. externe Webservices in die eigene Datenarchitektur integriert. Der funktionale Umfang der Edition wurde damit erheblich erweitert und erlaubt der historischen Forschung eine bisher nicht erreichte Verknüpfung spezifischer Forschungsdaten und -kontexte (Abb. 8).

Die *edition humboldt digital* steht mit dem hier formulierten Anspruch an Datenmodellierung und Datennachnutzung zweifelsohne noch am Anfang. Es gehört zu den unbestreitbaren Vorteilen des deutschen Akademienprogramms, dass für die weitere Entwicklung der Edition und damit für einen Beitrag zur Entwicklung digitaler Editionen allgemein noch rund fünfzehn Jahre Projektlaufzeit zur Verfügung stehen. Dadurch kommen Entwicklungshorizonte in den Blick, die in einer üblichen Projektfinanzierung von wenigen Jahren Umfang nie über einen Beta-Status hinauskommen würden.

Hierzu gehört ein Nachdenken über die Grenzen der verwendeten Technologien und Datenspezifika. TEI-XML etwa ist als Datenformat keineswegs über jeden Zweifel erhaben und kommt aufgrund seiner strengen hierarchischen Struktur an Grenzen der Vereinbarkeit des für Editoren wichtigen Verhältnisses von Befundtreue, textkritischer Genauigkeit und funktionaler Darstellbarkeit. Dieses Verhältnis

aber streng auszulegen, ist selbst Zeugnis eben jenes typographischen Erbes, das digitale Editionen nicht bedienen sollten, wenn sie zu ihrem eigenen medialen Ausdruck finden wollen. Der romanistische Philologe Christof Schöch hat jüngst einen kompakten Vergleich verschiedener Textformate vorgelegt und diese unter den Aspekten proprietär/offen, wenig/stark strukturiert, darstellungs- oder strukturorientiertes Encoding analysiert. Schöch's Studie zeigt die Überlegenheit von TEI-XML für eine literaturwissenschaftlich adäquate digitale Repräsentation literarischer Texte:

*TEI repräsentiert explizit die Zeichensequenz, die Textstruktur, lokale editionswissenschaftliche Annotationen und dokumentbezogene Metadaten. Es ist langfristig nutzbar, da es auf dem weit verbreiteten Standard XML beruht und eine große, institutionalisierte Gemeinschaft von Nutzerinnen und Nutzern hinter ihm steht [8].*

Wer das funktionale Wachstum einer digitalen Edition plant, sollte auch über die Skalierbarkeit dieses Wachstums nachdenken. Wie viel kann und darf in Zukunft dazu kommen? Ressourcenentwickler wissen, dass sich dieser Prozess nur vernünftig steuern lässt, wenn man ihn nicht additiv, sondern inkrementell denkt. Jede bereits in die Edition integrierte Komponente verlangt eine zukünftige Betreuung, Wartung und Kuratierung. Dabei handelt es sich keineswegs nur um eine organisatorische Frage: Wenn es nicht gelingt, digitale Editionen als dauerhafte Wissensressourcen, also als wissenschaftlich referenzierbare Primärquellen, zu etablieren, weil die Stabilität ihres Angebots nicht sichergestellt werden kann oder in Zuständen fortgesetzter Vorläufigkeit verharret, dann ist das Ziel einer damit verbundenen digitalen (Editions)-Philologie reichlich verfehlt. Strategisch lässt sich die hier aufgeworfene Problematik nur beantworten, indem man sorgfältig abwägt, welche technischen Komponenten die Entwicklung digitaler Editionen nicht nur qualitativ verbessern, sondern auch strukturell absichern. Je standardisierter die Strukturelemente einer vernetzten Edition sind, desto eher arbeitet man mit verlässlichen Komponenten. Die GND und Georeferenzen als Entitäten-IDs, TEI-XML als Auszeichnungssprache, strukturierte Metadaten als Objektbeschreibung und generell die Anwendung möglichst permanenter Referenzierungen sind sol-

che Komponenten und Strukturmerkmale. Welche weiteren Komponenten zu dieser überschaubaren (und nicht originellen) Liste in naher und nicht so naher Zukunft hinzukommen, lässt sich derzeit nicht mit gutem Gewissen sagen. Entscheidender aber ist das Problem der nachhaltigen Wartung, also Bereithaltung und Kuratierung dieser Komponenten zu lösen.

Eine Lösung dieses Problems ist in Sicht: Derzeit werden die wissenschaftspolitischen Voraussetzungen dafür geschaffen, dass es in naher Zukunft (auch) in Deutschland eine nationale Forschungsdateninfrastruktur (NFDI) [6] geben wird. Hier entsteht eine neue zivile Schnittstelle, durchaus in Analogie zum Aufbau eines professionellen, für die Versorgung der Öffentlichkeit zuständigen Bibliotheks- und Archivwesens im späten 19. und 20. Jahrhundert. Beide Infrastrukturen, die für die Bewahrung analoger wie digitaler Informationen, Daten und Datenträger entstanden und entstehen nicht im Rahmen einer Legislaturperiode. Es sind – wie viele Neugründungen im Zuge und am Ausgang der digitalen Sattelzeit – Aufgaben für eine ganze Generation. Diese neue Infrastruktur für die Sicherung unserer wissenschaftlichen Produktivität wird digitalen Einzelprojekten, auch geisteswissenschaftlicher Grundlagenforschung wie sie die *edition humboldt digital* verkörpert, eine dauerhafte Versorgung garantieren, egal, ob diese zwei, fünf oder fünfundzwanzig Jahre lang aktiv entwickelt werden.

## Literatur

1. Deutsche Forschungsgemeinschaft (2015) Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft. [http://dfg.de/download/pdf/foerderung/grundlagen\\_dfg\\_foerderung/informationen\\_fachwissenschaften/geisteswissenschaften/foerderkriterien\\_editionen\\_literaturwissenschaft.pdf](http://dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf), letzter Zugriff: 9.12.2018
2. Deutsche Forschungsgemeinschaft (2013) DFG-Praxisregeln „Digitalisierung“. [http://www.dfg.de/formulare/12\\_151/12\\_151\\_de.pdf](http://www.dfg.de/formulare/12_151/12_151_de.pdf), letzter Zugriff: 9.12.2018
3. <http://edition-humboldt.de>, letzter Zugriff: 9.12.2018
4. <http://edition-humboldt.de/api/v1/tei-xml.xml>, letzter Zugriff: 9.12.2018
5. Jabr F (2013) The Reading Brain in the Digital Age: The Science of Paper Versus Screens. *Scientific American*
6. Rat für Informationsinfrastrukturen (2016) Positionspapier „Leistung aus Vielfalt“. <http://www.rfii.de/de/themen/>, letzter Zugriff: 9.12.2018
7. Sahle P (2013) Digitale Editionsformen – Teil 1: Das typografische Erbe. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. *Books on Demand*, Norderstedt
8. Schöch C (2016) Ein digitales Textformat für die Literaturwissenschaften: Die Richtlinien der Text Encoding Initiative und ihr Einsatz bei Textkonstitution und Textanalyse. *romanische studien* 2:325–364
9. TEI Consortium (2016) TEI P5: Guidelines for electronic text encoding and interchange. <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>, last access: 9.12.2018
10. Wulf A (2015) *The Invention of Nature: Alexander von Humboldt's New World*. Alfred A. Knopf, New York

# *e-Science in Scandinavia*

*The Case of the Swedish e-Science Research Center*

Olivia Eriksson · Erwin Laure  
Erik Lindahl · Dan Henningson  
Anders Ynnerman

## **The Swedish e-Science Research Centre (SeRC)**

The Swedish e-Science Research Centre (SeRC) is based on a collaboration between four Swedish universities: The KTH Royal Institute of Technology (KTH), Stockholm University (SU), Karolinska Institutet (KI) and Linköping University (LiU). SeRC's mission statement is to develop state-of-the-art e-Science tools and provide e-infrastructure support to existing and emerging e-Science research communities to help bring about scientific breakthroughs in Sweden. SeRC was founded in 2010 as the result of the Strategic Research Area (SRA) initiative launched by the Swedish Government Bill on Research Policy in 2008, where a total of 24 different strategic research areas were defined – one of which was e-Science. Initially SeRC was granted funding for 5 years. During those first 5 years, SeRC built up an organization for e-Science research, which has been highly successful. This was reflected in the excellent grades that SeRC received when the SRAs in Sweden were evaluated in 2015, and the fact that after this, SeRC received funding for at least 5 more years. This new phase of SeRC partly focuses on activities relating to emerging technologies (such as exascale systems and data-driven science) while also consolidating SeRC's ongoing efforts in working towards a long-lasting e-Science environment in Sweden.

## **e-Science and SeRC**

We are right in the middle of a revolution where computers are redefining the process of scientific research as we know it. While this trend started in the natural and engineering sciences, it has evolved to become a pervasive component of virtually all

scientific fields: modern aeroplane design depends more on computational fluid dynamics than on wind tunnels, bioinformatics has turned biology into a quantitative subject, and we are seeing entirely new data-driven research, for instance in medicine.

Many of the challenges that are being faced within e-Science are similar across different scientific disciplines; all these research directions rely on having access to large computational resources and the ability to work with and archive huge amounts of data, and they also need new algorithms (for performing both computations and large-scale analysis of data). Consequently, there is also a demand for e-Science experts, computer experts who are highly skilled in a particular applied research area and also in e-infrastructures or method development.

## **SeRC's scope, model, and mission**

The concept of e-Science, in its most basic form, is simply the notion of using digital information and processing that information to gain new scientific insights. This means that a lot of e-Science research can be conducted using the standard hardware and software that are widely available. However, there are also research problems that require additional resources, which is where SeRC comes in.

---

<https://doi.org/10.1007/s00287-018-01133-2>  
© The Author(s) 2018. This article is available on SpringerLink.com with Open Access.

---

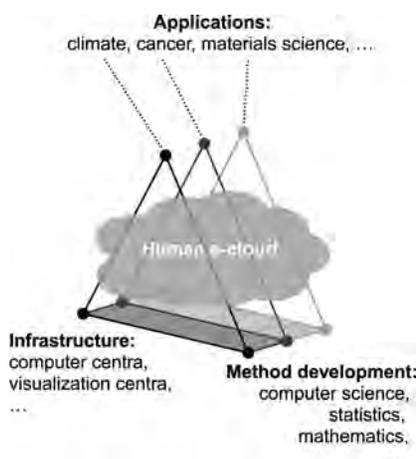
Olivia Eriksson · Erwin Laure · Erik Lindahl  
Dan Henningson · Anders Ynnerman  
SeRC, SE-100 44 Stockholm, Sweden  
E-Mails: olivia@mech.kth.se, erwinl@pdc.kth.se,  
erik.lindahl@scilifelab.se, henning@mech.kth.se,  
anders.ynnerman@liu.se

SeRC's scope is to address the e-Science problems that need solutions which go beyond the standard ones that are currently available, and that require researchers to have access to a state-of-the-art e-infrastructure as well as new e-Science methodology and technology. SeRC is thereby pushing the development of new e-Science tools forward. In this way, SeRC is paving the path for new high-end e-Science that will eventually mature and benefit not only the groups involved in SeRC but also the whole academic base of e-Science users.

To address these kinds of problems, SeRC is based on a three-pillar model (Fig. 1). The three pillars of existing excellence in application areas, core e-Science and method development, and computational infrastructures interact to find new e-Science solutions. In particular, this is done through collaborations and e-Science experts active in at least two of these pillar areas. For the long term, SeRC is working on developing a "human e-cloud" in order to better address challenges that are common to different application areas, and also to provide clear career paths for this new kind of e-Science researchers with dual expertise.

### Pillar 1: application areas

One of SeRC's central goals is to have a direct impact in application areas of strategic relevance for our partner universities, and we also want to focus on research topics where SeRC has world-leading research groups. These SeRC application areas include fields



**Fig. 1** Three pillars of existing excellence in application areas, method development, and e-infrastructures, connected through a "human e-cloud"

such as computational fluid dynamics, materials science, bioinformatics, cancer epidemiology and screening, neuroinformatics, molecular simulation, and climate modeling.

### Pillar 2: e-infrastructures

SeRC wants to facilitate the development of a world-class computational infrastructure at the PDC Center for High Performance Computing at KTH, the National Supercomputer Center at Linköping University, and at the visualization centers affiliated with SeRC (which are based in Norrköping and at KTH in Stockholm). The development of this infrastructure includes competitive hardware investments in collaboration with the Swedish National Infrastructure for Computing (SNIC), although it is equally important to recruit advanced infrastructure experts to develop key simulation codes further.

### Pillar 3: core e-Science and method development

An integral part of SeRC is leading-edge research in core e-Science and method development, for instance in fields such as numerical analysis, visualization, parallelization, acceleration, and data engineering. SeRC has several core research groups that provide both the latest research results for e-Science tools, as well as a broad competence base in e-Science tool usage and methodology.

### The human e-cloud

For SeRC, the central challenge is to get these three pillars to strengthen and support each other through collaborations. Infrastructures and core e-Science do not exist in a vacuum but need to engage directly in research on important applications. In particular, SeRC focuses on supporting application areas where we see an opportunity to translate advances in fundamental e-Science into research impact.

Our primary tool to achieve this is the recruitment of cross-disciplinary e-Science experts that work between these areas – we call them the "human e-cloud" in Fig. 1. These experts combine a deep knowledge of an application area with expertise either in method development or large-scale computational infrastructures. By integrating the application experts into the applied research environments (while they work close to the method development groups or the infrastructure), it



**Fig. 2 Example of SeRC collaboration**

provides opportunities for close collaborations between e-Science experts and applied researchers, as well as between e-Science experts facing similar challenges (for instance, in parallelization, data management, or method development) in different application areas.

Figure 2 introduces “e-Science for Cancer Prevention and Control” as an example of SeRC collaboration. This program has an e-Science team that includes epidemiologists, bioinformaticians, image analysts, and computer scientists who develop, implement, and evaluate methods for using large amounts of biomedical data for risk-stratification in cancer screening and diagnosis. More details are available at <http://www.e-science.se/ecpc>.

### Organization

SeRC is based on a number of research communities, together with the SeRC e-Science infrastructure. Some of the SeRC research communities focus on core e-Science methodology, like numerical analysis, parallel software and data engineering, and visualization, whereas others focus on applying the methodology in a particular area, like bioinformatics, neuroinformatics, cancer epidemiology and screening, molecular simulation, electronic structure, computational fluid dynamics, or climate research. The infrastructure on the other hand consists of:

- two supercomputer centers – the PDC Center for High Performance Computing at KTH (PDC) and the National Supercomputing Center at LiU (NSC), both of which are part of SNIC.

- Two visualization centers – the Visualization Center C (VC), located in Norrköping, and the Visualisation Studio at KTH (VIC).
- A number of e-Science experts who are employed at the supercomputer center, and who work closely with the research communities.

SeRC has been co-financing about 30 projects and 15 faculty positions per year but is now in the process of reallocating the funding to a number of larger multidisciplinary collaboration programs initiated in order to increase the level of collaboration between researchers in the areas of applications, methodology, and infrastructure.

To consolidate the organization, in 2013 SeRC was turned into a research center with KTH as the host, and shortly after we recruited an external chair of the board, Professor Morten Dæhhlen, who is Dean of Mathematics and Natural Sciences at the University of Oslo and a prominent international e-Science leader. More information about the SeRC organization can be found at <http://www.e-science.se/org>.

### SeRC faculty

Since the inception of SeRC, we have recruited a substantial number of highly skilled assistant professors. In addition, we have selected a number of e-Science investigators who were either hired recently or who recently joined SeRC with support from other funding. Together, these researchers form the SeRC faculty.

The main goal of the SeRC faculty is to facilitate interdisciplinary exchange between e-Science

communities in application areas and core e-Science and method development, and to provide SeRC with a network of e-Science researchers that straddle the various research communities and multidisciplinary collaboration programs.

### **Collaborations and scientific output**

SeRC hosts a large number of outstanding scientists who are pushing the state-of-the-art, in application areas as well as method development, with several high-impact results. However, SeRC's mission is to go a step beyond merely providing some marginal funding and instead build a scientific environment that promotes excellence within research disciplines through multidisciplinary collaboration. SeRC makes it possible to initiate new, ambitious and risky collaborative projects that have the potential to redefine research fields. In particular, the center has acted as a catalyst in the formation of several strong research environments that work on the application, method development, and infrastructure aspects of computational research (for example, see Fig. 2). These multidisciplinary collaborative projects, for instance, make it possible to engage computational experts in long-term development efforts that have previously been very difficult to fund.

The new initiatives have had tremendous scientific impact: teams that did not even know each other prior to SeRC have merged, and the collaborations between infrastructures and researchers have led to a range of new high-profile international research projects being funded, not to mention the award of several prestigious European Research Council grants. However, instead of being content with these achievements, SeRC will keep raising the bar by continuously evaluating our collaborations with a view to further improving the center's effectiveness. SeRC's goal is not to maintain the status quo, but to promote discoveries and new applications by helping and encouraging talented researchers to team up and leave their comfort zone. Collaboration, multidisciplinary, change, and impact are the genes of SeRC.

### **Computing resources**

e-Science is critically dependent on efficient large-scale computing and storage resources. Within SeRC, these resources are provided by the leading Swedish high performance computing centers, PDC and NSC, which offer more than two-thirds of the

Swedish national computing capacity. A strategic partnership between these centers has been established. This allows for better coordination and alignment of their services with the needs of SeRC researchers.

Since 2015, SeRC researchers have had access to petascale computing resources through these centers. However, it is noteworthy that PDC and NSC not only provide hardware, they also provide expertise in many application areas in terms of computing experts who work closely with certain scientific domains, including bioinformatics, climate research, computational fluid dynamics, materials sciences, molecular dynamics, and neuroinformatics.

In January 2015, the SeRC infrastructure received a significant boost with the inauguration of Beskow at PDC – a 2-petaflop Cray XC40 system that is more powerful than all the other SNIC systems together. With Beskow, SeRC researchers had a world-class petascale system at their disposal for the first time – it was much needed for their research and also as a stepping stone towards using the larger European-level HPC infrastructure. Beskow was designed to cater for the needs of SeRC and important SeRC applications were used as benchmarks to guide the selection process.

In the autumn of 2017, NSC's flagship system, Triolith, was replaced with another petascale system, ensuring that SeRC researchers and other e-Science researchers in Sweden continue to have access to the latest hardware technologies.

### **The future and e-Science agenda**

During its first 5 years, SeRC had tremendous success in establishing e-Science in Sweden and fostering the e-Science paradigm in many research areas. e-Science is now well established and is a priority area at the SeRC partner universities. Starting up the next phase of SeRC we are consolidating these efforts by working towards a long-lasting e-Science environment and, on the other hand, increasing SeRC activities relating to emerging technologies for e. g. exascale systems and data-driven science.

The report Swedish Science Cases for e-Infrastructure (<http://www.e-science.se/sscei>) is an important document when it comes to shaping the future of e-Science in Sweden. This document discusses strategic e-Science development in Sweden and makes a number of recommendations that

are of help in guiding future SeRC developments. Key findings, from a SeRC perspective, include the following.

- “Development of methods, tools and software within core disciplines is necessary to make breakthroughs.”
- “Advanced and long-term user support and human infrastructures are keys to e-Science adoption.”
- “The simulation paradigm dominates the current Swedish needs for e-infrastructure. A complementary and more data-centric aspect of e-Science should be promoted.”
- “e-Social science and e-humanities are potentially very large users, but need active support, like other research communities new to e-Science.”
- “e-Science methods and tools are in increasing demand and will be instrumental in increasing interaction between tool makers and tool users.”

Responding to these findings entails work at several levels, ranging from infrastructure and tool development to policy definitions. In its strategy process, SeRC has defined an agenda for the second phase. This agenda addresses many of the items from the report and describes how SeRC is in a unique position to spearhead this work.

### The impact of e-Science

The impact of e-Science can be seen in all areas of academia. The use of basic e-Science in new research domains can generate fundamentally new and important results. At the same time, access to advanced e-Science tools and infrastructures enables the leading research groups that are spearheading Swedish e-Science to compete at the international forefront. In view of this, SeRC intends to intensify the collaboration with relevant Swedish groups and universities that are currently not part of SeRC, in domains like bioinformatics, numerical analysis, and materials science. Also, the broader adoption of e-Science methodology cannot be overestimated. Through extensive collaborations with industry and governmental agencies, SeRC has a direct impact in society and contributes to societal development and commercial competitiveness. SeRC has several industrial representatives on its advisory board, and we are delighted with their interest and engagement in learning how computing can change what they do,

for instance at our annual meetings. One of the primary indirect mechanisms for SeRC to have an effect is through knowledge transfer in the form of recent Ph.D. graduates, who have unique experiences of e-Science and form a network consisting of the next generation of e-scientists.

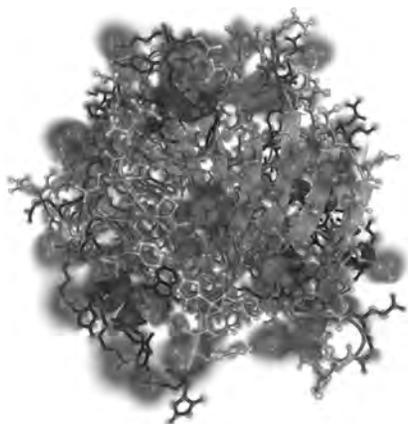
The single largest overall success of SeRC is that it has helped application researchers and computational experts to find each other and speak with one voice. Large-scale data analysis and simulation have been introduced in areas where they have never been used before, the strongest computational groups in the country suddenly collaborate internationally instead of competing nationally, and we will keep re-defining these boundaries by identifying areas where the combination of computational and application research excellence makes a field ripe for disruptive changes to the science.

### SeRC initiative: multidisciplinary collaboration programs

In the first phase of SeRC, larger, multidisciplinary constellations were prototyped through a multidisciplinary collaboration program (MCP) promoting multidisciplinary approaches to strategic research areas with high potential for gains. The e-Science for Cancer Prevention and Control (eCPC) was SeRC's first MCP, which was followed by “Visualisation in e-Science Applications” and “SESSI-SeRC Exascale Simulation Software Initiative”.

The illustration from the SeRC Multidisciplinary Program “Visualisation in e-Science Applications” (Fig. 3) shows a visualization of a protein, combining hybrid data from different sources using a hybrid visualization of geometric protein representations and volumetric data. The improved A-buffer algorithm is capable of rendering the entire scene with a performance increase of 5.9 times compared to prevalent techniques.

The illustration from the SeSSI MCP (Fig. 4) shows results from computational modeling of the skin barrier (the lipid matrix of the stratum corneum) using molecular dynamics simulations, which gives us further insight into the largest organ of the human body and will also further clinical experiments. Thanks to the highly optimized heterogeneous parallelization in the GROMACS code, complex computational studies can be carried out quickly and efficiently. The image on the right shows the modeled molecular system with ceramide



**Fig. 3 Illustration from the SeRC multidisciplinary program “Visualisation in e-Science Applications”**

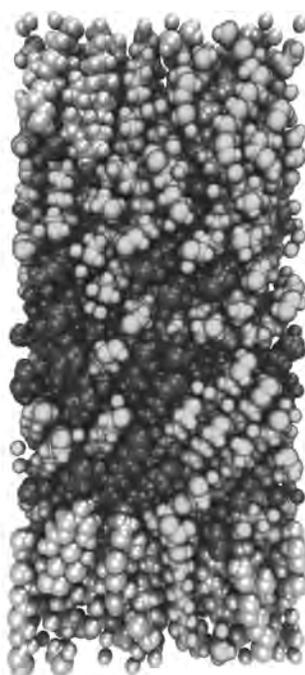
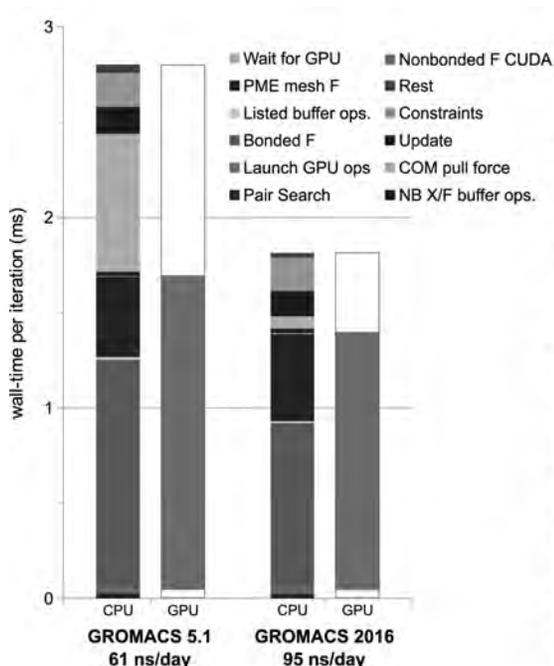
molecules in green, cholesterol in white, and fatty acids in red. Recent work in SeSSI (that focused on improving SIMD, GPU, and thread parallelization) resulted in speeding up the calculations by up to 50 %. The figures on the left show the execution time breakdown of the CPU and GPU tasks in GROMACS versions 5.1 and 2016; improved performance of multiple tasks leads to an increase in simulation throughput from 61 ns/day to 95 ns/day. The simulations were performed on a workstation equipped

with a Core i7-5960X CPU and a GeForce TITAN X GPU.

SeRC will now build on these successes and further develop the MCPs by consolidating the SeRC efforts into larger constellations. A key requirement of MCPs is the existence of external third-party support contributing to the program goals. These multiple funding streams are one component to ensure the sustainability of SeRC efforts. Recently, three additional MCPs have been started: Data-Driven Computational Materials Design, SeRC Data Science, and Brain-IT.

### SeRC initiative: scientific computing lab

SeRC groups are engaged in the development of computational methods and infrastructures, which are necessary ingredients for world-class computer-based scientific research. With architectural changes, which are being driven partly by the challenge of building exascale systems, these efforts need to be reinforced and strengthened. Developments – such as massively increased core counts, large vector units, deep memory hierarchies, and the like – require significant efforts in the development of efficient, scalable methods and implementations.



**Fig. 4 Illustration from the SeSSI MCP. (Source: <https://doi.org/10.6084/m9.figshare.3423395.v1> (Magnus Lundborg and Szilárd Páll) – image reused under CC-BY)**

Such work is planned in the MCPs referred to earlier. These constitute ideal cases for the SeRC vision of integrating researchers (from the three pillars that were described earlier) with each other in such a way that collaboration between application area, e-infrastructure, and methods researchers is being facilitated by the e-Science experts in the human e-cloud.

Following international examples, for instance from the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, we intend to prototype a scientific computing lab, where researchers working on applications, methods, and infrastructure can effectively collaborate in such a manner. A physical interaction space that has been created at PDC is a key component in this endeavor. With this lab, SeRC will build the nucleus of the human e-cloud identified as a key ingredient above.

### **SeRC initiative: data-driven science**

The data deluge that is arising from new scientific instruments, large simulations, and data on the internet, together with new technologies for handling and analyzing this data, has significant impact on e-Science methodology. SeRC is, therefore, embracing big data, with an initial emphasis on the life sciences. This includes research regarding storage, manage-

ment, integration, visualization, and analysis of big data.

In recent years, new platforms and tools have appeared that can efficiently and cost-effectively store and process up to petabytes of data. This allows many organizations to handle data that was previously considered to be too expensive to store and manage. This data-driven science trend is transforming scientific research, by making data-driven discovery and prediction possible. The requirements for data analytics on a large scale include data management on public and private clouds, data-parallel algorithms, and proficiency with a complex ecosystem of tools and platforms such as Hadoop. Currently, these are not widely available within the existing SeRC research communities, so one of the goals for SeRC is to help transfer knowledge about the use of methods and tools from data science to facilitate the use of new research methodologies in other SeRC communities.

**Open Access.** This article is distributed under the terms of the Creative Commons Attribution Non-commercial License (<http://creativecommons.org/licenses/by/4.0/deed.de>) which permits any non-commercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

# *eScience development and experiences in The Netherlands*

Wilco Hazeleger · Tom Bakker  
Rob van Nieuwpoort

## **Introduction**

Historically, researchers have been aided by scientific instruments and methodologies to address their scientific questions. In the modern era, digital technology and methods are omnipresent, and their use is quintessential to current scientific discovery. It has been recognized that ever growing data sizes and increased availability of disparate data sets pose tremendous opportunities for research, but effective use of these data challenges scientific researchers [5]. Moreover, computer capabilities keep increasing, while computer architectures continuously change. Data and computer challenges are tightly coupled as analyses of large and disparate data sets require modern computational resources.

The “big” sciences, such as high energy physics, astronomy, materials science, earth and climate science, are prime examples of research domains that have developed strongly over the past decades due to the availability and development of computational resources [2, 3] With the rise of data availability and with developments in novel applied mathematical techniques, such as machine learning, new domains are increasingly profiting from the digital revolution. In different aspects of the life sciences this started about a decade ago, drug discovery [7] and genomics being well-known examples, and increasingly also scholars in social sciences and humanities (e. g. [12]) are using digital technology and analytics techniques. In general, scientific research is becoming more data intensive and inductive of character. So, in modern research, nearly all researchers also become data scientists, the complexity of projects increases, and research is inherently multidisciplinary.

eScience effectively applies modern e-infrastructure (computing, storage, networks) and utilizes research software and data to address challenging scientific questions posed in research disciplines. eScience is developing into a scientific field itself as well. eScience has engineering aspects, similar to civil or chemical engineering, by applying novel technology and methods to other application domains. These approaches themselves, developed at the interface of domain challenges and computer and data science, result in questions that are addressed with a scientific approach. For instance, scaling up or scaling out using high performance and distributed data and computer resources is a typical eScience challenge. Moreover, new analytics methods and access to distributed data, developed in the domains of data and computer sciences, generate new scientific challenges when effectively applied.

In the Dutch research landscape, The Netherlands eScience Center (the eScience Center from now on) was set up in 2011 to effectively build an interface between domain research and e-infrastructure, as well as an interface between computer and data science and domain research. This need was recognized after successful large projects on eScience

---

<https://doi.org/10.1007/s00287-018-01132-3>  
© The Author(s) 2018. This article is available on SpringerLink.com with Open Access.

---

Wilco Hazeleger · Tom Bakker · Rob van Nieuwpoort  
Netherlands eScience Center,  
Amsterdam, The Netherlands  
E-Mail: {w.hazeleger, t.bakker, r.vannieuwpoort}  
@esciencecenter.nl

Wilco Hazeleger  
Wageningen University,  
Wageningen, The Netherlands

Rob van Nieuwpoort  
University of Amsterdam,  
Amsterdam, The Netherlands

## Abstract

The Netherlands eScience Center is the national expertise center for the development and application of research software. Collaborating with researchers from all academic disciplines, it extends the breadth and depth of research opportunities by exploiting the latest insights from computer and data science, as well as making optimal use of hardware, software, and data infrastructures. It does so through problem-driven research projects where eScience research engineers, employed by the eScience Center, collaborate with researchers in all disciplines at Dutch academic institutions. Project software is generalized and made available for reuse for other disciplines and goals. The center has three main technological competences: efficient computing, optimized data handling, and data analytics. Furthermore, on the national level it coordinates and contributes to science policies on computing, data, and applications thereof. With its two main assets, a staff of highly educated and multi-disciplinary eScience Research Engineers and an open online directory of research software tools and knowledge, it successfully contributes to the Dutch scientific landscape and enhances and accelerates all research in The Netherlands and beyond.

(e.g., VL-e; Virtual Laboratory on eScience) and the successful integration of academic e-infrastructure facilities and services of SURF. SURF (the collaborative organisation for ICT infrastructure in Dutch education and research) and NWO (The Netherlands Organization for Scientific Research, effectively the Dutch research council) jointly set up the eScience Center. The eScience Center is positioned between scientific research and e-infrastructure in the national landscape and connects the national e-infrastructure with academic disciplines (see Fig. 1).

Its strategy and implementation will be discussed below, as well as examples of eScience projects.

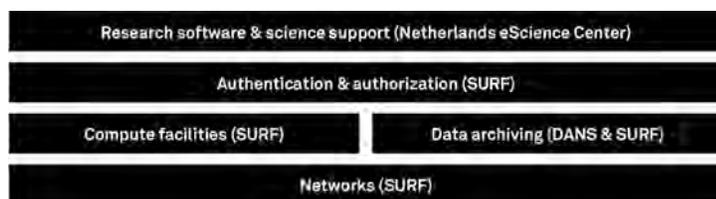
## Strategy

In response to the changing research environment described in the Introduction, we developed a strategy that positions the eScience Center as an organization with interfaces to domain research, computer and data science, and e-infrastructure, while delivering eScience services and output. Its strategy is based on the mission, which is to *enable digitally enhanced research through efficient utilization of data, software, and e-infrastructure*. This translates into four tasks: to enable scientific breakthroughs, to collaborate on problem-driven projects, to develop versatile cross-disciplinary eScience tools, and to coordinate eScience activities (see the strategy: <https://www.esciencecenter.nl/about/strategy>).

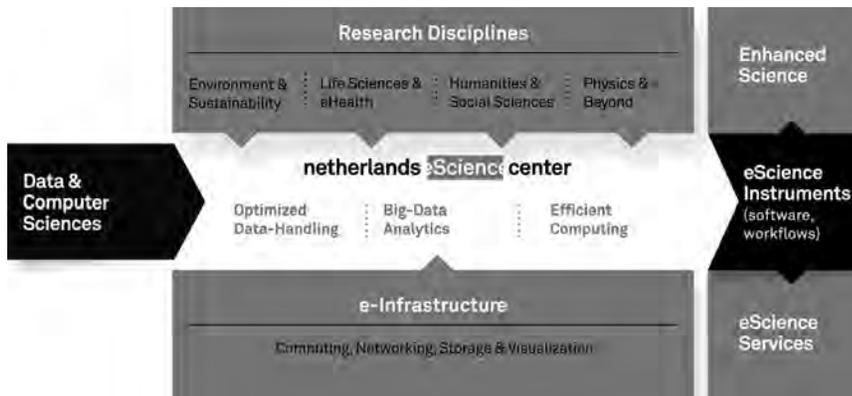
It is crucial that the eScience Center takes the domain research perspective as a starting point and works in a problem-driven manner, primarily through calls for collaborative project proposals. Figure 2 summarizes the strategy and shows the different interfaces. It should be read as an arrow from left to right, with the eScience Center in the middle and with different domains and institutions feeding into the eScience Center. The right-hand side shows the outcome of our activities. The strategy is explained in detail below.

## Interface with domain research

The eScience Center addresses all research disciplines (top part of Fig. 2), but distributes its efforts in the domains over Physics & Beyond (this includes high and low energy physics, fluid dynamics, materials sciences, and astronomy), Environment and Sustainability (this includes climate research, ecology, energy research), Life Sciences and eHealth (this includes genomics, proteomics, medicine), and Humanities and Social Sciences. The eScience Center's work is driven by scientific problems from these disciplines. We grant, in competition, collaborative



**Fig. 1 Positioning of the Netherlands eScience Center in the national digital infrastructure for science and education**



**Fig. 2 Schematic of the strategy of The Netherlands eScience Center, focusing on the eScience technical competences (optimized data handling, big data analytics, and efficient computing) and the different interfaces**

project proposals for 1 to 4 years, in which one or more eScience Research Engineers work with PhD students or postdocs at a university or academic institution (see “Implementation of the strategy” below). The four domains are deliberately chosen such that there is a balance in the level of literacy with digital technologies. That is, the four domains do not compete amongst each other for eScience support, but are all addressed. This also facilitates reuse of technologies and cross-fertilization of digital technologies among research disciplines.

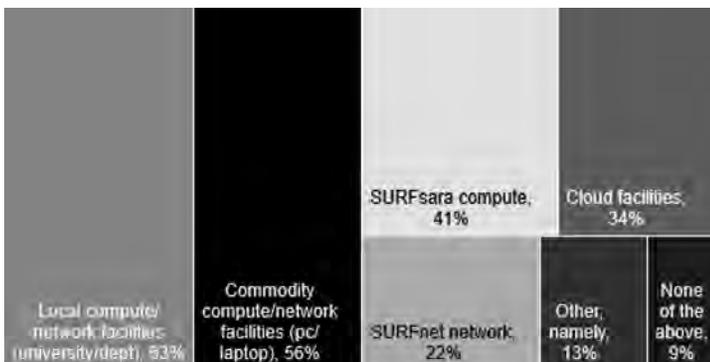
### Interface with computer and data science

Computer and data sciences are very quickly developing research disciplines. New opportunities in digital technologies and analytics techniques are investigated and developed in computer and data sciences. For the eScience Center to continue to be up to date with recent developments and scout relevant technologies and techniques for applications in domain research, collaborative projects are set up with computer and data scientists. These are also competitive multiyear projects, similar to projects

with application domains, but with the specific objective to strengthen our own capabilities in our three main technological competence areas: efficient computing, data analytics, and data management.

### Interface with e-infrastructure

Computing, data, and network facilities are important tools in digitally enhanced research. The eScience Center does not have its own hardware facilities but uses state-of-the-art e-infrastructure facilities that match the research needs at hand. SURF is the main national e-infrastructure provider (with SURFsara providing high performance computational and storage facilities), but we also use local facilities at universities and commercial (cloud) providers. There is a preference to use national public facilities, but other facilities can be used when needed and when they match the scientific problem at hand better. In each project, a consultant from SURF is available for individual support and advice. In total, over half of the projects use SURFs e-infrastructure (Fig. 3), while there is also a lot of use of local facilities and commercial providers of



**Fig. 3 Use of e-infrastructure in eScience projects based on a survey of 38 projects conducted with support from the eScience Center**

infrastructure, in particular the different commercial cloud providers. There are large differences in e-infrastructure use in the different disciplines, so a tailored approach is needed.

## Outcomes of the eScience Center

The outcomes of eScience Center projects are peer-reviewed papers and research software, both in the domain sciences and in eScience. A leading principle of the eScience Center is to develop and apply reusable and sustainable open-source software within multiple projects. To promote this principle and to increase the chance for growth and adoption, we set up an online research software directory to increase the adoption of research software that we develop, enhance, and have expertise with. Some of the projects reach beyond the realm of academic research and develop into services for communities (e. g., in the health domain, an infrastructure for translational research has been set up).

## Implementation of strategy

### Call strategy

The eScience Center fulfils its mission primarily by carrying out collaborative scientific research projects. These projects are selected after a call for proposals to the academic community. Potential project leaders from academia are invited to develop proposals, matching their purely scientific ambition with the digital technology and methods needed to achieve that. The typical project size is 500 k€, half of which is funding for a PhD or postdoc position at the academic partner institution, while the other half constitutes in-kind contribution to the project by one or more eScience Research Engineers employed by the eScience Center. The project proposals are evaluated on their scientific quality and the eScience quality. This last criterion includes the potential for impact on other disciplines, the sustainability of the tools beyond the duration of the project, and the innovation for the problem at hand. The current success rate of applications is low due to the very large demand (~10–15 % of the proposals are granted funding). Typically, one project in each of our four scientific “domains” is funded, while also two collaborative projects are granted with researchers from computer and data science. Recently, in collaboration with other funding programs of NWO, the number of granted projects has increased. By col-

laborating with existing funding programs, more and larger-impact projects can be set up in existing research communities. Recent examples include a joint call with CLARIAH, a humanities research infrastructure, and joint calls with the FOM/Shell program on computational sciences for energy research. Since 2011, up to 60 large projects have been conducted, distributed over all scientific domains, and covering almost all Dutch universities.

## eScience Research Engineers

The *eScience Research Engineers* require special attention in this article, as they play an important role in the implementation of eScience in The Netherlands. They are digital specialists at academic level (MSc to PhD level up to associate professor level). They are employed by the eScience Center, where they follow a career track with a technological, research or managerial emphasis. They are specialists in one or more of our three major technological competence areas (efficient computing, data analytics, and data management). They typically spend two-thirds of their time at the academic research institution where eScience projects are executed, while the remaining time is spent at the eScience Center in Amsterdam. Here, they collaborate with other eScience Research Engineers, sharing best practices, tackling technological project challenges, and performing joint research and development activities. On purpose they have been labeled *research engineers*, as they perform not only software development and engineering tasks but are part of a research team. For instance, they jointly publish the results in peer reviewed literature. Each project has been assigned one or more eScience Research Engineers, depending on the skills and competences required. This allows us to effectively combine and contribute disparate skills to an individual project (e. g., data visualization and high-performance computing). For regular academic groups, this is difficult to achieve. The eScience Center currently (mid-2018) employs more than 50 eScience Research Engineers.

Other advantages of our approach are that the research engineers can identify and exploit opportunities for novel methodologies and the reuse of existing tools *across different disciplines*. Because they collaborate in many inter-disciplinary projects in different domains, they have a birds-eye view of technology in science, allowing them to make quick and efficient progress, using the right tool for the

task. Moreover, they develop software with general applicability and reuse in mind from the start.

### eScience technologies

As mentioned before, the eScience Center's competences are focused on efficient computing, big data analytics, and optimized data management. The required skills largely depend on project demands, which requires a flexible and adaptive staff. This is accomplished through continuous training and working in teams leading to consistent internal knowledge management and through mobility of staff to and from the eScience Center.

eScience technologies vary from algorithms to the tools for utilization of advanced computer infrastructures and virtual research environments (e.g., computer kernels, libraries, scientific workflows, data tools and applications to use storage facilities, high-speed networks, visualization equipment). Another relevant development is the emergence of the user-friendly Jupyter Notebooks, which has a significant impact on reuse and reproducibility. This technology supports the open science development, which is stimulated by the Dutch government and European Committee, for instance as part of the European Open Science Cloud.

### Reuse and generalization

An important aspect of the development and application of eScience technologies is generalizing tools and making them applicable and re-usable in different disciplines. The importance of academic software is increasingly recognized (e.g. [1]). To enable reuse of our software, we make it publicly accessible, engage in active online and offline promotion, and use open software licenses. To increase and maintain software quality, we use standard off-the-shelf cloud-based software testing and source code quality assessment tools. To maximize transparency, all code and test results are publicly available on our GitHub repositories. The eScience Center created a checklist<sup>1</sup> containing the essential steps for creating high quality software, which is applied to all tools developed. The software is – if possible – released under the permissive open-source Apache 2.0 license<sup>2</sup>. For documentation and

data sets, we use the permissive Creative Commons licenses.

To further improve findability and re-usability, we set up the Research Software Directory: <https://www.research-software.nl> (Fig. 4). The aim is to provide an extensive and curated online portal of advanced scientific software, including (1) our own software, (2) external software to which we have made contributions, and (3) externally developed software that we have significant expertise with. The added value compared to our GitHub repositories is that it helps researchers to get a high-level overview of available relevant software and related eScience Center projects. Moreover, it has increased the visibility of our contributions to leading external software projects, such as ROOT, Amuse, OpenDA, and libLAS. We associate DOI's with all software releases, stimulating software citation. This is equally important for scientific reproducibility, and for giving credit to software as research output. The latter is key for the career perspectives of research software engineers.

A major recent development in this respect is the use of Jupyter Notebooks. These offer a web-based interface for interactive programming in Python, R, and a wide range of other languages. This has vastly improved accessibility to a wider audience with less experience in using digital tools. Furthermore, we often package our software. Because many programming languages have package managers that offer both a central location to find software and simple tools to install it, we release software via package managers such as PyPI and Anaconda for Python, NPM for JavaScript, and Maven for Java. Successful examples are ROOT Conda recipes, to which we contributed, and which have been very successful in the high-energy physics community. Finally, the eScience Center increasingly uses Docker virtualization to distribute end-user applications. "Dockerizing" software decreases the dependency problems often encountered in legacy applications.

In addition to the Research Software Directory, a shared knowledge base is maintained (<http://knowledge.esciencecenter.nl>) to share guides for software development, tutorials, and technical reports. One of its goals is to retain knowledge when projects end, or engineers leave the eScience Center.

### Example eScience projects

In this section, we highlight a few of the eScience collaborative projects in different domains.

<sup>1</sup> <https://nlesc.gitbooks.io/guide/content/>.

<sup>2</sup> Apache 2.0 allows anyone to re-use our software in any way, including commercial application.

## Research Software Directory

Encouraging the re-use of research software

The Research Software Directory aims to promote the impact, the exchange and re-use of research software. Please use our tools! [Read more](#)

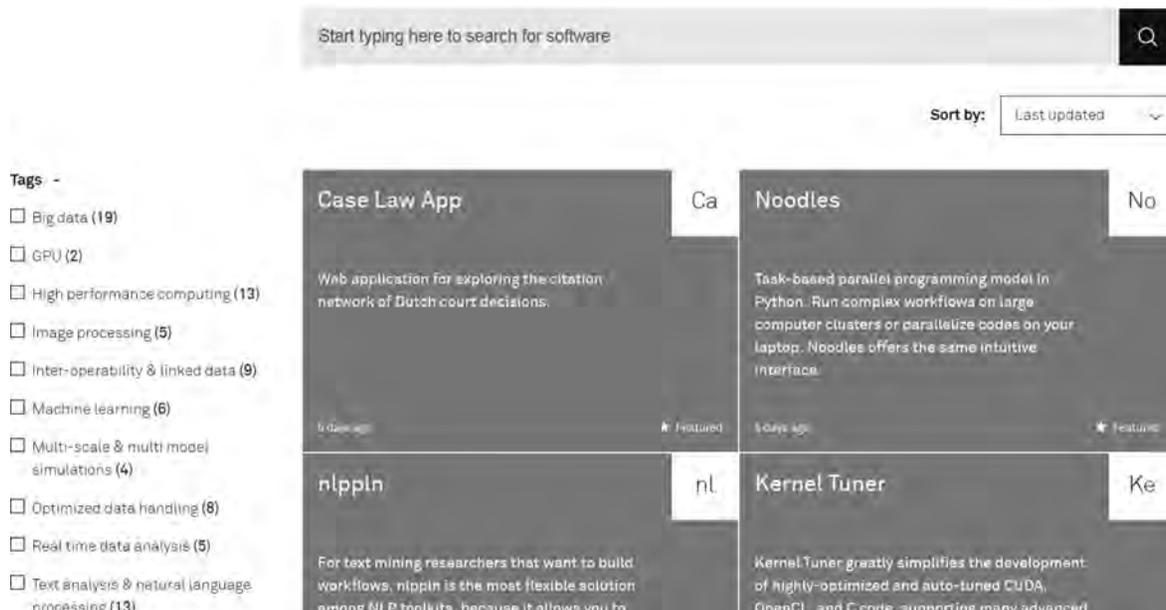


Fig. 4 Screenshot of the research software directory

In the *Summer in the City* project, we co-developed a very detailed weather model of the city of Amsterdam in collaboration with Wageningen University, integrating data sources of the earth's surface and near surface atmosphere. The project used existing software (the WRF mesoscale model), with the eScience challenge to obtain realistic bottom boundary conditions (i. e., the land surface characteristics) by integrating data sets. These include satellite data, weather and hydrologic data, and socio-economic data. Such improved boundary conditions and higher resolution improved the quality of local weather forecasts at the sub-kilometer scale. The scientific results published within the project were followed up by a successful grant application by one of the researchers in the project, which ensures further scientific development of meteorology in an urban environment. The results

are described in detail in a recent paper by Ronda et al. [8].

In the *eSalsa* project, we enabled global ocean modeling at very high spatial resolution using GPU computing technologies in collaboration with Utrecht University. An existing numerical ocean model code was optimized for execution on the CPU-GPU Cartesius supercomputer of SURF-Sara [13]. Century scale simulations, up to 2100 under climate change, are extremely challenging at the eddy-resolving spatial resolution (equivalent to about 10 km horizontal scale; [4]). Also, natural variability arising in the climate system needs to be sampled to detect changes forced by anthropogenic activities. Hence, to accelerate the numerical simulations, using eScience efficient computing expertise is highly relevant. Scientific publications on regional sea level rise projections were published,

and the work on high resolution global coupled climate modeling is now followed up in the H2020 funded PRIMAVERA project. Also, the GPU expertise developed here, including a kernel tuner to find optimal set ups in a large parameter space, is reused in many other projects where accelerated computing is addressed, such as radio astronomy (e. g. [10]).

The FAIR data principles (FAIR = findable accessible interoperable re-usable) were initiated with eScience at the heart. An eScience Center sponsored workshop in 2014 resulted in a white paper focusing on data stewardship. Later, a paper appeared in Nature Scientific Data on the FAIR principles with eScience co-authorship [14]. We applied the FAIR approach in the *CANDYGENE* project funded by the eScience Center. Integrating different software tools, we set up a FAIR data port for biological data [11], from genotype to phenotype of tomatoes. Currently, we are implementing aspects of FAIR in many other domains, such as in the eScience funded *AAAlert* project for radio astronomical data and in the *MAGIC* project, funded by Copernicus Climate Change Services (C3S), for climate data.

In the *Embodied Emotions* project, the eScience Center collaborated with VU University in Amsterdam and demonstrated that a multi-label text classification approach to learning complex emotion models on historical text is feasible. In particular, we analyzed theater texts with respect to the development of emotions used in theater plays. The performance of the Historic Embodied Emotion Model (HEEM) developed by the project team is similar to the performance for simpler emotion models. Comparing HEEM to LIWC, an existing dictionary-based sentiment analysis tool, HEEM yields finer-grained results. The project's novel approach in sentiment analysis has the potential to have a transformative impact on the humanities domain. The substantially more detailed and fine-grained analysis method enables researchers working in sentiment mining to apply the method to their own research questions and possibly generate new and more detailed theoretical insights. The analyses over time of the use of emotions were not possible before [6].

In the *eAstronomy* project, we accelerated the pipeline of signal processing in radio astronomy. We developed novel GPU algorithms for the real-time signal processing system used in LOFAR, the largest radio telescope in the world. In collaboration with

ASTRON, we also created new methods to remove interference from the signal in real time. The IBM Blue Gene supercomputer that was initially used could be replaced with a small GPU cluster, reducing operational cost and power usage, and making the instrument more sensitive. In addition, we developed a pulsar searching pipeline on GPUs, which is now used in production for Apertif in the Westerbork telescope. This pipeline is an order of magnitude faster than all earlier pipelines, allowing astronomers to survey a much larger part of the sky [9].

### Coordinating activities

As part of the national e-infrastructure, the eScience Center is also engaged in coordinating and training activities. The eScience Center has set up and leads ePLAN (<https://escience-platform.nl/>), the national platform for eScience and data research centers. This brings together institutions engaged in data research and eScience to share insights and have a single voice in the policy arena. Example activities include a survey among Dutch researchers on e-Infrastructure needs<sup>3</sup>, a survey on needs for the European Science Clouds, and workshops on, for instance, software sustainability, and the FAIR principles in different domains.

A similar activity has been initiated at the European scale. PLAN-E is the platform of European eScience Centers (<https://plan-europe.eu/>). Its members form a vital link between domain researchers and e-infrastructure providers in particular. Members from over 20 countries are represented. PLAN-E is a stakeholder to the European Committee and, for instance, advises the Committee on the development of the European Open Science Cloud. It is crucial that its view on current and future e-infrastructures is based on the perspectives from multiple scientific disciplines. Members meet at least once per year to discuss ongoing matters and to align their activities.

The eScience Center is strongly involved in the national science programming arena, in particular when it is related to digitization of research. It co-leads the National Science Agenda activities on big data and jointly works with SURF on implementing the FAIR principles in scientific research.

Training the next generation scientists on data science and data stewardship is done by university

<sup>3</sup> [https://www.esciencecenter.nl/pdf/Infrastructuur\\_duurzaam\\_op\\_maat\\_versie\\_1.0B\\_DEF\\_03-03-2016.pdf](https://www.esciencecenter.nl/pdf/Infrastructuur_duurzaam_op_maat_versie_1.0B_DEF_03-03-2016.pdf).

partners. The eScience Center contributes by offering Data and Software Carpentry workshops where essential data and programming skills are trained. Furthermore, we provide tutorials of research software packages that we develop and maintain at the eScience Center.

Beyond Europe, the eScience Center is visible at domain and computer science conferences, as well as the international eScience Conference. With IEEE, we are co-organizing the 14<sup>th</sup> International Conference on eScience in Amsterdam, where both eScience applications in specific scientific fields and advancements in computer and data science will be addressed.

## Summary and conclusions

The Netherlands eScience Center is the Dutch national expertise center for the development and application of research software to pioneer new scientific horizons. It has been successful in collaborating with domain scientists since its inception in 2011, building upon earlier activities that were more programmatic of nature. Crucial is that it takes the application-domain perspective by translating research needs into software that takes advantage of available technology and data from all disciplines, computer and data science, and providers of hardware infrastructure. An important part of the success is the close collaboration of eScience Research Engineers, employed by the eScience Center, with researchers at Dutch academic institutions. The eScience Research Engineers have a skill set that will be increasingly required to achieve scientific advancements in all academic fields that depend on digital methods. Combining the physical central location where staff work and learn, together with distributed local environments where engineers collaborate with academic partners, has contributed to the success of the eScience Center. The demand for eScience projects substantially exceeds our financial capacity and has led to the advice to the national research council to augment the role of the center. The eScience Center offers career perspectives to eScience Research Engineers within and outside the center. Given the increased use of digital tools and the increased multidisciplinary character of scientific practice, such a career development is essential for the future development of scientific research. In this context, software citation is equally important for scientific

reproducibility, and for giving credit to software as research output.

eScience crucially contributes to open science through its application of open science standards (open access and open source of data, software, and publications), but also through the way it works. Re-using software, workflows, and data is at the heart of the center. The tools and applications developed are generalized and made available for reuse. The reuse and the sustainability of research software is a major challenge, though. While data management, data stewardship, and related data reuse are high on research policy agendas and accepted by many researchers, the sustainability and reuse of software and workflows is still underdeveloped. Software sustainability, although supported by the eScience Center, does not have a clear enough place in the national and international e-infrastructure, while it is very much needed for open and reproducible science.

The eScience Center has been very successful in enhancing and accelerating research in a wide variety of scientific disciplines, based on the steep rise of scientific publications to which eScientists contributed. However, it remains a challenge to have a much wider impact beyond individual scientists. Setting up multi-year large projects with one research group that is well connected to other researchers may partially remedy this. A more coordinated international effort is needed as well, though.

The link with the e-infrastructure is crucial. While the national e-infrastructure is used extensively, which can be considered to be a success, the feedback from the research to the e-infrastructure development itself needs attention. By taking the domain perspective and making work problem driven, eScience can provide feedback about e-infrastructure development. Coordinated activities, such as ePLAN and PLAN-E, do so, but this needs to be strengthened. For instance, the development towards data-intensive science and the increased use of AI needs to be rapidly responded to by the national e-infrastructure developments.

Further opportunities beyond academic research exist. Although the focus is on scientific research, collaboration with the private sector may further increase the impact of eScience. While the eScience Center has been engaged in some activities, this is still in development.

In summary, the Dutch eScience activities, culminating in the activities of The Netherlands

eScience Center, have successfully contributed to digitally enhanced scientific research. It is unique as a national institution and expertise center, employing its own highly skilled staff that collaborates with academia. With research software at its heart, it has developed as a crucial part of the Dutch national research infrastructure.

**Open Access.** This article is distributed under the terms of the Creative Commons Attribution Non-commercial License (<http://creativecommons.org/licenses/by/4.0/deed.de>) which permits any non-commercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

- Allen A, Aragon C, Becker C, Carver J, Chis A, Combemale B, Croucher M, Crowston K, Garjo D, Gehani A, Goble C, Haines R, Hirschfeld R, Howison J, Huff K, Jay C, Katz DS, Kirchner C, Kuksenok K, Lämmel R, Nierstrasz O, Turk M, van Nieuwpoort RV, Vaughn M, Vinju JJ (2017) Engineering Academic Software (Dagstuhl Perspectives Workshop 16252). Dagstuhl Manifestos 6. <https://doi.org/10.4230/DagMan.6.1.1>
- Bauer P, Thorpe A, Brunet G (2015) The quiet revolution of numerical weather prediction. *Nature* 525:47–55, <https://doi.org/10.1038/nature14956>
- Bird I (2011) Computing for the Large Hadron Collider. *Annu Rev Nucl Part Sci* 61:99–118, <https://doi.org/10.1146/annurev-nucl-102010-130059>
- Brunnabend SE, Dijkstra HA, Kliphuis MA, Bal HE, Seinstra FJ, van Werkhoven B, Maassen J, van Meersbergen M (2017) Changes in extreme regional sea level under global warming. *Ocean Sci* 13:47–60, <https://doi.org/10.5194/os-13-47-2017>
- Hey T, Tansley S, Tolle K (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, WA, ISBN 978-0-9825442-0-4
- Leemans I, van der Zwaan JM, Maks I, Kuijpers E, Steenbergh K (2017) Mining embodied emotions: a comparative analysis of sentiment and emotion in Dutch texts, 1600–1800. In: *Digital Humanities Quarterly* 11 <http://www.digitalhumanities.org/dhq/vol/11/4/000343/000343.html>
- Lusher SJ, McGuire R, van Schaik RC, Nicholson CD, Vlieg J de (2014) Data-driven medicinal chemistry in the era of big data. *Drug Discov Today* 19:859–868, <https://doi.org/10.1016/j.drudis.2013.12.004>
- Ronda RJ, Steeneveld GJ, Heusinkveld BG, Attema JJ, Holtslag AAM (2017) Urban finescale forecasting reveals weather conditions with unprecedented detail. *Bull Amer Meteor Soc* 98:2675–2688, <https://doi.org/10.1175/BAMS-D-16-0297.1>
- Sclocco A, Bal HE, van Nieuwpoort RV (2015) Finding Pulsars in Real-Time. In: *2015 IEEE 11th International Conference on e-Science*, pp 98–107, <https://doi.org/10.1109/eScience.2015.11>
- Sclocco A, van Leeuwen J, Bal HE, van Nieuwpoort RV (2016) Real-time dedispersion for fast radio transient surveys, using auto tuning on many-core accelerators. *Astron Comput* 14:1–7, <https://doi.org/10.1016/j.ascom.2016.01.001>
- da Silva Santos LB, Wilkinson MD, Kuzniar A, Kaliyaperumal R, Thompson M, Dumontier M, Burger K (2016) FAIR data points supporting big data interoperability. In: Zelm M, Doumeings G, Mendonca JP (eds) *Enterprise Interoperability in the Digitized and Networked Factory of the Future*. ISTE, London, pp 270–279, ISBN 978-1-8470404-4-2
- Warwick C, Nyhan J, Terras MM (eds) (2012) *Digital Humanities in Practice*. Facet, London, ISBN 978-1-8560476-6-1
- van Werkhoven B, Maassen J, Kliphuis MA, Dijkstra HA, Brunnabend S-E, van Meersbergen M, Seinstra FJ, Bal HE (2014) A distributed computing approach to improve the performance of the Parallel Ocean Program (v2.1). *Geosci Model Dev* 7:267–281, <https://doi.org/10.5194/gmd-7-267-2014>
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC 't, Hoofst R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik RC, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:16–18, <https://doi.org/10.1038/sdata.2016.18>

# *To achieve the goals of e-Science, we must change research culture globally*

Neil Chue Hong

## Introduction

### An ambitious plan

In the early 2000s the UK's research funders embarked on an ambitious initiative that brought together domain scientists, computer scientists, and software engineers in collaboration to solve grand challenges of research. As coined by John Taylor, at that time the Director General of the UK's Office of Science and Technology, this £120m "e-Science" program invested in large scale, collaborative projects to develop infrastructure, software and resources to support researchers working in all areas of research [18].

This approach was picked up by other countries, including the United States of America through the National Science Foundation's Office of Cyber-Infrastructure [17] and The Netherlands through The Netherlands e-Science Center<sup>1</sup>. Although e-Science is funded differently across the world, it represents a paradigm shift, identifying the need for infrastructure and tools to support data-intensive research workflows, supporting the case for open data and open science to make research more efficient, and supporting research integrity as experiments become more complex.

### Software as a shaky foundation

Underpinning this is software: if e-Science is defined as the use of computational tools and platforms to enable novel research, almost all of modern research is e-Science. This is evidenced by surveys conducted of UK researchers at research-intensive

universities in 2014 [11] (Fig. 1) and US post-doctoral researchers in 2017 [16]. These found that the majority of respondents (69 % in UK, 63 % in US) would find it impossible to conduct their research without software.

A key challenge identified in these surveys is that a large proportion of researchers (54 % in US; 45 % in UK, including 21 % of those who develop their own research software) had received no training in software development, despite researchers often having to extend or modify the software they are using. This was backed by a 2014 study [3] by the UK Biotechnology and Biological Sciences Research Council and Medical Research Council on vulnerable skills, which noted that "informatics skills are applicable to many areas of both the biosciences and the medical sciences" but "data analytics, especially bioinformatics, appear to be particularly vulnerable." Without the correct skills and experience, the threat is that research may be reaching a point where researchers are conducting experiments without truly knowing what is happening.

This crisis of "reproducibility" – the ability to take the data and methodology (often in the form of software) and reproduce the results published in a paper – is of particular concern given the implications it can have in certain fields. A 2012 study by Begley and Ellis [4] reviewed a decade of landmark

---

<https://doi.org/10.1007/s00287-018-01134-1>  
© The Author(s) 2018. This article is available on SpringerLink.com with Open Access.

---

Neil Chue Hong  
Software Sustainability Institute, EPCC, University of Edinburgh,  
Edinburgh, United Kingdom  
E-Mail: N.ChueHong@software.ac.uk

---

<sup>1</sup> <https://www.esciencecenter.nl/>.

## Abstract

The e-Science program was initiated in the United Kingdom in the early 2000s with the aim of bringing together researchers in large scale, collaborative projects involving software and computation to solve grand challenges. A legacy of this program has been an understanding of the importance of the people behind the software, the researchers and research software engineers, as well as the challenges of developing and maintaining code that is reusable given the problems of software decay.

The Software Sustainability Institute was established in the UK to provide support and direction for the research software community through consultancy, training, engagement, and policy campaigns. Through this it has worked with an international community of collaborators, in the UK, in Europe, and across the world to support reusability, research integrity, and transparency, recognizing that to achieve the goals of e-Science, we must change research culture globally.

cancer research papers and found that 47 out of 53 were irreproducible, often for simple-to-correct reasons such as failure to repeat experiments, failure to show all data, and inappropriate use of statistical tests. Each of these issues should benefit from the increased use of computational techniques, and yet similar failures can be seen happening in disciplines such as genetics [13] and computer science [6], where even getting hold of the code used – let alone verifying its correctness – was an issue.

### A culture of challenges

The question is why this culture exists. Is it because of the pressures of “publish or perish” rewarding speed and quantity over rigor and quality? Certainly, this plays a part, as discussions at the Software Sustainability Institute’s Collaborations Workshops<sup>2</sup> have noted. Because much code is written by students and early career postdocs, the emphasis is getting the software only to the point where it can be used to produce results for a paper. Additionally, because academic reviewers (both of research papers

and research grants) favor novelty, there is no motivation to go back and make software more reusable or robust.

A big issue is that whilst it is believed that there are many errors introduced by software in the literature, there is no incentive to fix them. The chances of an author being found out diminish, if their code is not published, and for those who do notice an error and try to correct it, the dreaded “retraction” is often the only way to do it [10, 15]. There are discussions amongst journals and publishers to change this [9], and it will be necessary to change the culture around scientific integrity and correctness before software quality improves across all codes.

Finally, software is hard to work with. Forget making research reproducible, it is hard enough to use it 6 months later. Konrad Hinsien has coined the term “software collapse” [12] due to the fact that software stops working eventually if is not actively maintained. Hinsien notes that software stacks used in computational science have a nearly universal multi-layer structure:

- Project-specific software: whatever it takes to do a computation using software building blocks from the lower three levels; scripts, workflows, computational notebooks, small special-purpose libraries, and utilities.
- Discipline-specific research software: tools and libraries that implement models and methods which are developed and used by research communities.
- Scientific infrastructure: libraries, and utilities used for research in many different disciplines, such as LAPACK, NumPy, or Gnuplot.
- Non-scientific infrastructure: operating systems, compilers, and support code for I/O, user interfaces, etc.

where software in each layer builds on and depends on software in all layers below it, and any changes in any lower layer can cause it to collapse.

The reproducible research community has traditionally focused on the project-specific software layer, where the main obstacle (as described earlier) is the unavailability of the code. This is also the software that receives the least attention after a project ends – it may become the starting point for software specific to another project – but it is rarely used by anyone outside the project that developed it. Again, it is down to incentive for improvement, for reusing

<sup>2</sup> <https://www.software.ac.uk/workshops>.

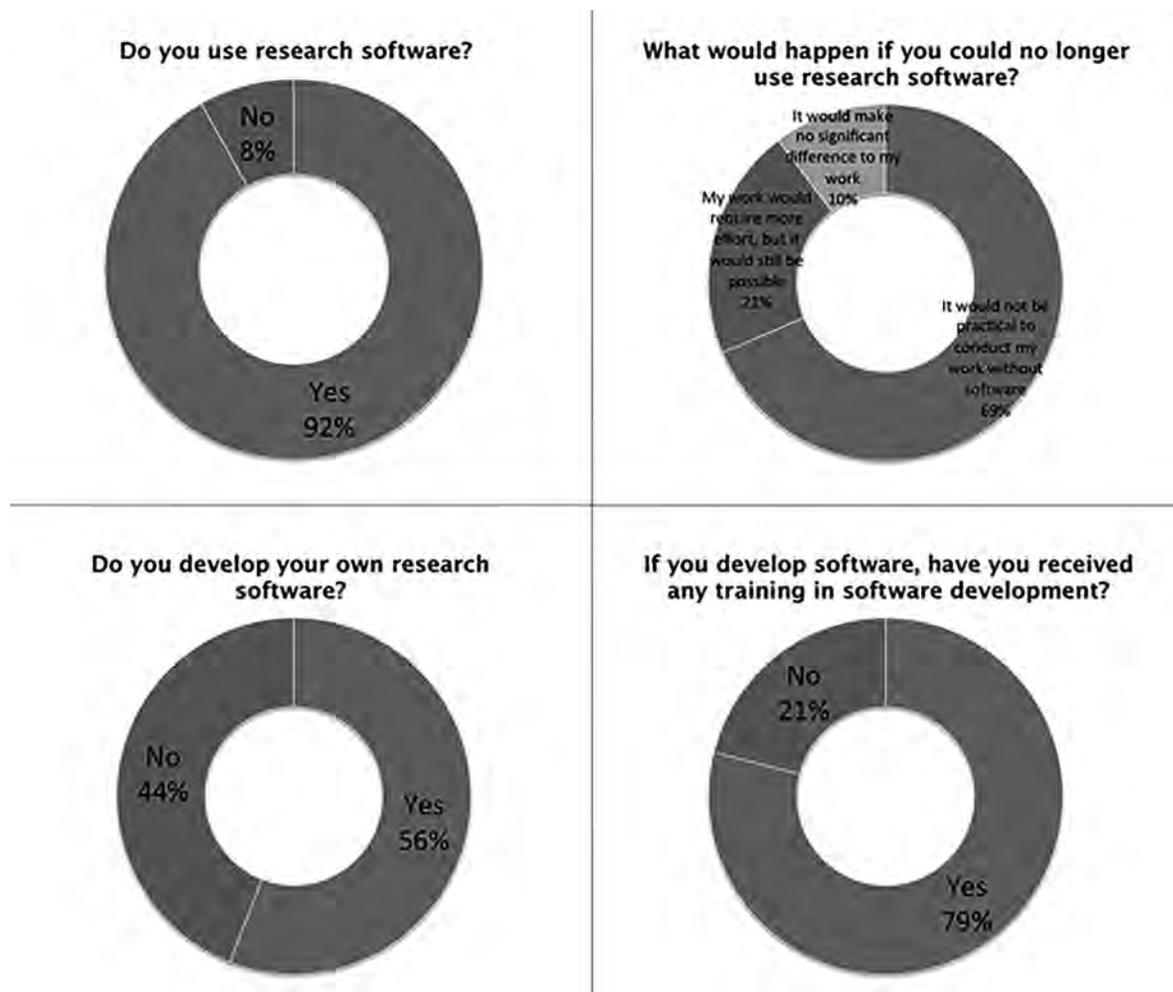


Fig. 1 UK Research Software Survey Results (2014). doi: 10.5281/zenodo.14809

others' work, for contributing back to the "communal maintenance". Whilst some software achieves this sustainability, making its way perhaps to Hinsen's lower levels of discipline-specific research software or scientific infrastructure, the majority is discarded because of the effort to keep it running. This is not necessarily a problem – much of this software is used to solve a single, specific problem – but it is an issue if results are to be revisited later, when even if the code has been made available, it cannot be easily run because of changes in the underlying libraries, operating systems, and other dependencies. Although virtual machines and, more recently, containers have been posited as solutions for this problem of decay, in reality, they suffer the same issues. C. Titus Brown, a regular commentator in this area, concludes [5] that "on a decadal time scale, we

cannot rely on software to run repeatedly." Daniel S. Katz, who was previously NSF's Program Director for their Software Infrastructure for Sustained Innovation program, asks "Is software reproducibility possible and practical?" [14]. This pushback against the vision of perfect reproducibility is perhaps the most important debate in the research software community at present, because it forces us to ask the question about the trade-off between the costs of making research results reproducible, the usefulness of doing so for a particular research output, and the potential new research that could be accomplished with those resources.

In the end, perhaps the pragmatic approach is to make it easier (not easy) to inspect research. After all, the consequences of mistakes in software can be large. In the case of economists Carmen Reinhart

and Kenneth Rogoff, their research paper “Growth in a Time of Debt” [19] was used as evidence to support the implementation of austerity measures by several governments, including in the UK. However, a software error in an Excel spreadsheet – identified by a student [7] – meant that the results were not as conclusive as they first appeared. However, this discovery could not have happened if Reinhart and Rogoff had not been willing to provide their data and code to the student to analyze. Perhaps open science is the way to go for practical reproducibility?

## The work of the Software Sustainability Institute

### Building a cross-disciplinary community

The Software Sustainability Institute was set up [8] in 2010 as a partnership between the universities of Edinburgh, Manchester, Oxford, and Southampton. Its goal is to cultivate world-class research with software, by overcoming the problems that beset research software and changing the way that researchers view it; a way to address the challenges identified in the previous section.

Over the last 8 years, it has evolved (with funding from the Engineering and Physical Sciences Research Council, Economic and Social Sciences Research Council, and Biotechnology and Biological Sciences Research Council) to bring the community together in tackling these problems. It works on several scales, providing consultancy and advice direct to researchers who are developing software, enabling researchers across the UK to access appropriate training, and running events and campaigns to support and facilitate champions of research culture change through best practice [23] and policy [1].

This work is inherently cross-disciplinary. The Institute works with researchers from every discipline. Although individual requirements are different, many of the issues faced are the same, including software reuse, reproducibility, and accumulation of software technical debt. To understand how to address these and other issues, an understanding of them must first be built from within the research community.

The Institute’s Fellowship Program [20] is a cost-effective approach to obtain community intelligence, recruiting members of the software research community in a way that benefits them and the Institute. The program provides bursaries to researchers

across many domains, institutions, and career stages in exchange for their expertise and advice, although it focuses on early career researchers. Institute Fellows receive funding for attending conferences and workshops that focus on different aspects of software development and use in their research area. In exchange, they supply the domain intelligence used to develop wider policy objectives. As domain specialists, they are also ideally placed to identify other opportunities (such as consultancy projects) and they also provide feedback on activities undertaken by the Institute. Over 5 years, the Institute has created a network of more than 100 Fellows, spanning subjects including history, engineering, imaging and earth sciences, to name a few.

Given the limited budget, the Fellowship Program represents a significant return on investment, with the Fellows having established a keen group identity, vision, and set of activities that assist the UK research community in many areas. A remarkable outcome of the program is that many Fellows go beyond their initial remit, taking initiative in organizing events within their respective communities, delivering presentations, advocating best practices, and suggesting other activities that have relevance and benefit to the Institute. A major event organized and attended by four Fellows was a “Software & Research Town Hall Meeting” at the American Geophysical Union (AGU) Fall Meeting; this is one of the AGU’s largest community events, attended by 22,000 researchers from all over the world. Fellows have also led events that have focused on addressing the software development skills gap and discussing cultural issues. For example, in coordination with the training team, our Fellows have been instrumental in the growth of Software Carpentry [22] and Data Carpentry<sup>3</sup> in the UK, with two of them becoming fully qualified instructor trainers and having been involved in the creation of new training as part of Library Carpentry<sup>4</sup>. As the program has progressed, our Fellows have also progressed to prominent positions, including Head of Research Services at the British Library, Director of Research Engineering at the Alan Turing Institute, and Vice President (Promotion of Computing) of the British Computer Society, all helping to influence policy and promote the objectives of the Institute.

<sup>3</sup> <http://www.datacarpentry.org/>.

<sup>4</sup> <https://librarycarpentry.github.io/>.

## Recognizing the right people

The Institute also aims to bring together the right people to understand and address topical issues. From the flagship Collaborations Workshop un-conference event [21] to workshops on themes such as software credit, measuring the impact of workshops, and research data visualization, our aim is to ensure that the widest possible input is sought for key software challenges facing research. This has particularly been the case with two topics: recognition of research software engineers and computational skills for researchers.

In 2012, following a discussion at the Collaborations Workshop 2012, a number of attendees published a paper (including this author) on “The research software engineer” [2] which noted that:

*“Computational work must reflect the committed attitude of experimentalists towards caring about precise, professional, repeatable, meticulous work – no-one with the same casual attitude to experimental instrumentation as many researchers have to code would be allowed anywhere near a lab.”*

This returned to the themes of e-Science, and in particular to the idea that research should be carried out by teams of varied specialists, including individuals who “combine a professional attitude to the exercise of software engineering with a deep understanding of research topics” – *research software engineers*. The paper made three recommendations to address the challenges identified:

- “The REF [Research Evaluation Framework for assessing the quality of research in UK institutions] allows for recognition of software deliverables through its system for impact measurement. In practice, however, this depends on the decisions of individual panels. REF panels should be given clear guidance on the importance of weighing software as a ‘first class’ research output.
- Research software engineers are a new role in academic institutions. Institutions and funding panels should recognize the value of this role in funding research proposals and in providing career progression and succession for such individuals.
- All researchers must be exposed to best practices in software development. The fundamentals of good software engineering should form part of every researcher’s basic training.”

The first is still in progress and awaits the imminent publication of guidance for the 2021 REF assess-

ment process. The third has been achieved through the success of the international carpentries open training initiative, for whom the Institute is the UK coordinator. Since the first event in 2012, over 120 workshops have been run in the UK, teaching thousands of researchers the basic software development (through software carpentry) and data management and analysis (through data carpentry) skills they require to conduct research professionally. More than this, there are now over 100 trained instructors spread across the UK (Fig. 2), ready to pass on their knowledge to new researchers and students.

Finally, the second point, on recognition for research software engineers, has become one of the success stories of the last years. The initial paper sparked a debate into the role, quickly followed by the setting up of the UK Research Software Engineer Association<sup>5</sup> and recognition by the Engineering and Physical Science Research Council through the funding of RSE Fellowships – putting the position on par with academics. Since then, many Research Software Engineering groups have been set up at universities, providing services and support to researchers developing code, and changing the business of research software. These groups are actively involved in funding proposals, ensuring that the effort required to help address Hinsen’s software decay is properly understood and costed. 5 years since the first discussions, and a second round of EPSRC RSE Fellowships have been awarded, job adverts for RSEs are posted every week; the RSE Association is about to incorporate into a legal professional body, and chapters have opened in The Netherlands, Canada, South Africa, and, of course, Germany<sup>6</sup>.

## e-Science is international

It seems obvious, but the main finding that the Institute has learned from working with researchers in the UK is that research is global. Not only are research teams at UK universities drawn from all nations, but research itself is often conducted in teams that span countries. The most obvious of these are those associated with high energy physics, such as the experiments on the Large Hadron Collider, but international collaboration, even when it is just between two people, is very much the norm. The Institute itself has many international collaborators;

<sup>5</sup> <https://rse.ac.uk/>.

<sup>6</sup> <http://www.de-rse.org/>.



**Fig. 2** Map of UK carpentry instructors affiliations (1/2018)

in addition to the RSE groups and Carpentries, the Institute works with groups like the Center for Trustworthy Scientific Cyberinfrastructure<sup>7</sup> in the US to tap into their expertise on security and resilience, Software Heritage<sup>8</sup> in France for software preservation, the Future of Research Communications and e-Scholarship<sup>9</sup> and Research Data Alliance<sup>10</sup> on software citation and credit, and the International Coalition on Science Gateways<sup>11</sup>. Working to understand the similarities and differences between how research software is treated in different countries also helps us to work together to change the research culture, for the benefit of all.

This is the ultimate goal of e-Science, as first defined nearly 20 years ago: to bring together teams from across disciplines, roles, and countries in collaboration to solve research's grand challenges. It is also the challenge of e-Science, which was perhaps misunderstood in those early years of the UK

e-Science program, that to be successful it must concentrate on the people and knowledge transfer, more than on the hardware and software. The practice of research has changed substantially, it is our role as users and developers of research software to ensure that we continue to strive for research integrity and transparency over blind reproducibility and recognize the essential role of those who help make software reusable.

**Open Access.** This article is distributed under the terms of the Creative Commons Attribution Non-commercial License (<http://creativecommons.org/licenses/by/4.0/deed.de>) which permits any non-commercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

1. Aleksic J, Alexa A, Attwood TK, Chue Hong N, Dahlö M, Davey R (2015) An open science peer review oath. F1000Research. <https://doi.org/10.12688/f1000research.5686.2>
2. Baxter R et al (2012) The research software engineer. In: Proceedings of the Digital Research Conference, Oxford

<sup>7</sup> <https://trustedci.org/>.

<sup>8</sup> <https://www.softwareheritage.org/>.

<sup>9</sup> <https://www.force11.org/>.

<sup>10</sup> <https://www.rd-alliance.org/>.

<sup>11</sup> <http://www.icsciencegateways.org/>.

3. BBSRC and MRC (2014) Review of vulnerable skills and capabilities. <https://www.mrc.ac.uk/documents/pdf/review-of-vulnerable-skills-and-capabilities/>, last access: 1.2.2018
4. Begley CG, Ellis LM (2012) Drug development: Raise standards for preclinical cancer research. *Nature* 483(7391):531–533, <https://doi.org/10.1038/483531a>, PMID, 22460880
5. Brown CT (2017) How I learned to stop worrying and love the coming archivability crisis in scientific software. <http://ivory.idyll.org/blog/2017-pof-software-archivability.html>, last access: 1.2.2018
6. Collberg C, Proebsting TA (2016) Repeatability in computer systems research. *Commun ACM* 59(3):62–69, <https://doi.org/10.1145/2812803>
7. Coy P (2013) FAQ: Reinhart, Rogoff, and the Excel Error that changed history. Bloomberg. <https://www.bloomberg.com/news/articles/2013-04-18/faq-reinhart-rogooff-and-the-excel-error-that-changed-history>, last access: 1.2.2018
8. Crouch S, Chue Hong N, Hettrick S, Jackson M, Pawlik A, Sufi S, Parsons M (2013) The software sustainability institute: changing research software attitudes and practices. *Comput Sci Eng* 15(6):74–80, <https://doi.org/10.1109/mcse.2013.133>, last access: 1.2.2018
9. Enserink M (2017) Rethinking the dreaded r-word. *Science* 356(6342):998, <https://doi.org/10.1126/science.356.6342.998>
10. Gallego Llorente M, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, Stock JT, Coltorti M, Pieruccini P, Stretton S, Brock F, Higham T, Park Y, Hofreiter M, Bradley DG, Bhak J, Pinhasi R, Manica A (2016) Erratum for the Report “Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa” (previously titled “Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent”). *Science* 351(6275):aaf3945–aaf3945, <https://doi.org/10.1126/science.aaf3945>
11. Hettrick SJ et al (2014) UK Research Software Survey 2014, <https://dx.doi.org/10.5281/zenodo.14809>
12. Hinzen K (2017) Sustainable software and reproducible research: dealing with software collapse, <http://blog.khinzen.net/posts/2017/01/13/sustainable-software-and-reproducible-research-dealing-with-software-collapse/>, last access: 1.2.2018
13. Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, van Noort V et al (2008) Repeatability of published microarray gene expression analyses. *Nat Genet* 41(2):149–155, <https://doi.org/10.1038/ng.295>
14. Katz DS (2017) Is software reproducibility possible and practical? <https://danielskatzblog.wordpress.com/2017/02/07/is-software-reproducibility-possible-and-practical/>, last access: 1.2.2018
15. Mann RP, Perna A, Strömbom D, Garnett R, Herbert-Read JE, Sumpter DJT, Ward AJW (2012) Retraction: Multi-scale inference of interaction rules in animal groups using bayesian model selection. *PLoS Comput Biol* 8(8), <https://doi.org/10.1371/annotation/7bc3a37e-db82-4813-8242-7d34877125c5>
16. Nangia U and Katz DS (2017) Track 1 Paper: Surveying the US National Postdoctoral Association regarding software use and training in research, <https://doi.org/10.6084/m9.figshare.5328442.v3>
17. National Academies of Sciences, Engineering, and Medicine (2016) Future Directions for NSF Advanced Computing Infrastructure to Support US Science and Engineering in 2017–2020. The National Academies Press, Washington, DC, <https://doi.org/10.17226/21886>
18. RCUK Review of e-Science (2009) Building a UK Foundation for the Transformative Enhancement of Research and Innovation. <https://www.epsrc.ac.uk/newsevents/pubs/rcuk-review-of-e-science-2009-building-a-uk-foundation-for-the-transformative-enhancement-of-research-and-innovation/>, last access: 1.2.2018
19. Reinhart C, Rogoff K (2010) Growth in a time of debt. National Bureau of Economic Research, <https://doi.org/10.3386/w15639>
20. Sufi S (2016) The Software Sustainability Institute Fellowship Programme: Supporting the Social Side of Research. In: Proceedings of the 4th Workshop on Sustainable Scientific Software: Practice and Experience. CEUR-WS, [http://ceur-ws.org/Vol-1686/WSSSP4\\_paper\\_27.pdf](http://ceur-ws.org/Vol-1686/WSSSP4_paper_27.pdf), last access: 1.2.2018
21. Sufi S, Peru G, Robinson J, Carr L, De Roure D, Goble C, Pawlik A (2014) Software in Reproducible Research. In: Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering – TRUST '14. ACM Press, <https://doi.org/10.1145/2618137.2618140>
22. Wilson G (2016) Software Carpentry: Lessons Learned. F1000Research, <https://doi.org/10.12688/f1000research.3-62.v2>
23. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Wilson P (2014) Best practices for scientific computing. *PLoS Biol* 12(1):e1001745, <https://doi.org/10.1371/journal.pbio.1001745>

# DFG-Förderung und der digitale Wandel in den Wissenschaften

Ein Wegweiser zu Fördermöglichkeiten und Leitlinien

Matthias Katerbow · Michael Royeck  
Andreas Raabe

## Einführung

Die Deutsche Forschungsgemeinschaft (DFG) ist die größte Förderorganisation und zentrale Selbstverwaltungsorganisation der Wissenschaft in Deutschland; sie dient der Wissenschaft in allen ihren Zweigen. Die DFG ist als gemeinnütziger Verein organisiert, ihre Mitglieder sind Hochschulen, Forschungseinrichtungen, Akademien der Wissenschaft und weitere wissenschaftliche Verbände. Die Kernaufgabe der DFG besteht in der wettbewerblichen Auswahl der besten Forschungsvorhaben von Wissenschaftlerinnen und Wissenschaftlern an Hochschulen und Forschungsinstituten und in deren Finanzierung. Hierfür verfügt die DFG über ein jährliches Budget von ca. 3,2 Mrd. €, mit dem über 30.000 Projekte gefördert werden. Zu den Voraussetzungen für exzellente Wissenschaft zählt auch die wissenschaftliche Infrastruktur, sodass die DFG die Erprobung, den Aufbau und die Konsolidierung wissenschaftlicher Literaturversorgungs- und Informationssysteme ebenso fördert wie wissenschaftliche Geräte und Informationstechnik. Der offensichtliche Zusammenhang zwischen Forschung und Forschungsinfrastruktur spiegelt sich im Förderhandeln der DFG in der Forschungsförderung und der Infrastrukturförderung wider.

Das wissenschaftliche Arbeiten in allen Disziplinen ist durch stetige Veränderung gekennzeichnet, alleine schon aufgrund der wissenschaftlichen Neugier von Forscherinnen und Forschern. Diese Veränderungen sind – in den einzelnen Fächern durchaus unterschiedlich intensiv und schnell – seit mehreren Jahren vor allem durch den Digital Turn geprägt. Dies wird in der täglichen

Forschungsarbeit ebenso deutlich (z. B. Analyse und Weiterverarbeitung von Forschungsdaten, Nutzung von Forschungssoftware und neuen Publikationsformen) wie in der Informationsversorgung (Zugriff, Suche, Rezeption, Weitergabe und Austausch von wissenschaftlichen Informationen). Die Veränderungen des wissenschaftlichen Arbeitens betreffen den gesamten Forschungszyklus und in jeder Phase des wissenschaftlichen Arbeitens zeichnen sich umfassende Bedarfe und Herausforderungen an wissenschaftliche Infrastrukturen ab.

Die durch den Digital Turn hervorgerufenen Veränderungen sind derart umfassend und andauernd, dass sich sowohl die Vorstellungen und Sachverhalte als auch die sprachlichen Bezeichnungen für diese noch nicht gefestigt haben. Es lässt sich daher in vielen Debatten und Veröffentlichungen sehr anschaulich beobachten, wie sprachliche Aushandlungsprozesse und Deutungshoheiten ebenso allgegenwärtig sind wie variierende sprachliche Bezeichnungen (z. B. Digitalisierung, Digitalität, Digital Science, Science 2.0, E-Science, Open Science, Digitale Wissenschaft, usw.). Grundsätzlich umfasst der *Digitale Wandel in den Wissenschaften* alle Änderungen und Auswirkungen, die in den Wissenschaften mit der Nutzung und dynamischen Weiterentwicklung digitaler Technologien zu tun haben. Dazu gehören sowohl Transformationsprozesse

<https://doi.org/10.1007/s00287-018-01135-0>  
© Springer-Verlag Berlin Heidelberg 2018

Matthias Katerbow · Michael Royeck · Andreas Raabe  
DFG, Bonn  
E-Mails: {matthias.katerbow, michael.royeck,  
andreas.raabe}@dfg.de

## Zusammenfassung

Der digitale Wandel prägt die verschiedenen Bereiche der Gesellschaft in hohem Maß und auf vielfältige Weise. Sofern dadurch auch die Wissenschaften betroffen sind, sieht sich die Deutsche Forschungsgemeinschaft (DFG) in der Verantwortung, diesen Wandel wissenschaftsgeleitet mitzugestalten. Und dies insbesondere in den drei zentralen Handlungsfeldern der DFG: Selbstverwaltung der Wissenschaft, Förderung von Forschung und wissenschaftlicher Infrastruktur sowie Politikberatung. Der vorliegende Beitrag dient als Wegweiser durch das aktuelle Förderportfolio der DFG, um so für unterschiedliche Vorhaben die passenden Fördermöglichkeiten aufzuzeigen. Darüber hinaus bietet der Beitrag eine Übersicht bisheriger Aktivitäten, Leitlinien und Stellungnahmen der DFG zum Thema digitaler Wandel und IT in den Wissenschaften. Die fachübergreifenden Stellungnahmen basieren auf den Erfahrungen einer Vielzahl von Förderentscheidungen und zeigen notwendige Rahmenbedingungen sowie Strategien für erfolgreiches wissenschaftliches Arbeiten im digitalen Zeitalter auf.

als auch alle neuen Voraussetzungen und Möglichkeiten, die durch digitale Technologien überhaupt erst geschaffen werden.

Eine besondere Eigenschaft des digitalen Wandels ist seine große Dynamik und (annähernde) Ubiquität. Die durch die genuin digitale Arbeitsweise entstandenen Herausforderungen betreffen die Wissenschaft voll umfänglich – wissenschaftspraktisch, fachlich, epistemisch, organisatorisch, institutionell, technisch, infrastrukturell, finanziell, politisch, rechtlich, ethisch und kulturell. Vor diesem Hintergrund hat der Senat der DFG 2016 einen Prozess initiiert, der zu einer grundlegenden Positionierung der DFG zur „Wissenschaft im Digitalen Zeitalter“ führen soll. Diese Positionierung soll der DFG ermöglichen, die Interessen der Wissenschaft in allen – also nicht nur in den auf Infrastruktur bezogenen – Fragen des digitalen Wandels in den Wissenschaften noch wirkungsvoller zu vertreten.

Der digitale Wandel in den Wissenschaften war unlängst zentrales Thema der Neujahrsrede des Präsidenten der DFG, Prof. Dr. Peter Strohschneider.

In dieser hebt er hervor, dass die DFG als „Förderinstitution wie als Selbstverwaltungsorganisation der Wissenschaften in der Bundesrepublik [...] hier ihre Verantwortung [sieht]. Wir nehmen uns vor, ihr in dreierlei Hinsicht gerecht zu werden: Erstens in der Eröffnung von Foren und in der Begleitung fachspezifischer Reflexion auf den digitalen Wandel in allen Bereichen der Wissenschaft (und was dabei Fachspezifik heißt, wird sich unter den Bedingungen von Digitalität selbst wandeln). Zweitens wird unter diesen Bedingungen unser Förderhandeln, werden seine Instrumente und Verfahren weiterentwickelt werden müssen. Und drittens kommen auf die DFG neue Aufgaben bei der Beratung von Politik und Gesellschaft im Hinblick auf die Entwicklung der Wissenschaften im digitalen Zeitalter zu. Um dieser dreifachen Verantwortung gerecht werden zu können, führen wir in der Geschäftsstelle ein umfangreiches, mehrjähriges Strukturierungsprojekt durch. Das Präsidium der DFG wird überdies eine hochrangige Expertenkommission ‚Wissenschaft im digitalen Zeitalter‘ einsetzen.“

Auf der Basis dieser – aktuell zu erarbeitenden umfassenden Positionierung – wird der Senat der DFG auch über mögliche Aktualisierungen des Förderhandels der DFG entscheiden. Unabhängig von diesen künftigen Neuerungen bietet die DFG bereits aktuell sowohl in der Forschungsförderung als auch der Infrastrukturförderung zahlreiche Möglichkeiten, um im Rahmen der projektorientierten Drittmittelförderung den Bedarfen und Herausforderungen des digitalen Wandels begegnen zu können. Die Darstellung dieser Fördermöglichkeiten im folgenden Abschnitt ist als Wegweiser für die Wahl des passenden Förderangebots gedacht. Darüber hinaus erarbeitet die DFG – insbesondere anhand der Arbeit der Fachkollegien und anderer Gremien sowie mit weiteren Akteuren (z. B. Allianz der deutschen Wissenschaftsorganisationen) – Empfehlungen, Leitlinien und Stellungnahmen zu unterschiedlichen Themen und Aspekten des digitalen Wandels in den Wissenschaften; diese sind in Auswahl im Abschnitt „Stellungnahmen und Leitlinien der DFG“ dargestellt.

## Fördermöglichkeiten der DFG

Hinsichtlich der DFG-Fördermöglichkeiten des digitalen Wandels in den Wissenschaften sind zwei Ansätze zu unterscheiden: Einerseits die Forschungsförderung, deren zentrales Ziel die

Förderung herausragender wissenschaftlicher Forschung ist. Und andererseits die Infrastrukturförderung, deren Ziel die Verbesserung der Bedingungen zur Durchführung bestmöglicher Forschung ist; dazu zählen Informationsinfrastrukturen und Technologien sowie Geräte und Gerätetechnologien.

### Forschungsförderung

Die Förderung herausragender wissenschaftlicher Forschung ist der zentrale Auftrag der DFG. Im Rahmen der verschiedenen Programme der Forschungsförderung können originelle, grundlegende Forschungsvorhaben zu allen Themen und Fragestellungen beantragt werden.

Ein Vorhaben mit Bezug zum digitalen Wandel sollte dann im Rahmen der Forschungsförderung beantragt werden, wenn es eine originäre wissenschaftliche Fragestellung adressiert. Dies ist zum Beispiel der Fall, wenn digitale Innovationen originäre Treiber wissenschaftlicher Erkenntnis sind oder wenn digitale Artefakte Forschungsgegenstand eines Projektes darstellen. Auch Forschungsvorhaben zu Auswirkungen oder Rahmenbedingungen des Einsatzes digitaler Methoden oder des digitalen Wandels können im Rahmen der Forschungsförderung beantragt werden.

Über die gesamte Breite des Fächerkanons ist derzeit eine deutliche Zunahme erfolgreicher Vorhaben zu verzeichnen, die Methoden der digitalen Informationsverarbeitung in den Fokus ihrer Forschung rücken oder diese als zentrale Methode zur Erzielung von Erkenntnisgewinn einsetzen.

**Möglichkeiten im Rahmen der Einzelförderung.** In der Einzelförderung ist die Sachbeihilfe das Schweizer Taschenmesser unter den Förderinstrumenten der DFG (Merkblatt 50.01). Hier können zu jedem beliebigen Zeitpunkt Projekte jeder Forschungsrichtung beantragt werden. Qualität des beantragten Vorhabens und dessen zu erwartende Erkenntnishöhe sind die wichtigsten Begutachungskriterien. Personenbezogene Förderung ermöglicht die DFG im Rahmen von Forschungsstipendien (Merkblatt 1.04), den Programmen zur Förderung des wissenschaftlichen Nachwuchses Emmy-Noether (Merkblatt 50.01) und Heisenberg (Merkblatt 50.03) und den Reinhart Koselleck-Projekten (Merkblatt 1.18). Mit rund zwei Dritteln der eingegangenen Anträge und einem Drittel des Gesamtfördervolumens

der DFG ist die Einzelförderung das mit Abstand größte Programm der DFG (siehe DFG Jahresbericht 2016).

In den vorgenannten Formaten prägt sich der digitale Wandel insbesondere durch eine interdisziplinäre Vermischung von Wissenschaft jeder Fachrichtung mit neuen digitalen Methoden aus. Um der fachlichen Mischung jedes Antrags gerecht zu werden, erfolgt die Auswahl von Gutachterinnen und Gutachtern in enger Zusammenarbeit aller durch ein Antragsthema berührten Fachbereiche. Bei der Bewertung und bei der abschließenden Förderentscheidung arbeiten Wissenschaftlerinnen und Wissenschaftler über die Fachkollegienstrukturen und Fächerzugehörigkeiten hinweg bei den Begutachtungsprozessen Hand in Hand. Um Wissenschaftlerinnen und Wissenschaftlern bei der Antragstellung individuelle Unterstützung zu geben, beraten die Fachreferate der Geschäftsstelle gerne zu diesem Themenfeld.

**Koordinierte Verfahren.** Für die Bearbeitung interdisziplinärer Fragestellungen, wie sie gerade im Kontext des digitalen Wandels auftreten, sind häufig größere Konsortien notwendig. Sonderforschungsbereiche (SFB) (Merkblatt 50.06) und Graduiertenkollegs (GRK) (Merkblatt 50.07) sind bei vorwiegend lokalen Konsortien geeignete Förderprogramme, wenn mit dem gewählten Thema zugleich ein Schwerpunkt an der Hochschule ausgebaut werden soll. Um infrastrukturelle Notwendigkeiten abbilden zu können, gibt es in SFBs einen einschlägigen Teilprojekttypus für Informationsinfrastrukturen. Konkret dient dieses Modul der Realisierung von projektspezifischen Forschungsdaten- oder Informationsplattformen. Die Einbindung solcher Informationsinfrastrukturteilprojekte liegt mittlerweile bei knapp 20 % aller geförderten Sonderforschungsbereiche. Ein häufig im Kontext des digitalen Wandels thematisierter Themenkomplex ist die Methodenkompetenz des wissenschaftlichen Nachwuchses. Mit dem Programm Graduiertenkollegs bietet die DFG ein Programm zur strukturierten Ausbildung von Promovierenden anhand erkenntnisgeleiteter Forschung, in dem je nach gewähltem Thema auch die digitale Methodenkompetenz ein wichtiges Ausbildungselement sein kann.

Ist nicht örtliche Schwerpunktbildung, sondern die Bearbeitung einer gemeinsamen Forschungs-

aufgabe durch bestmögliche Kooperationen die Zielsetzung des Vorhabens, dann sind Forschergruppen (FOR) (Merkblatt 50.04) ein geeignetes Programm.

In allen koordinierten Programmen erfolgt die Auswahl der besten Anträge nach strenger Begutachtung durch Wissenschaftlerinnen und Wissenschaftler verschiedener Fachrichtungen; Hauptkriterium ist stets die wissenschaftliche Qualität.

**Schwerpunktprogramme.** Schwerpunktprogramme (Merkblatt 50.05) dienen dazu, die deutsche Wissenschaft in der Erkundung völlig neuer Forschungsfelder durch die Ausbildung thematischer Netzwerke frühzeitig zu unterstützen. In einem Wettbewerb werden die aktuellsten und wichtigsten von der Wissenschaft vorgelegten Themen durch den Senat identifiziert; nachfolgende Ausschreibungen ermöglichen eine offene Beteiligung aller, die an dem Themengebiet Interesse haben. In vielen Fachwissenschaften führt gerade die Adaptierung digitaler Forschungsmethoden zur Erschließung solcher „Emerging Fields“. In jüngerer Zeit haben sich daher vermehrt Themen an der Schnittfläche zwischen fachspezifischen Forschungsfragen und datenbezogenen bzw. simulationsorientierten Methoden oder anderen Forschungsthemen aus dem Umfeld des digitalen Wandels im Schwerpunktprogramm durchgesetzt. Auch ein deutlich verstärktes Interesse an Forschung, die sich mit Auswirkungen des digitalen Wandels auf alle Lebensbereiche beschäftigt, ist zu verzeichnen.

### Infrastrukturförderung

Die Förderung der DFG zielt auf die Schaffung optimaler Voraussetzungen zur Durchführung von Forschungsvorhaben und somit insgesamt auf eine Stärkung des gesamten Wissenschaftssystems ab. Da zahlreiche Forschungsvorhaben spezielle infrastrukturelle Voraussetzungen, Bedarfe und Entwicklungen berühren, fördert die DFG in mehreren Programmen dediziert wissenschaftliche Infrastrukturen und Technologien. Die Fördermöglichkeiten umfassen sowohl wissenschaftliche Informationssysteme als auch Geräte und Gerätetechnologien. Beide Aspekte sind intensiv durch den fortschreitenden digitalen Wandel betroffen.

**Wissenschaftliche Literaturversorgungs- und Informationssysteme.** Ziel dieses Förderbereichs ist der Aufbau eines abgestimmten Systems von Informationsinfrastrukturen für die Wissenschaft; es werden ein für Nutzerinnen und Nutzer freier und umfassender Zugang zu analogen und insbesondere digitalen wissenschaftlichen Informationen sowie die Vernetzung von wissenschaftlichen Informationen und Daten angestrebt. Das Förderangebot in diesem Bereich richtet sich an Bibliotheken, Archive, Museen, Rechen- und Medienzentren sowie Wissenschaftlerinnen und Wissenschaftler und besteht derzeit aus den folgenden sieben Förderprogrammen:

- *e-Research-Technologien* (Merkblatt 12.19)
- *Open-Access-Publizieren* (Merkblatt 12.20)
- *Informationsinfrastrukturen für Forschungsdaten* (Merkblatt 12.14)
- *Elektronisches Publizieren und digitale Wissenschaftskommunikation* (Merkblatt 12.11)
- *Fachinformationsdienste für die Wissenschaft* (Merkblatt 12.10)
- *Überregionale Lizenzierung* (Merkblatt 12.18)
- *Erschließung und Digitalisierung* (Merkblatt 12.15)

Im Rahmen aller Förderprogramme werden zusätzlich Ausschreibungen veröffentlicht, um gezielt Schwerpunkte zu setzen, Entwicklungen zu stimulieren und Themen zu positionieren. Diese Ausschreibungen haben meist einen direkten Bezug zum digitalen Wandel in den Wissenschaften, wie an folgenden Beispielen deutlich wird: „Forschungsdaten in der Praxis (2015)“, „Nachhaltigkeit von Forschungssoftware (2016)“ und „Open-Access-Transformationsverträge (2017)“.

Zahlreiche Herausforderungen und Bedarfe, die durch den digitalen Wandel in den Wissenschaften entstehen, können mit Projektanträgen im Förderprogramm *e-Research-Technologien* adressiert werden. Gerade dem Anspruch, für Projekte aus dem gesamten Themenspektrum der wissenschaftlichen Informationsversorgung offen zu sein, insbesondere auch für unkonventionelle und grundlegend neue Fragestellungen, wird mit dem Programm *e-Research-Technologien* begegnet. In dem Programm können Projekte, die sich sowohl auf unterschiedliche Teilaufgaben der Informationsversorgung als auch auf unterschiedliche Entwicklungsstadien von Infor-

mationsinfrastrukturen beziehen, beantragt und gefördert werden. Insbesondere bietet es auch einen Förderrahmen für Projekte, die sich in forschender und experimentierender Weise mit aktuellen Frage- und Aufgabenstellungen wissenschaftlicher Informationsinfrastrukturen befassen wollen. Des Weiteren können Studien zur Analyse von mit der digitalen Wissenschaft zusammenhängenden Themen beantragt werden, sofern die Ergebnisse einer solchen Studie in die Entwicklung, Implementierung oder Konsolidierung von e-Research-Technologien fließen.

**Wissenschaftliche Geräte und Informationstechnik.** Erfolgreiche Wissenschaft benötigt Zugang zu aufwendigen und speziellen Gerätetechnologien. Die DFG fördert projektspezifische Geräte innerhalb der allgemeinen Forschungsförderung. Forschungs Großgeräte mit infrastrukturellem Charakter, die eine Voraussetzung für mehrere Forschungsvorhaben darstellen, können bei der DFG im Programm Forschungs Großgeräte nach Art.91b GG beantragt werden. Zusätzlich können im Programm Großgeräte der Länder Anträge für Investitionsvorhaben gestellt werden, deren Ziel der Einsatz von Großgeräten in Lehre, Forschung und Krankenversorgung ist. In diesem Programm werden auch Anträge zur Vernetzung sowie zu medizinischen Softwaresystemen an Hochschulen begutachtet, die Strategien für eine zeitgemäße, wissenschaftlich nutzbringende und kosteneffiziente IT-Infrastruktur verfolgen. In beiden Programmen können Hochschulen je nach vorgesehenem Einsatz auch Rechner- und Speichersysteme als Großgerät beantragen. Zusätzlich zu den Hauptgeräten können bei datenintensiven Großgeräten und bei Technologien, die aufwendige Auswertungen erfordern, Computer, Speichersysteme oder Software auch als Zubehör mitbeantragt werden.

Neben den etablierten Programmen der Geräte- und IT-Infrastrukturförderung wurden Ende 2017 drei neue Programme der gerätebezogenen Forschungsförderung ins Förderportfolio der DFG aufgenommen. *Neue Geräte für die Forschung* soll Wissenschaftlern und Wissenschaftlerinnen ermöglichen, neuartige Geräte für den Einsatz in der Wissenschaft bis hin zum Prototypen selbst zu entwickeln. Das Programm ist themenoffen und bietet im Hinblick auf digitale Technologien auch die

Möglichkeit, die zugehörige Softwareentwicklung zu fördern. Im Programm *Gerätezentren/Core Facilities* wird an Hochschulen die Einrichtung und Weiterentwicklung offener, wissenschaftlich betreuter Zugangsmöglichkeiten zu aufwendigen Geräteinfrastrukturen und -technologien gefördert. Bei bisher geförderten Anträgen waren dabei Datenspeicherung, Analyse oder Simulation in das Serviceangebot des Zentrums bereits vielfach wesentlicher Bestandteil. In *Großgeräteinitiativen* werden bislang noch wenig erprobte, aufwendige Gerätetechnologien durch dedizierte Ausschreibungen gefördert. Zur Themenauswahl für die Ausschreibung einer Großgeräteinitiative nimmt die DFG Vorschläge aus der Wissenschaft in Form von strukturierten Konzepten an. Bei der Betrachtung vergangener Großgeräteinitiativen wird offenkundig, dass die Einbeziehung spezifischer informationstechnischer Expertise und entsprechender IT-Infrastrukturen stetig an Bedeutung zugenommen hat.

Darüber hinaus kann die DFG auf spezielle Impulse aus der Wissenschaft reagieren. So wurde im Jahr 2016 eine Ausschreibung zum *Performance Engineering für wissenschaftliche Software* durchgeführt. Die Förderung zielt darauf ab, Wissenschaftler und Wissenschaftlerinnen unterschiedlicher Disziplinen an die speziellen Softwareanforderungen der Hochleistungsrechner heranzuführen, um Codeoptimierungen zu ermöglichen, die einen effizienten Einsatz der begrenzten Rechenressourcen in Deutschland begünstigen. Die Ausschreibung kann als eine Reaktion auf die zunehmende Bedeutung von Simulationen und rechenintensiven Anwendungen in bislang weniger einschlägigen Wissenschaftsbereichen angesehen werden. Informationen zu den geförderten Projekten finden sich unter dem Link [www.gepris.de/PE](http://www.gepris.de/PE).

In den Lebenswissenschaften gewinnen OMICs-Ansätze zunehmend an Bedeutung. Die Erkenntnis, dass insbesondere der Einsatz des Next Generation Sequencings (NGS) die Infrastruktur an Hochschulen an ihre Grenzen bringt und große Projekte nur schwer finanzierbar sind, führte zur Entscheidung des Senats, eine Ausschreibung zum NGS umzusetzen. Im Rahmen des Programms werden Investitionsmittel bis zur Höhe von insgesamt 10 Mio. € finanziert sowie zahlreiche Personalstellen in der Bioinformatik. An diesem Beispiel wird deutlich, welches Gewicht einer einschlägigen qualifizierten Betreuung von Projekten

der Lebenswissenschaften durch die Informatik zukommt.

## Stellungnahmen und Leitlinien der DFG

### Forschungsdaten

Eine wesentliche Auswirkung des digitalen Wandels in den Wissenschaften ist der fachübergreifende Trend einer stetigen und teils rasanten Zunahme der Menge an verfügbaren Forschungsdaten. Die Gründe für die Zunahme der Datenmenge sind fachspezifisch verschieden: In Teilen der Geisteswissenschaften werden Texte digitalisiert, in vielen experimentell arbeitenden Fächern entstehen umfangreiche Datenmengen durch apparative und methodische Weiterentwicklungen sowie durch algorithmische Berechnungen, Modellierungen und Simulationen. Selbst die Menge an digital verfügbaren wissenschaftlichen Publikationen wächst zunehmend und diese sind ebenso wie alle anderen digitalen Informationen selbst auch Forschungsdaten. Zusätzlich erhöht sich durch die stetig wachsende Verfügbarkeit von Informationen aus dem Internet die Menge an Daten für ganz unterschiedliche wissenschaftliche Fragestellungen. Geeignete Beispiele finden sich in der Bioinformatik, der Medizininformatik, der Linguistik oder der Wirtschaftsinformatik. Da jede wissenschaftliche Arbeit auf qualitativ hochwertige Forschungsdaten angewiesen ist, finden zunehmend intensive Debatten über Qualitäts- und Standardisierungsprozesse für Forschungsdaten statt.

Die DFG hat aus diesem Grund im Jahr 2015 übergeordnete Leitlinien zum Umgang mit Forschungsdaten veröffentlicht und die unterschiedlichen Fachbereiche aufgefordert, den Diskurs fortzusetzen und fachspezifische Mindestanforderungen für das Forschungsdatenmanagement auszuarbeiten. Dieser Aufforderung sind inzwischen zahlreiche Fach-Communities gefolgt: Die Biodiversitätswissenschaften, die Erziehungswissenschaften, die Sozial-, Verhaltens- und Wirtschaftswissenschaften und die Psychologie haben konkretere Empfehlungen veröffentlicht, die bei der Bewertung von Forschungsanträgen Berücksichtigung finden. Mit diesem Vorgehen versucht die DFG, sowohl den differenzierten Anforderungen der fachlichen Communities als auch der (Gesamt-)Verantwortung für die Sicherstellung von Qualitätsmaßstäben im wissenschaftlichen Arbeiten gerecht zu werden.

### Forschungssoftware

Komplementär zur stetig wachsenden Datenmenge ergibt sich die ebenfalls fachübergreifende Notwendigkeit der Nutzung von Forschungssoftware, um die verfügbaren Forschungsdaten zu analysieren, kommentieren, annotieren, sie letztlich adäquat nutzen sowie – in einzelnen Fächern – überhaupt per Berechnung, Modellierung oder Simulation erzeugen zu können. Somit sind in fast allen Fächern softwarebasierte Methoden zur Datenerhebung, -auswertung und -visualisierung entwickelt worden, ebenso wie Software, mit der auf Basis vorhandener Daten Simulationen und Modelle ermöglicht werden. Die steigende Bedeutung von Data-Mining-Technologien, von Modellierungen, Simulationen und anderen Methoden zur Verarbeitung großer Datenmengen erfasst immer mehr Fächer und nimmt auch keine Rücksicht auf die bislang „üblichen“ fachlichen Abgrenzungen. Im Rahmen der Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen werden Handlungsempfehlungen für die Entwicklung, die Nutzung und das Anbieten von Forschungssoftware formuliert. Auch im Rahmen der Initiative „Knowledge Exchange“ hat die DFG an der Aufarbeitung zum Verständnis der Schlüsselrolle von Forschungssoftware für das heutige wissenschaftliche Arbeiten mitgewirkt. Vor diesem Hintergrund hat die DFG im Jahr 2016 die Ausschreibung „Nachhaltigkeit von Forschungssoftware“ durchgeführt. Ziel der Ausschreibung sind Aufbau und Erprobung von Infrastrukturen, um Forschungssoftware für einen größeren Anwenderkreis nutzbar zu machen und nachhaltig anzubieten. Die Vorhaben sollen als Best-Practice-Beispiele eine positive Signalwirkung für die wissenschaftliche Softwareentwicklung und für Infrastruktureinrichtungen haben.

### Digitale Informationsinfrastrukturen

Die DFG versteht digitale Informationsinfrastrukturen in Anlehnung an die Definition des Rates für Informationsinfrastrukturen (RfII) als „technisch und organisatorisch vernetzte Dienste und Angebote für den Zugang zu und die Erhaltung von Daten-, Informations- und Wissensbeständen. Im Sinne des RfII dienen sie primär Forschungszwecken, sie sind häufig Forschungsgegenstand und haben stets eine ermöglichende Funktion.“ Die inhaltliche Auseinandersetzung mit der stetigen Weiterentwicklung

von Informationsinfrastrukturen leistet in der DFG der Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme (AWBI). Etwa alle fünf Jahre unterzieht der AWBI das Förderhandeln der DFG in diesem Bereich einer Bestandsaufnahme und setzt diese in Kontrast zu den neu entstandenen Herausforderungen der wissenschaftlichen Informationsversorgung. Im Jahr 2012 ist unter dem Titel „Die Digitale Transformation weiter gestalten“ das letzte Positionspapier erschienen, aktuell ist nun das Positionspapier „Förderung von Informationsinfrastrukturen für die Wissenschaft“ veröffentlicht worden.

### **Digitale Editionen und Sprachkorpora**

Auf Initiative des Fachkollegiums 104 „Sprachwissenschaften“ der DFG haben in 2012 und 2013 zwei Rundgespräche zur Erhebung von mündlichen und schriftlichen Sprachkorpora stattgefunden. Als Ergebnis dieser Rundgespräche haben sich Arbeitsgruppen gebildet, die Empfehlungen zur Erhebung von Sprachkorpora formuliert haben. Diese Empfehlungen beziehen sich auf „Datentechnische Standards und Tools bei der Erhebung von Sprachkorpora“ und auf „Rechtliche Aspekte bei der Handhabung von Sprachkorpora“. Beide Papiere sollen als Handreichung für alle in diesem Bereich Forschenden dienen, insbesondere auch für Antragstellende und Gutachtende in DFG-Programmen.

Das Fachkollegium 105 „Literaturwissenschaft“ der DFG beschreibt grundlegende „Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft“ für die Antragstellung und Begutachtung. Diese Kriterien sollen helfen, einen doppelten Anspruch an die Wissenschaftlichkeit von Editionen zu erfüllen, denn wissenschaftliche Editionen haben einerseits die fachwissenschaftlichen Anforderungen zu beachten und sollten andererseits den etablierten editionswissenschaftlichen Standards genügen, die die Editionsphilologie als eine transdisziplinäre Grundlagenwissenschaft entwickelt hat.

### **Informationsverarbeitung an Hochschulen**

Alle fünf Jahre veröffentlicht die Kommission für IT-Infrastruktur der DFG eine Stellungnahme zur Informationsverarbeitung an Hochschulen – Organisation, Dienste und Systeme. In dieser Stellungnahme werden die aus der vielfältigen IT-bezogenen Antrags- und Projektbewertung ge-

machten Erfahrungen in Form von Empfehlungen zusammengefasst. Diese Empfehlungen tragen zu einem qualitätsgesicherten, nach bundeseinheitlichen Maßstäben erfolgenden Infrastrukturaufbau in der Informationstechnologie bei. Die Stellungnahme umfasst eine strategische, eine organisatorische und eine technische Sicht auf den Themenkomplex: Die strategische Sicht beinhaltet Empfehlungen, welche Punkte ein standortbezogenes IT-Konzept adressieren sollte und wie eine organisatorische Abstimmung zwischen verantwortlichen IT-Beauftragten und der Hochschulleitung erfolgen kann. Die organisatorische Sicht äußert sich zu der Frage, welche Dienste und Mindeststandards zu einer zeitgemäßen Hochschulausstattung gerechnet werden können bzw. erfüllt sein sollten und richtet ein Augenmerk auf Themen wie Datenschutz und die Bereitstellung von Cloud-Diensten durch Drittanbieter. Die technische Sicht befasst sich mit den an Hochschulen benötigten technischen Grundlagen, startend mit den Netzen, Rechnern der unterschiedlichen Leistungsklassen bis hin zu Höchstleistungsrechnern. Sie adressiert dabei gleichwohl Speicher- und Archivsysteme sowie Energiebetrachtungen, welche heute zu den relevanten Faktoren bei der Betrachtung der Betriebskosten gehören. Weiterhin trifft die Stellungnahme Aussagen zu Möglichkeiten der Standortprofilbildungen und den besonderen Herausforderungen in der Universitätsmedizin. Zusammenfassend zielt das Papier darauf ab, einen sich entwickelnden Blick auf die Herausforderungen bei der Planung, Finanzierung und Unterhaltung wissenschaftlicher Informationstechnikinfrastrukturen zu ermöglichen, um so den einzelnen Hochschulen eine unterstützende Richtschnur anzubieten, entlang derer die eigene Umsetzung erfolgreich gestaltet werden kann.

### **Zusammenfassung**

Der digitale Wandel prägt Wissenschaft und Gesellschaft in hohem Maß und auf vielfältige Weise. Die DFG führt derzeit eine Vielzahl von Aktivitäten durch, um dessen Auswirkungen auf die Forschung nachhaltig zu gestalten. Dazu zählen fachspezifischen Reflexionen, Weiterentwicklungen der Förderinstrumente und auch die Beratung von Politik und Gesellschaft. Der vorliegende Beitrag dient zudem als Wegweiser durch das aktuelle Förderportfolio der DFG, um so für unterschiedli-

che Vorhaben die passenden Fördermöglichkeiten aufzuzeigen. Darüber hinaus bietet der Beitrag eine Übersicht bisheriger Aktivitäten, Leitlinien und Stellungnahmen der DFG zum Thema digitaler Wandel und IT in den Wissenschaften. Die fachübergreifen-

den Stellungnahmen basieren auf den Erfahrungen einer Vielzahl von Förderentscheidungen und zeigen notwendige Rahmenbedingungen sowie Strategien für erfolgreiches wissenschaftliches Arbeiten im digitalen Zeitalter auf.

# Informatik als Treiber für E-Science

Positionspapier der Gesellschaft für Informatik

Gesellschaft für Informatik e. V. (GI)

## Informatik – Grundlage für alle Wissenschaften

Albert Einstein nutzte eine Kreidetafel. Heutige Wissenschaftlerinnen und Wissenschaftler brauchen Informatik. In den Erd- und Umweltwissenschaften liefern Satelliten und Sensornetzwerke kontinuierlich massive Datenströme. Von den Sentinel-Satelliten des Europäischen Copernicus-Programms werden beim Deutschen Fernerkundungsdatenzentrum (DFD) bis zu zehn Terabyte pro Tag archiviert und verarbeitet. Am Deutschen Klimarechenzentrum (DKRZ) wird durch Klimasimulationen ein jährliches Datenwachstum von etwa 75 Petabyte erzeugt. Lebenswissenschaften wie Molekularbiologie, Biomedizin oder Lebensmitteltechnologie können durch die Untersuchung molekularer Prozesse auf der Basis von Genom-Daten ein neues Verständnis von Lebensvorgängen erreichen. Dies kann zu maßgeschneiderten Therapien in der Medizin führen oder die Züchtung robuster und ertragreicher Nutzpflanzen ermöglichen. Die dafür notwendige Datenmenge ist allein im European Bioinformatics Institute (EBI) von 15 auf 25 Petabyte im Jahr 2015 angestiegen.

Um die Datenmengen aktueller Forschung verarbeiten zu können, stellt z. B. die European Grid Infrastructure (EGI) 530 000 logische CPUs, 200 Petabyte Festplattenspeicher und 300 Petabyte Bandlaufwerke zur Verfügung. Diese Ressourcen werden gemeinschaftlich von 350 Rechenzentren in 56 Ländern der Welt betrieben. EGI ist nur möglich durch leistungsfähige Datenbanken, schnelle Rechnernetze und hohe Parallelität.

Die Ergebnisse der Informatik verändern in allen Wissenschaftsbereichen technologische

Rahmenbedingungen und die Art und Weise, wie Forschung durchgeführt wird. Sie verändern damit auch Kostenstrukturen und schaffen verbesserte Grundlagen für Forschung und Innovation. Studien zeigen, dass der Mehrwert einer effizienten IT-Infrastruktur um das Zwanzigfache höher sein kann als deren operationelle Kosten [2].

## Warum ist E-Science wichtig?

E-Science beschreibt eine Vision des zukünftigen wissenschaftlichen Arbeitens und die mittel- bis langfristige Entwicklung der wissenschaftlichen Praxis dorthin. E-Science betrachtet im engeren Sinn die Anwendung von Informations- und Kommunikationstechnologie (IKT) in der modernen Wissenschaft und umfasst ein Spektrum von der Vorbereitung und Durchführung von Untersuchungen, über die Datenerfassung, Verarbeitung und Verbreitung von Ergebnissen bis hin zur Langzeitspeicherung und dem Zugang zu relevanten Materialien und Artefakten [3]. Damit fördert E-Science Innovation in der kollaborativen, rechen- und datenintensiven Forschung in allen Disziplinen über den gesamten Forschungslebenszyklus hinweg [1]. Dieser Transformationsprozess, die Umsetzung der Digitalen Agenda in der Wissenschaft, betrifft sowohl Natur- und Ingenieurwissenschaften als auch Sozial- und Geisteswissenschaften.

<https://doi.org/10.1007/s00287-018-01129-y>  
© Springer-Verlag Berlin Heidelberg 2018

Gesellschaft für Informatik e. V. (GI)  
Ahrstr. 45, 53175 Bonn  
E-Mail: [gs@gi.de](mailto:gs@gi.de),  
[www.gi.de](http://www.gi.de)

Ein wichtiger Treiber der Entwicklung zu E-Science sind die großen gesellschaftlichen Herausforderungen wie Klimawandel und Energieversorgung, Gesundheit und demographischer Wandel oder der nachhaltige Umgang mit natürlichen Ressourcen. Von der Wissenschaft werden zunehmend konkrete Beiträge zum Umgang mit diesen Herausforderungen erwartet. Die Grundlagen dazu werden in global verteilten, kollaborativen und interdisziplinären Kooperationen erforscht, deren Erfolg von leistungsfähigen digitalen Arbeitsplattformen und einer dadurch möglichen integrierten Nutzung vorhandener Hard- und Software abhängig ist. Sensornetzwerke und Monitoringsysteme werden mit Hochleistungsrechnern und Speichersystemen durch leistungsfähige Middleware zu integrierten Forschungsplattformen verbunden. Die schnelle Analyse und Auswertung sehr großer experimenteller und empirischer Datenmengen schafft erst die Voraussetzungen, komplexe Zusammenhänge und Wechselwirkungen auf Systemebene zu erkennen und mit theoretischen Modellen abzugleichen. Diese Datenmengen müssen in global verteilten Wissenschaftsteams gemeinschaftlich interpretiert werden, um daraus neue Ergebnisse und Handlungsoptionen abzuleiten.

Die Fähigkeit zum Umgang mit der digitalen Transformation in Forschung und Lehre stellt eine Schlüsselkompetenz für unsere Universitäten, Forschungseinrichtungen und unsere Wissensgesellschaft insgesamt dar. E-Science setzt eine hochwertige technische Infrastruktur voraus. Die Förderung, Gestaltung und Umsetzung des Transformationsprozesses geht jedoch weit über die Bewältigung von technischen Herausforderungen hinaus. Das Themenfeld E-Science ist aktuell durch eine Vielzahl von nationalen und internationalen Aktivitäten gekennzeichnet. Die konkrete Umsetzung ist von spezifischen forschungspolitischen Rahmenbedingungen und komplementären Fördermaßnahmen abhängig.<sup>1</sup> Trotz der treibenden Kraft der IKT für E-Science wird die wesentliche Bedeutung der Wissenschaftsdisziplin Informatik für die Gestaltung und Umsetzung dieser für

den Forschungsstandort Deutschland wichtigen Entwicklungen noch nicht wahrgenommen.

## Handlungsbedarf – was ist zu tun?

Die Entwicklung von E-Science erfolgt aktuell in einer Vielzahl von Projekten, die von unterschiedlichen Organisationen im Rahmen von Programmen oder Einzelprojekten gefördert werden. Das Ziel, wissenschaftliche Daten und Software konsistent und nachhaltig zu verwalten sowie zwischen unterschiedlichen Fachdisziplinen austauschen und unabhängig von der Quelle nutzen zu können, wird durch fachbezogene Einzelentwicklungen nicht immer befördert. Gleiches gilt für Softwarewerkzeuge. Außerhalb des jeweiligen Faches sind diese Ressourcen nur mit erheblichem Aufwand verfügbar oder gar unzugänglich.

Wissenschaftsdaten sind heterogen, dynamisch und folgen ebenso wenig einer strengen Normierungslogik wie interdisziplinäre Forschungsprojekte. Der Nutzen von Wissenschaftsdaten ergibt sich aus deren „fitness for use“, dem Nutzen für zukünftige Forschung, deren Anforderungen, z. B. an die Struktur der Daten, wir heute noch gar nicht kennen (können). Zur Prozessierung und zur Analyse von Daten, aber auch für die Modellierung und Simulation werden von Wissenschaftlern verbreitet Algorithmen und Programme entwickelt. Resultierende Softwarewerkzeuge werden nach Projektende häufig weder weiterverwendet noch weiterentwickelt. Ihr Entwicklungsaufwand ist aber meist genauso hoch wie der des wissenschaftlichen Experiments selbst. Während wir bei Mikroskopen, Teilchenbeschleunigern, Forschungsschiffen und anderen wissenschaftlichen Großgeräten eine möglichst vielfältige Nutzung bereits bei der Anschaffung einbeziehen, steckt dies bei wissenschaftlicher Software noch in den Kinderschuhen.

Als wissenschaftliche Fachdisziplin ist die Informatik dazu prädestiniert, derartige Entwicklungen aufzuzeigen und zugleich in ihrer Eigenschaft als konstruktive Disziplin belastbare Alternativen anzubieten. Kernthemen der Informatik waren und sind seit Jahren die Schaffung von Interoperabilität und die Verknüpfung heterogener Infrastrukturen. Konzepte und Architekturen zur Integration von Software und Hardware wurden ebenso von Informatikern entwickelt wie weltweite Standards zur Datenerhaltung, zur Vernetzung und zum Management globaler Informationssysteme. Aus Sicht von In-

<sup>1</sup> Beispiele:

- DE: Rat für Informationsinfrastrukturen <http://www.rfii.de>
- GB: JISC <https://www.jisc.ac.uk/>; Software Sustainability Institute <http://www.software.ac.uk/>
- NL: Netherlands eScience Center <https://www.esciencecenter.nl/>
- EU: Platform of National eScience Centers in Europe <https://plan-europe.eu/>

formatikerinnen und Informatikern ist es möglich, sich diesen Problemen wissenschaftlich zu nähern und bestehende Konzepte der Informatik und angrenzender Bereiche anzuwenden und weiterzuentwickeln. Es geht also nicht nur um eine technische Unterstützungsleistung für bestimmte Fachgebiete, sondern um die Chance, durch intensive Zusammenarbeit und gelebte Interdisziplinarität Fachsilos und Corpora miteinander zu vernetzen. Dies würde die Forschung sowohl im jeweiligen Fachgebiet als auch in der Informatik voranbringen. Daraus leiten wir Anforderungen an Wissenschaftspolitik, Universitäten und Forschungseinrichtungen sowie die Informatik selbst ab.

### **Herausforderungen für die Wissenschaftspolitik**

Für die zielgerichtete Entwicklung von E-Science sind positive forschungspolitische Rahmenbedingungen notwendig. Die deutsche Wissenschaftspolitik sollte sich klar zu E-Science und dem koordinierten Aufbau einer neuen Generation von Infrastrukturen und Anwendungen für Forschung und Lehre bekennen. Diese Aktivitäten sollten grundsätzlich offenen Standards folgen, z. B. Zugang zu Daten und Software (Open Science) oder Vernetzung von Daten (Linked Open Data). Aufbauend hierauf können fach- und domänenspezifische Standards und Konventionen etabliert werden, die jedoch mit anderen Disziplinen interoperabel bleiben müssen. Gerade im Hinblick auf zukünftige interdisziplinäre Forschungsanstrengungen sollte daher eine gemeinsame methodische und technische Basis für einen nachhaltigen Zusammenhalt und Erweiterungsmöglichkeiten sorgen.

Die bereits existierenden, vielfältigen Wissenschaftsprogramme, welche auf die Digitalisierung einzelner Wissenschaften zielen, könnten durch eine solche Initiative gestärkt werden. Ein mögliches nationales Programm sollte u. a. folgende übergreifende Themen adressieren:

- Wissenschaftliches Datenmanagement
- Etablierung von Referenzarchitekturen und Standards
- Semantische Interoperabilität und Integration von Daten
- Software Engineering für wissenschaftliche Anwendungen

- Integrierte Modellierung und Systemwissenschaften, Data Science und Simulation
- Aufbau von Forschungsumgebungen mit kollaborativer Mensch-Maschine-Kooperation
- Einbeziehung von E-Science-Praktiken in die Graduiertenausbildung

### **Herausforderungen für Universitäten und Forschungseinrichtungen**

Wissenschaft wird heute in globalem Kontext und internationalem Wettbewerb durchgeführt. Qualität und Innovationsgeschwindigkeit sind gleichermaßen wichtige Anforderungen. Neben aller Kreativität im Finden und Beantworten von Forschungsfragen entscheidet dabei auch der gekonnte Umgang mit Methoden und Werkzeugen einer digitalisierten und datenorientierten Wissenschaftswelt sowie die effiziente Vernetzung mit anderen global verteilten Wissenschaftlerinnen und Wissenschaftlern über Erfolg oder Misserfolg. Auch die Einbindung in und die Nutzung von nationalen<sup>2</sup> und internationalen Forschungsinfrastrukturen<sup>3</sup> hat hier einen zunehmenden Einfluss. Universitäten und Forschungseinrichtungen müssen dabei zwei Entwicklungen beachten:

Der *Betrieb* eigener IT-Infrastrukturen ist eine Grundvoraussetzung für die aktuelle und zukünftige Wissenschaftspraxis. Das ist eine Herausforderung sowohl in finanzieller und personeller, aber auch in konzeptioneller und organisatorischer Hinsicht. Rechenzentren und IT-Servicezentren sind prinzipiell unverzichtbare Dienstleister für die Wissenschaft und sollten die Sprache der Wissenschaft verstehen können. Sollten sie fachlich oder technisch dazu nicht fähig sein, werden die Fachbereiche das Heft des Handels selbst in die Hand nehmen und parallele Strukturen aufbauen, mit negativen Konsequenzen hinsichtlich interdisziplinärer und interinstitutioneller Kommunikation sowie Kosten, Sicherheit und Nachhaltigkeit. Der Schlüssel zum Erfolg liegt deswegen in einer zunehmenden Professionalisierung und größeren Anwendernähe zentraler Serviceeinrichtungen an Universitäten und außeruniversitären Forschungsorganisationen.

<sup>2</sup> BMBF: <https://www.bmbf.de/de/roadmap-fuer-forschungsinfrastrukturen-541.html>.

<sup>3</sup> European Commission: <http://ec.europa.eu/research/infrastructures/>.

Die *Entwicklung* neuer Wissenschaftsfelder benötigt zunehmend Kenntnisse in Werkzeugen und Methoden der Informatik. In vielen Wissenschaftsbereichen sind Projekte nur noch mit erheblichem IT-Aufwand erfolgreich durchführbar. Herausragende Kenntnisse in der Anwendung von Informatik auf die eigene Fachwissenschaft können ein Berufungskriterium für junge Wissenschaftlerinnen und Wissenschaftler werden. Für die Ausbildung junger Wissenschaftlerinnen und Wissenschaftler stellt E-Science einen notwendigen Baustein der zukünftigen Graduiertenausbildung dar, ebenso wie wissenschaftliches Publizieren oder Methodenkenntnisse. Die Ausprägung zukünftiger Formate der Graduiertenunterstützung erfordert neben der Bereitstellung von Werkzeugen auch die personelle Einbindung von Informatikkompetenzen. Einschlägig qualifizierte und interessierte Informatikerinnen und Informatiker an Universitäten sollten sich in eine solche Graduiertenausbildung einbringen.

## **Herausforderungen für Informatikerinnen und Informatiker**

Informatikerinnen und Informatiker, in den Wissenschaftsdomänen arbeitende „Bindestrich“-Informatiker (z. B. Medizin-, Bio-, Wirtschafts-, Geo-Informatiker) sowie Domänenexpertinnen und -experten sind aufgefordert, Plattformen zu schaffen, in denen übergreifende Fragen gemeinsam diskutiert und Lösungswege identifiziert werden können. Anfängliche Mehrinvestitionen rentieren sich durch höhere Nachhaltigkeit und Wiederverwendung gefundener Lösungen. Interessierte Informatikerinnen und Informatiker sollten sich an der Entwicklung offener Standards für den Austausch und die Verarbeitung wissenschaftlicher Daten beteiligen, die Interoperabilität mit anderen Fachdomänen in den Vordergrund stellen. Wir können und sollten als Wissenschaftlerinnen und

Wissenschaftler in Deutschland nicht erwarten oder darauf vertrauen, dass eine einzelne große Institution im Alleingang Standards und Interoperabilität durchsetzt. Eine breit angelegte Koalition von Fachdomänen mit entsprechend breit gefächerten Fach- und IT-Kompetenzen scheint allerdings essenziell, um eine offene, leistungsfähige und effiziente wissenschaftliche Dateninfrastruktur zu erreichen.

## **Die Rolle der Gesellschaft für Informatik e. V.**

Die Gesellschaft für Informatik steht bereit, einer solchen Entwicklung Raum zur Diskussion und Entfaltung zu bieten. Wir veranstalten Konferenzen und Workshops, die digitale und persönliche Plattformen der Kommunikation sein können. Wir führen gemeinsame Projekte mit der Wirtschaft, den Wissenschaften und der Gesellschaft durch. Wir sind Ansprechpartner für die Politik und Medien. Wir arbeiten zusammen mit anderen Fachgesellschaften, den Anforderungsgebern und Anwendern unserer Technologien. Wenn E-Science mehr sein soll als ein Begriff mit vielen Bedeutungen, sind Informatikerinnen und Informatiker der Schlüssel für fach- und domänenübergreifende Ansätze zum Austausch und zur Analyse von Daten aus unterschiedlichen Bereichen anhand gemeinsam entwickelter Prozess- und Datenstandards. Die Gesellschaft für Informatik kann als Katalysator für entsprechende Projektansätze ihren Teil beitragen.

## **Literatur**

1. IEEE International Conference on eScience, <https://escience-conference.org/>, last access: 30.11.2018
2. Beagrie (2016) The Value and Impact of the European Bioinformatics Institute. Full Report January 2016. <https://beagrie.com/static/resource/EBI-impact-report.pdf>, last access: 30.11.2018
3. Bohle S (2013) What is E-science and how should it be managed? Nature.com, Spektrum der Wissenschaft (Scientific American). [http://www.scilogs.com/scientific\\_and\\_medical\\_libraries/what-is-e-science-and-how-should-it-be-managed/](http://www.scilogs.com/scientific_and_medical_libraries/what-is-e-science-and-how-should-it-be-managed/), last access: 30.11.2018

# Eine nationale Daten- und Analyseinfrastruktur als Grundlage digitaler Souveränität

Volker Markl

## Hintergrund, Herausforderungen und Potenziale

Die Vielfalt und die exponentiell wachsende Menge digitaler Daten, die heute in Wirtschaft, Medizin, Mobilität sowie in vielen weiteren Lebensbereichen anfallen, bieten zusammen mit neuen Verfahren des maschinellen Lernens und der künstlichen Intelligenz gänzlich neue Chancen, automatisiert neue Muster und Zusammenhänge zu erkennen, Frühwarnungen zu erzeugen oder Prozesse zu steuern. Maschinen können – im Gegensatz zu Menschen – aus großen Datenströmen in Echtzeit lernen, sie helfen uns, die neue Flut an Daten und Informationen zu beherrschen und produktiv zu nutzen. *Dabei spielen Menge und Qualität der Daten über die Möglichkeiten und Mächtigkeit der maschinellen Lern- und KI-Verfahren und Anwendungen eine entscheidende Rolle.*

Dauer und Kosten von Analyseprozessen sind dank der neuen Datenquellen und neuer KI-Verfahren für viele Berufe, für die Wissenschaft und den Bürger drastisch gesunken. Relevante Inhalte können schneller gefunden und wichtige Zusammenhänge schneller erkannt werden. Die Geschwindigkeit der Wissenserzeugung aus Daten ist heute mehr denn je zum entscheidenden Wettbewerbsvorteil geworden.

Wenn Daten die Produktionsfaktoren der Zukunft sind, dann sind *Datenzugang* und *Datenanalysekompetenz* die Schlüsselfaktoren für zukünftige Wettbewerbsfähigkeit. In beiden Belangen ist Deutschland leider im internationalen Vergleich ins Hintertreffen geraten.

Deutschland nimmt zwar nach wie vor eine Spitzenposition in der (wissenschaftlichen) Techno-

logieforschung ein, jedoch sind es derzeit vor allem US-amerikanische und zunehmend auch asiatische Unternehmen, die wichtige Daten- und Analyseinfrastrukturen für die neue digitale Wirtschaft bereitstellen. Die Angebote der großen Software-as-a-Service(SaaS)-Anbieter Amazon Web Services, Google Data Flow, Microsoft Azure, IBM Bluemix dominieren aufgrund ihrer überlegenen Funktionalität, Performance und Usability sowohl in wissenschaftlichen wie auch in wirtschaftlichen Anwendungen. Entsprechend machen sich europäische Akteure immer stärker von diesen kostenpflichtigen Angeboten abhängig. Und nicht nur das: In der Folge liegen die meisten europäischen Daten – Konsumentendaten wie auch Unternehmensdaten – außerhalb Europas und werden von Software nichteuropäischer Unternehmen in Drittländern analysiert. Wichtigste Voraussetzung für digitale Souveränität ist jedoch uneingeschränkte Verfügungsgewalt über unsere Daten und den Quellcode der Analysesoftware sowie die Durchführung der Analysen auch geographisch in unserem Rechtsrahmen.

Dem Markt gelingt es nicht, eine unabhängige deutsche oder europäische Alternative mit einem eigenen, gleichwertigen Angebot zu entwickeln, obwohl das technologische Potenzial dafür durchaus vorhanden ist, wie Beispiele international anerkannter und eingesetzter Technologieplatt-

<https://doi.org/10.1007/s00287-018-01136-z>  
© Springer-Verlag Berlin Heidelberg 2018

Volker Markl  
Technische Universität Berlin und DFKI,  
Fachgebiet Datenbanksysteme und  
Informationsmanagement,  
Einsteinufer 17, 10587 Berlin  
E-Mail: volker.markl@tu-berlin.de

## Zusammenfassung

Daten sind der wesentliche Produktionsfaktor der Zukunft. Informatik und Data Science werden zur Grundlage der meisten anderen wissenschaftlichen Fächer wie Physik, Chemie, Wirtschaft. Sie sind in ihrer Bedeutung damit der Mathematik vergleichbar. Datenzugang und Datenkompetenz sind die Schlüsselfaktoren für die zukünftige Wettbewerbsfähigkeit. In beiden Belangen ist Deutschland ins Hintertreffen geraten. Zur Erlangung von Datensouveränität benötigt Deutschland bzw. Europa eine nachhaltig (und öffentlich) betriebene Daten- und Analyseinfrastruktur. Diese stellt den Zugang zu Daten zu großen, qualitativ hochwertigen Datenmengen (Internet, Forschungsdaten, öffentliche Daten, z. B. mCloud) sicher und ermöglicht es, deren Analyse und Visualisierung für Unternehmen, Schulen, Universitäten, Forschungseinrichtungen und Bürger zu demokratisieren. Nur auf diese Weise lässt sich die Lücke zu den monopolisierten Datensammlungen amerikanischer IT-Konzerne oder dem staatlich regulierten Zugang zu Daten in China schließen. Eine derartige nationale, allgemein zugängliche Infrastruktur sollte nicht nur Daten, Analysen und Visualisierungen verwalten und kontinuierlich in Echtzeit bereitstellen, sondern gleichzeitig die Algorithmen und Werkzeuge entlang der gesamten Datenwertschöpfungskette (Quellenauswahl, Informationsextraktion und Integration, Analyse und Modellbildung sowie Anwendung und Visualisierung) einfach nutzbar (webbasiert, open-source, wiederverwendbar) bereitstellen, um auf diese Weise durch „Daten und Analysen aus der Steckdose“ Forschung sowie die Innovation in datengetriebene Anwendungen in Wirtschaft, Wissenschaft und Gesellschaft zu befeuern.

formen wie Apache Flink [1], Rapidminer [8] und OpenML [7] zeigen. Grund ist, dass eine vergleichbare kritische Masse hier von keinem Anbieter erreicht werden kann und in Deutschland insbesondere auch die Kompetenz zum Aufbau großer Daten- und Analyseinfrastrukturen bei keinem Unternehmen vollumfänglich vorhanden ist. Um den Leistungsabstand zu den dominierenden Plattfor-

men zu verkürzen und eine eigenständige deutsche oder europäische Entwicklung mit dem damit verbundenen Kompetenzaufbau bei Erstellung und kontinuierlichem Betrieb von großen Daten- und Analyseinfrastrukturen anzustoßen, wird ein Anschlag benötigt. Den Anfang für ein solches Angebot könnte eine nachhaltig (und öffentlich) betriebene Datenanalyseinfrastruktur sein, die im Gegensatz zu den amerikanischen Angeboten weniger auf der Auswertung und Vermarktung von Konsumdaten basiert, sondern vielmehr breitere, für die europäische Wirtschaft und Gesellschaft relevantere Domänen wie beispielsweise Mobilität und Gesundheit in den Fokus stellt und gleichzeitig den Zugang zu großen, qualitativ hochwertigen Datenmengen (Internet, Forschungsdaten, öffentliche Daten) und deren *Analyse und Visualisierung* in Echtzeit für Unternehmen, Universitäten, Forschungseinrichtungen, Schulen und Bürger ermöglicht und demokratisiert.

Damit entsteht eine ausreichend große „Spielwiese“, ein Nährboden für datengetriebene Technologieinnovationen und Unternehmensgründung entlang der Wertschöpfungskette der Daten- und Analyseinfrastruktur, von Datenquellenauswahl über Veredelung der Rohdaten durch Informationsextraktion und Integration bis hin zur Erstellung von Datenanalysen und Modellbildung durch Verfahren des maschinellen Lernens und der künstlichen Intelligenz. Voraussetzung ist die auf Dauer angelegte Etablierung und der Betrieb dieser Infrastruktur. Heutige Forschungsprogramme sind zumeist auf drei Jahre angelegt, in der einzelne Projekte mühsam eine eigene Daten- und Analyseinfrastruktur aufbauen, Daten kurieren und analysieren, um nach Projektabschluss den Server und die Dienste abzuschalten und die gewonnenen wertvollen Daten zu verwerfen. Dieser Prozess erzeugt nicht nur erheblichen Mehrfachaufwand, er verschenkt auch das Potenzial, durch die kontinuierliche Arbeit an einer gemeinsamen Daten- und Analyseinfrastruktur große Datenbestände aufzubauen, Analysewerkzeuge kontinuierlich weiterzuentwickeln und alle Projektergebnisse inklusive Daten und Code nachhaltig auch nach dem Ende eines Projektes für die Allgemeinheit nutzbar zu machen.

Eine *logisch zentrale, nationale, allgemein zugängliche Infrastruktur* könnte dagegen nicht nur eine große Vielzahl und Vielfalt an Daten *kontinuierlich* bereitstellen, sondern gleichzeitig *Werkzeuge*

der gesamten Datenwertschöpfungskette (Aufbereitung, Analyse und Visualisierung) einfach nutzbar (webbasiert, Plug & Play, Kombination von öffentlichen und privaten Daten in einer Analyse) anbieten und über ihre Nutzung stetig weiterentwickeln. Eine derartige Infrastruktur kann nicht nur von Akteuren in der Wirtschaft genutzt werden, sondern auch von Forschungseinrichtungen, Universitäten, Schulen und Bürgern insgesamt. Auf diese Weise kann ein derartiges System ein Innovationsmotor für Ausbildung sowie datenorientierte Wertschöpfung, Geschäftsmodellinnovationen und Unternehmensgründungen werden.

Ein wesentlicher Vorteil einer öffentlich betriebenen Infrastruktur wäre die konsequente Umsetzung offener Standards, mittels derer der derzeit gängige Vendor-Lock-in für Daten- und Analyseinfrastrukturen gebrochen werden könnte. Damit könnte eine nationale bzw. europäische Datensouveränität erreicht und die Abhängigkeit von Anbietern aus Drittländern beim in der Zukunft immer wichtigeren Produktionsfaktor Daten verringern oder ganz abgebaut werden.

Im Bereich der Wissenschaft ist als ein bereits existierender Anwendungsfall einer derartigen Infrastruktur der Sloan Digital Sky Server (SDSS) [12] zu nennen, der als domänenspezifische Infrastruktur mit Daten und Analysefunktionen in einem System bereits seit Jahren in Betrieb ist und vor seiner Einführung auch von einigen Skeptikern aufgrund von Datenvolumen, Anfragekomplexität, etc. als unmöglich angesehen wurde. Wesentliche Protagonisten waren der leider inzwischen verstorbene Turing-Preisträger Jim Gray, der den Begriff des 4. Paradigmas der Wissenschaft [13] geprägt hat. Der SDSS hat zu vielen spannenden wissenschaftlichen Erfolgen geführt, neben der Nutzung durch die wiss. Community auch zum spielerischen Umgang mit den Daten durch Outsider und Hobbyisten, was u. a. in dem Auffinden von neuen Galaxien und anderer Phänomene durch interessierte Laien (Citizen Science) gipfelte. Aufbauend auf den Erfahrungen mit SDSS entsteht derzeit in den USA unter Federführung der Johns Hopkins Universität, Prof. Alex Szalay, ein Big Data Open Storage Network [3], das die logisch zentrale Datenverwaltung für einen Großteil der öffentlich geförderten Forschungsprojekte übernehmen soll.

Bezüglich wirtschaftlicher Anwendungen ist zu befürchten, dass die meisten großen deutschen

Unternehmen den Wert einer nationalen Daten- und Analyseinfrastruktur nicht sofort erkennen werden. Ich würde hier (leider) im Vergleich zum angelsächsischen und nordeuropäischen Raum eine langsamere Adoption erwarten, analog zu der verspäteten Adoption von Internet, Big Data, KI-Methoden, Cloud, etc.

Es besteht jedoch riesiges Potenzial für eine erste Nutzung bei „neuen“ Unternehmen, also Start-ups und jungen Unternehmen, bei denen Daten den wesentlichen Teil des Kerngeschäfts darstellen bzw. massive Wettbewerbsvorteile liefern, z. B. bei Marktforschungsdaten bzw. im Bereich Mobilität, Logistik, Handel. Diese könnten durch Skaleneffekte schneller am Markt aktiv werden, ohne eine Daten- und Analyseinfrastruktur selbst aufzubauen und unterhalten zu müssen. Die Erwartung ist, dass sich aufgrund von Start-ups durch schnelle Umsetzung in diesen Bereichen durch Positivbeispiele die Innovationszyklen und Technologieadoption auch in den langsameren großen Industrieunternehmen verkürzen. Ferner, dass die Technologien auch in den Mittelstand wirken, aufgrund einer geringeren Eintrittsschwelle bei der Nutzung einer Daten- und Analyseinfrastruktur. Außerdem würde eine derartige Infrastruktur die Wiederverwendung von Algorithmen und Daten in Wirtschaft und Wissenschaft fördern und somit die Kosten für neue Erkenntnisse bzw. die Etablierung von neuen Geschäftsmodellen senken. Gleichzeitig ist ein großer gesamtgesellschaftlicher Nutzen im Hinblick auf Data Literacy für die Bevölkerung insgesamt zu erwarten, da Schüler auf diese Weise, z. B. bei der Vorbereitung von Referaten oder Hausarbeiten, spielerisch an Programmierung, Datenanalyse und sogar an potenzielle Geschäftsmodelle oder Citizen Science herangeführt werden könnten, indem sie Apps zusätzlich zur Daten- und Analyseinfrastruktur entwickeln.

### **Anforderungen an eine solche Daten- und Analyseinfrastruktur**

Vom Vorbild kommerzieller Infrastrukturen mit Fokus auf Datenverarbeitung können folgende konkrete Anforderungen für eine derartige Infrastruktur abgeleitet werden:

1. Geringe Einstiegshürde für die Erstellung und Nutzung von Datenanalysen, Visualisierungen, Modellen durch deklarative Sprachen, in de-

- nen lediglich gewünschte Ergebnis beschrieben werden und nicht die Algorithmen zu dessen berechnen, gepaart mit visuellen Tools.
2. Ubiquitärer Zugang zur Datenanalyseinfrastruktur durch einen webbrowsersbasierten Zugang durch Mobiltelefonen, Laptops, etc. ohne die Anforderung, Software installieren zu müssen.
  3. Nutzung und Wiederverwendung des Wissens der Community durch Verfügbarmachen der Extraktions-, Integrations-, Analyse- und Visualisierungsmethoden in einem Repository, mit Bewertungen und Nutzungsbeispielen von Daten und Verfahren.
  4. Zugriff auf Datenströme in Echtzeit und die Möglichkeit, kontinuierliche Analyse auf einem oder mehreren Datenströmen gepaart mit anderen Datenquellen zu realisieren.
  5. Verantwortungsvolles Datenmanagement durch Sicherstellung von Daten(analyse)schutz durch Zweckbindung, Nachvollziehbarkeit und Reproduzierbarkeit der Analyseergebnisse.

Daraus leiten sich diverse Anforderungen an das technische System ab:

1. Benutzerfreundlichkeit, Funktionalität und Kosteneffizienz (deklarative Programmierschnittstellen und automatische Optimierung, modularer Aufbau).
2. Breite Verfügbarkeit von Quellen mit guter Qualität der Daten und Datenströme (Open Data, Datenspenden, Möglichkeit der Speicherung von Daten mit Zugriffsrestriktionen, Bereinigung, Kuratierung, Integration).
3. Skalierbarkeit des Systems bezüglich Datenmenge, Datenvielfalt und hoher Rate des Zugangs neuer Daten, Nutzerzahl und Analysekomplexität (durch Aufbau einer Rechnerarchitektur basierend auf moderner Hardware),
4. Community-Funktionen zum Teilen und Bewerten von Daten, Analysemethoden, Visualisierungen, durch Open-Source-Entwicklung, dokumentierte Open-Source-APIs, Offenheit.
5. Vertrauenswürdigkeit, Daten(analyse)schutz (Sicherstellung der Zweckbindung der Analysen), Zugriffsrechte, Sicherheit, Transparenz, Nachvollziehbarkeit, Reproduzierbarkeit.
6. Unterstützung des gesamten Datenlebenszyklus sowie der vollständigen, oft iterativen Verarbeitungsketten des maschinellen Lernens von Quellauswahl über Datenaufbereitung über Ana-

- lyse zur Modellanwendung und Visualisierung durch verschiedene, einander überführbare Programmierabstraktionen, von visuellen Tools über deklarativen Sprachen zu imperativen Code der Algorithmen, welcher parallelisiert, verteilt und auf der Datenanalyseinfrastruktur optimiert zur Ausführung gebracht wird
7. Aufbau einer logisch zentralen, massiv-parallelen, physisch verteilten, multi-tenancy Hardwareinfrastruktur zur Speicherung und kontinuierlichen Erweiterung der Daten sowie Verarbeitung der Datenströme und Analysealgorithmen durch intelligentes Caching mit Speicherung und Wiedernutzung bereits durchgeführter Analysen im Rahmen von neuen Analysen

Der Erfolg einer Daten- und Analyseinfrastruktur hängt vor allem von der Funktionalität, der Benutzerfreundlichkeit bei der Erstellung von Analysen und dem Datenangebot ab. Solange ausländische, kommerzielle Anbieter vergleichbare Funktionalität anbieten können und gleichzeitig dieselben Daten zur Verfügung stellen, wird eine neue Plattform nicht angenommen werden.

Um eine Nutzbarkeit von verfügbaren Daten zu ermöglichen ist es wichtig, Daten und Ströme nicht nur in ihrer Rohform zur Verfügung zu stellen, sondern auch Metadaten zu verwalten und sogar automatisch abzuleiten und kontinuierlich zu verfeinern, zum Beispiel durch Informationsextraktion (z. B. Erkennung von Ereignissen aus Textdatenquellen) oder fortgeschrittene Analyse (z. B. Erkennung von zeitabhängigen Mustern bei der Auslastung von Verkehrsmitteln). Durch die Integration und Abbildung von verschiedenen Datenquellen auf dieser Plattform können individuelle Datensätze nach Bedarf angereichert und erweitert werden. Dazu gehören auch Verfahren zur rechtssicheren Anonymisierung und Pseudonymisierung personenbezogener Daten, wie sie beispielsweise bei medizinischen Analysen gefordert sind.

Auf dieser Infrastruktur könnten dann kontinuierlich wachsende Datenmengen sowie Qualitätssicherungs- und Analyseverfahren entwickelt werden, die domänenspezifische Anwendungen unterstützen und durch eine wachsende Anzahl an einfach zu nutzenden, von einer Open-Source-Community entwickelten und auf der Infrastruktur bereitgestellten Data-Science-

Werkzeugen weiterverarbeitet werden. Eine wichtige Datenquelle stellen kontinuierliche Datenströme in Echtzeit dar, die es möglich machen, neue Erkenntnisse zum Zeitpunkt der Datenerhebung zu gewinnen. Die Entwicklung deklarativer Konzepte, die es auch Menschen ohne ausgeprägte Informatikkenntnisse ermöglichen sollen, Extraktions-, Integrations-, Analyse-, Stromverarbeitungs- und Informationsvisualisierungsmethoden in einem Baukastenprinzip anzuwenden, ist der Schlüssel zur breiten Akzeptanz und damit letztlich zur erfolgreichen Implementierung nachhaltiger digitaler Informationsinfrastrukturen.

Die rasante Entwicklung der Hardwarelandschaft in den letzten zehn Jahren erweitert hierbei die Datenverarbeitungsmöglichkeiten, wichtige Technologien sind hierbei Multi-Core-CPU, GPUs sowie spezielle Beschleuniger wie Xeon-Phi-Coprozessoren, FPGAs und TPUs. Die heutige Hardwarelandschaft erfordert, dass Daten- und Analyseinfrastrukturen verschiedene Prozessoren und folglich verschiedene Programmiermodelle unterstützen, um die Leistungsanforderungen von Benutzern zu erfüllen. Eine große Herausforderung in diesem Bereich ist eine gemeinsame Abstraktion für Daten- und Analyseinfrastrukturen, mit der Nutzer die Algorithmen mit heterogenen Prozessoren und Beschleunigern ohne großen Aufwand zur Ausführung bringen können. Weiterhin ist eine Unterstützung neuer Speichertechnologien wie SSDs und NVRAM (nicht flüchtiger Hauptspeicher) erforderlich, um auf große Datenmengen schnell zugreifen zu können und echtzeitnahe Analyse zu ermöglichen.

Eine erfolgreiche Daten- und Analyseinfrastruktur wird dem Benutzer alle Stufen der Datenverarbeitungskette des maschinellen Lernens anbieten und optimieren, dazu zählen Datenerhebung, Datenvorverarbeitung, Feature Extraction, Modellbildung und Training, Modellinstallation und -update, sowie die Inferenz. Die einfache Durchführung von Analysen im Webbrowser und die Verhinderung von Abhängigkeiten durch communitybasierte Open-Source-Entwicklung und Bereitstellung von Daten und Verarbeitungsketten sollte dabei ein entscheidender Vorteil einer öffentlichen, deutschen oder europäischen Lösung sein.

Die hier skizzierte Daten- und Analyseinfrastruktur kann auf eine Vielzahl an grundlegenden

Vorarbeiten aufbauen, sowohl im Bereich der Forschungsdateninfrastrukturen [6, 9, 10], der Wirtschaft (z. B. dem International Data Space [4]) und der Vorarbeiten im Kompetenzzentrum ScaDS und BBDC [2, 11]. Eine derartige Infrastruktur unterscheidet sich jedoch von klassischen Höchstleistungsrechnern in Bezug auf Fokus von effizienter Datenübertragung und Datendurchsatz im Gegensatz zu Höchstleistungsprozessoren sowie in Bezug auf Software- und Hardware-Co-Design mit Fokus auf Datenmanagement und Analyse sowie Management von Analysealgorithmen und der damit verbundenen Entwickler- und Nutzer-Community. Ferner unterscheidet sich eine derartige Infrastruktur auch von Open-Data-Lösungen, die häufig im Bereich der Forschungsdateninfrastrukturen bzw. für öffentliche Daten geschaffen werden bzw. dem Industrial Data Space, da sie die gesamte Datenwertschöpfungskette und die damit verbundenen Werkzeuge und Programmiermodelle sowie effiziente, ggf. parallele und verteilte Ausführung der Datenanalyseprogramme beinhaltet. Ferner unterscheidet sich eine derartige Infrastruktur auch von klassischen Datenbanksystemen, da die Analysen durch fortgeschrittene Methoden der mathematischen Programmierung, der Signalverarbeitung und des maschinellen Lernens durch Verwendung von Iterationen und komplexen benutzerdefinierten Funktionen über die Konzepte der relationalen Algebra hinausgehen. Von existierenden Vorhaben wie dem Industrial Data Space oder dem Smart Data Innovation Lab unterscheidet sich eine derartige Infrastruktur durch ihre generelle Ausrichtung und damit verbundene Skalierbarkeit und durch die Zielgruppe (Wirtschaft, Wissenschaft und Bürger insgesamt) sowie Zugriffsmöglichkeit (webbasiert, interaktiv).

Dabei ist anzumerken, dass aufgrund der verteilten Erzeugung der Daten eine zentrale Datenspeicherung aller anfallenden Rohdaten aus Datenströmen, z. B. alle Sensordaten von selbstfahrenden Autos mit derzeitigen Technologien, nicht sinnvoll ist und somit bei sehr großen, schnellen Datenquellen ggf. eine Vorverarbeitung in der Edge oder im Fog stattfinden wird. Hierbei gilt der Satz des ehemaligen CTOs von Teradata, Todd Walther: „Data is as elastic as brick wall“, d. h. die Daten können nicht jederzeit ad hoc übertragen werden. In einer derartigen Architektur sollten Metadaten, Modelle sowie Algorithmen und weitere durch Ana-

# { EINE NATIONALE DATEN- UND ANALYSEINFRASTRUKTUR

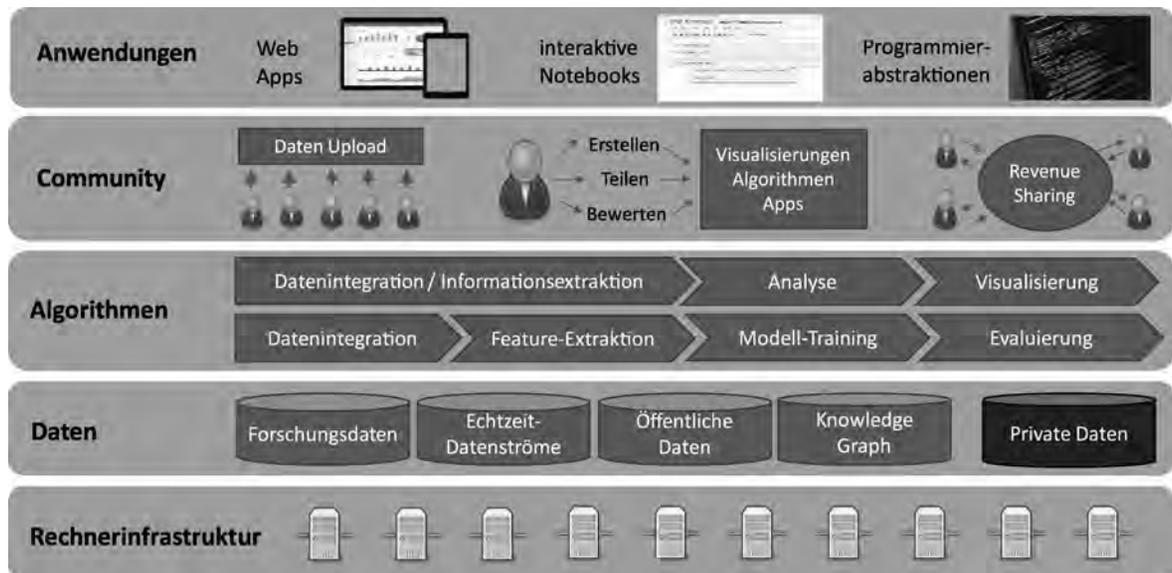


Abb. 1 Wesentliche Aspekte der Daten- und Analyseinfrastruktur

lysen abgeleitete Daten logisch zentral verwaltet werden, wobei Daten ggf. physisch verteilt vorliegen, und die Berechnung und Ableitung von Modellen sollte nahe an der Datenquelle durchgeführt werden. Bei großen, insbesondere kontinuierlichen Datenquellen kann dies durchaus gefördert, im Sinne des sogenannten „function shipping“ stattfinden, um „data shipping“ und Kommunikation nach Möglichkeit zu vermeiden bzw. zu reduzieren. Aus Latenzgründen sollten jedoch dabei je nach Netz- und Speicherkapazität so viele Daten „zentral“ bzw. nahe an der Analysedurchführung gespeichert werden, wie für Analysen nötig, ggf. mit Prefetching für avisierte Analysen von verteilten Daten und ggf. förderter Vorverarbeitung. Gleichzeitig sollte analog zu üblichen Data-Warehouse-Infrastrukturen der Drill-Through auf die Rohdaten möglich sein, falls erforderlich bzw. die (Vor-)Analyse der Rohdaten nahe der Datenquelle. Wichtig für die breite Nutzung der Infrastruktur von Wissenschaftlern, Unternehmen, in Schulen und von Bürgern insgesamt ist jedoch, dass die Infrastruktur die Verarbeitung von interaktiven Analysen in vielen Fällen ermöglicht, d. h. dass Analyseergebnisse und Visualisierungen für viele Klassen von Analysen im Rahmen einer „Denkzeit“/„Nahezu-Echtzeit“ produziert werden und Programme und Daten geteilt und wiederverwendet werden können. Dies erfordert das zentrale Caching/Vorhalten von bei Analysen benötigten Daten und die Verbindung von Computer

und Storage im Sinne einer gemeinsamen Architektur und des Software-/Hardware-Co-Designs, was im klassischen HPC-Bereich eher unüblich ist. Ferner muss die Infrastruktur alle Aspekte der Datenanalyse, von Informationsextraktion und -integration/Datenkuration über Analyse und Modellbildung bis hin zur Visualisierung und automatischem Feedback neuer, abgeleiteter Daten in die Infrastruktur beinhalten und all diese Funktionalität ohne Softwareinstallation beim Benutzer webbasiert bereitstellen.

Der Aufbau und nachhaltige Betrieb einer derartigen, in Abb. 1 skizzierten Daten- und Analyseinfrastruktur sollte als eine nationale Aufgabe im Rahmen der KI-Strategie der Bundesregierung angesehen werden, um die Rolle Deutschlands als starken Wirtschafts- und Wissenschaftsstandort in einer digitalisierten Welt nachhaltig zu sichern. Eine derartige Initiative könnte in Analogie zum Flugzeugbau als „IT-Airbus“ bezeichnet werden: Analog zum Flugzeugbau gilt es, den Rückstand gegenüber amerikanischen und asiatischen Unternehmen durch einen koordinierten und kraftvollen Anschlag des Staates aufzuholen. Gleichzeitig spielen Daten als Produktionsfaktor von nationalem Interesse eine Sonderrolle, so dass der Betrieb einer derartigen Infrastruktur als hoheitliche Aufgabe für die notwendige Neutralität und Vertrauen in Verfügbarkeit und Sicherheit sinnvoll sein könnte, insbesondere, um breiten Einsatz und Nutzung derselben

Infrastruktur in unterschiedlichen Branchen und durch potentiellen konkurrierende Unternehmen sicherzustellen. Die Realisierung wird konzertierte Beiträge aus nahezu allen Bereichen der Informatik in Forschung und Entwicklung erfordern und sollte ein kurz-, mittel- und langfristiges Ziel der Gesellschaft der Informatik sein. Im Rahmen der KI-Strategie der Bundesregierung ist eine derartige Infrastruktur vorgesehen [5].

### Über Prof. Volker Markl

Prof. Dr. Volker Markl leitet das Fachgebiet Datenbanksysteme und Informationsmanagement an der Technischen Universität Berlin und ist Chief Scientist und Leiter der Forschungsgruppe „Intelligente Analyse von Massendaten“ am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) in Berlin. Er ist Direktor des vom Bundesministerium für Forschung und Bildung eingerichteten Kompetenzzentrums „Berlin Big Data Center“ (BBDC) zum Umgang mit großen Datenmengen, Co-Direktor des Kompetenzzentrums „Berliner Zentrum für maschinelles Lernen“ (BZML) und Direktor des Smart Data Forums des Bundesministeriums für Wirtschaft und Energie.

Prof. Dr. Markl war vor seiner Tätigkeit in Berlin neun Jahre als Forscher und Innovator

bei IBM im Silicon Valley tätig. Parallel zu seiner wissenschaftlichen Laufbahn hat Prof. Dr. Volker Markl mehrere Start-ups begleitet bzw. mitgegründet, unter anderem Parstream (heute CISCO) und dataArtisans. Die dataArtisans sind aus dem Stratosphere-Projekt der TU Berlin hervorgegangen und haben die Datenstromverarbeitungstechnologie Apache Flink erfolgreich weiterentwickelt und kommerzialisiert. 2014 wurde Prof. Dr. Markl als einer der führenden „Digitalen Köpfe“ Deutschlands von der Gesellschaft für Informatik ausgezeichnet.

### Literatur

1. Apache Flink. <https://flink.apache.org/>
2. BBDC. <http://www.bbdc.berlin/home/>
3. Big Data Open Storage Network. <https://hub.jhu.edu/2018/06/07/big-data-open-storage-network-alex-szalay/>
4. International Data Space. <https://www.fraunhofer.de/de/forschung/fraunhofer-initiativen/industrial-data-space.html>
5. KI Strategie. <https://www.ki-strategie-deutschland.de/home.html>
6. NFDI Arbeitsgruppe. <https://www.akademienunion.de/arbeitsgruppen/ehumanities/nfdi-arbeitsgruppe/>
7. OpenML. <https://www.openml.org/>
8. Rapidminer. <https://rapidminer.com/>
9. Rat für Informationsinfrastrukturen. <http://www.rfii.de/de/themen/>
10. Rat für Wirtschafts- und Sozialdaten. <https://www.ratswd.de/forschungsdaten/fdi-ausschuss>
11. Scads. <https://www.scads.de/de/>
12. Sloan Digital Sky Survey. <http://www.sdss.org>
13. Hey T, Tansley S, Tolle K: The Fourth Paradigm. <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/>

# Vektorarchitektur

Was einen Vektorrechner ausmacht

Rudolf Fischer

## Einleitung

Zukünftige CPU-Generationen werden nicht mehr die Leistungssteigerung erbringen, die man bisher aufgrund des „Gesetzes“ von Gordon Moore („Moore’s Law“) auf Anwenderseite gewohnt war. Gleichzeitig steigt der Stromverbrauch wissenschaftlicher Großrechner und ein Ende dieser Entwicklung ist nicht absehbar. Das führt zu einer Stagnation, teilweise gar Reduktion der Taktfrequenzen der Prozessoren, denn die Leistungsaufnahme steigt überproportional mit der Taktfrequenz.

Ähnlich wie in den 1970er-Jahren, als der Stromverbrauch und die Kühlung der Systeme eine große Herausforderung darstellten und nicht zuletzt deshalb in den USA und in Japan Vektorrechner entwickelt wurden, hier sind besonders die Pioniere Seymour Cray<sup>1</sup> und Tadashi Watanabe<sup>2</sup> zu nennen, bedarf es heute wieder der Optimierung der Architektur von Prozessoren, um ein Maximum an Rechenleistung für die verfügbare Chip-Technologie bei gleichzeitig hoher Energieeffizienz zu erzielen. Deshalb ist in den letzten Jahren das Konzept der Vektorisierung wieder vermehrt realisiert worden.

## Datenparallelität

Vektorisierung beruht auf der Datenparallelität, welche in den meisten technisch-wissenschaftlichen Anwendungen vorliegt. Diese bringt zum Ausdruck, dass die gleiche Operation für viele unabhängig vorliegende Daten zur Erzeugung voneinander nicht abhängender Resultate durchgeführt werden muss (Abb. 1).

<sup>1</sup> [https://de.wikipedia.org/wiki/Seymour\\_Cray](https://de.wikipedia.org/wiki/Seymour_Cray).

<sup>2</sup> [https://de.wikipedia.org/wiki/Tadashi\\_Watanabe](https://de.wikipedia.org/wiki/Tadashi_Watanabe).

```
do i = 1, N
  y(i) = x(i) + y(i)
end do
```

Abb. 1 Datenparallele Schleife in FORTRAN

Eine solche einfache Struktur mag artifiziell erscheinen, ist sie aber nicht. Viele wissenschaftliche Programme beruhen auf einer mathematischen Beschreibung eines Problems durch diskretisierte partielle Differenzialgleichungen, d. h. die physikalischen Zusammenhänge werden auf allen Gitterpunkten individuell berechnet, eventuell unter Einbeziehung einer beschränkten Umgebung, und ein neuer Zustand der Variablen wird anschließend auf dem Gitter erzeugt. Letztendlich zeigt sich hier die Lokalität der Naturgesetze.

Als Resultat erhält man deshalb Schleifen, bei denen die Reihenfolge der Schleifendurchläufe mit unterschiedlichem „i“ wie in obigem Beispiel irrelevant ist: Einzelne Iterationen hängen nicht voneinander ab, jede Permutation der Indizes führt zu demselben Ergebnis. Das ist „Datenparallelität“.

## Pipelining

Die Berechnung beispielsweise einer Multiplikation ist nicht trivial in einem Takt zu bewerkstelligen.

<https://doi.org/10.1007/s00287-018-01139-w>  
© Springer-Verlag Berlin Heidelberg 2018

Rudolf Fischer  
NEC Deutschland GmbH,  
Fritz-Vomfelde-Straße 14, 40547 Düsseldorf  
E-Mail: [Rudolf.Fischer@emea.nec.com](mailto:Rudolf.Fischer@emea.nec.com)

Alle „Aktuellen Schlagwörter“ seit 1988 finden Sie unter:  
<http://www.is.informatik.uni-wuerzburg.de/as>

Stattdessen bedarf es mehrerer Schritte, sogenannter Segmente, die hintereinander durchzuführen sind. Die Anzahl der Segmente ist dabei abhängig von der Implementation (Abb. 2).

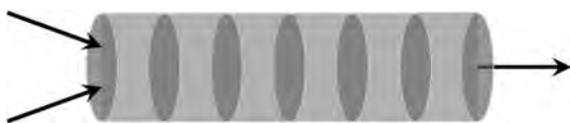


Abb. 2 Schematische Darstellung einer Recheneinheit

In einem Diagramm, beim welchem die horizontale Koordinate den Prozessortakten entspricht, die vertikale dem Segment in einer Pipeline, sieht eine Verarbeitung ohne Pipelining unter der Annahme, dass äußere Schleifeniterationen präpariert sind und nicht zu zusätzlichen Takten führen, wie in Abb. 3 dargestellt aus.

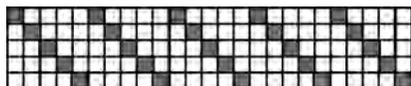


Abb. 3 Zeitlicher Verlauf der Nutzung der Segmente

Wenn verschiedene Operationen vorliegen, das ist auch innerhalb einer Schleife möglich, kann man von der Parallelität in den Segmenten Gebrauch machen, z. B. etwas komplexer mit einer Mischung von Addition und Multiplikationen wie im Beispiel der Abb. 4.

```
do i = 1, N
  y(i) = a * x(i) + b * y(i)
end do
```

Abb. 4 Etwas komplexere Schleife mit besserer Nutzung der Pipelines

Die beiden Multiplikationen können also, wenn alle Daten in Registern verfügbar sind, innerhalb eines Schleifendurchlaufs überlappen, während die Addition, die normalerweise in einem separaten Rechenwerk abgearbeitet wird, auf die Ergebnisse der Multiplikationen warten muss.

Jedes weiße Quadrat in der Abb. 5 stellt ungenutzte Rechenleistung und damit Energieverbrauch ohne Mehrwert dar. Offenbar ist diese Form der Abarbeitung der Schleife also nicht optimal.

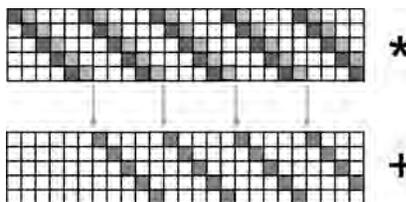


Abb. 5 Das Zeit-Segment-Diagramm mit besserer Nutzung der Pipelines

### Fließbandverarbeitung

Naheliegender kann man nun versuchen, unterschiedliche Schleifeniterationen so in die Pipeline zu leiten, dass die Segmente besser genutzt werden, ähnlich zur Fließbandverarbeitung in der Automobilindustrie. Schon seit vielen Jahren nutzt die Compiler-Technologie deshalb das sogenannte „Loop-Unrolling“. Für die obige Schleife wird unter der vereinfachenden Annahme, dass „N“ eine gerade Zahl ist, folgende Transformation des Codes durchgeführt, ein „Unrolling der Tiefe 2“ (Abb. 6).

```
do i = 1, N, 2
  y(i) = a * x(i) + b * y(i)
  y(i+1) = a * x(i+1) + b * y(i+1)
end do
```

Abb. 6 Dieselbe Schleife mit „Unrolling der Tiefe 2“

Schematisch ergibt sich folgende Nutzung der Rechenwerke (Abb. 7).

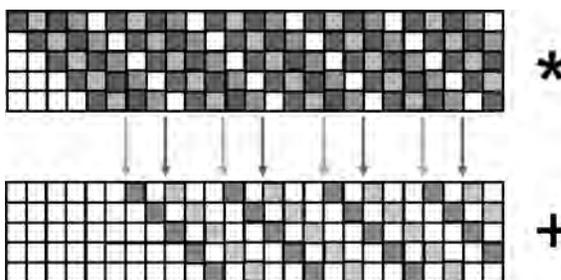


Abb. 7 Verbesserte Auslastung der Rechenwerke

Leider wird diese vergleichsweise gute Auslastung realistisch nicht zu erreichen sein. Erstens bedarf es zusätzlicher Operationen für jeden Schleifendurchlauf, um überhaupt die Rechenwerke nutzen zu können, und zweitens werden meist nicht genügend Daten in den Registern verfügbar sein. Speziell die Bandbreite zwischen Speicher und Pro-

zessor ist seit Jahrzehnten das dominierende Thema im Supercomputing.

## SIMD

Um eine schnellere Rechenleistung und bessere Energieeffizienz zu erzielen, kann man nun eine Hardwarearchitektur ebenfalls inhärent parallel entwerfen und entsprechende Instruktionen generieren, welche mehrere Schleifendurchläufe auf einmal abbilden und durch parallele Rechenwerke die notwendigen Operationen gleichzeitig abarbeiten. Natürlich benötigt dieses Vorgehen auch Register, die so viele Daten gleichzeitig enthalten können, wie es der Parallelität der Rechenwerke entspricht (Abb. 8).

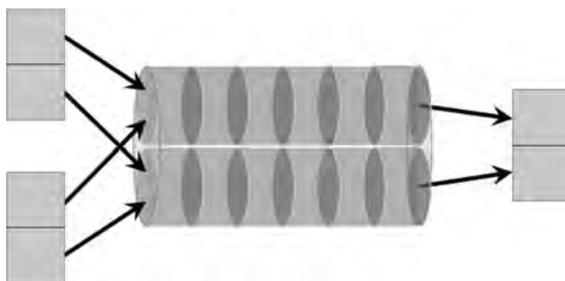


Abb. 8 Zweifach parallele Pipeline mit entsprechenden Registern

Der Vorteil liegt in der Tatsache, dass viele zusätzliche Operationen wie beispielsweise die Verwaltung von Adressregistern, Aktualisierung von Iterationszählern etc. nur einmal für mehrere Rechenoperationen anfallen. Zusätzlich müssen für jede Rechenoperation die Datenpfade des Prozessors so konfiguriert werden, dass die Inputdaten von Registern zur Recheneinheit und die Resultate in Ergebnisregister geleitet werden. Diese Maßnahmen kosten Zeit und Energie. Wenn die eigentliche Berechnung also parallel für mehrere Schleifeniterationen durch eine Instruktion durchgeführt werden kann, so reduziert sich der Aufwand pro Iteration.

Naheliegender nennt man diese Parallelisierung SIMD („single instruction multiple data“). Die Multiplizität der parallelen Operationen in der Hardware wird auch als „SIMD-Breite“ bezeichnet. Diese liegt bei den heutigen Mikroprozessoren typisch im Bereich zwischen zwei und acht Operationen, bei bekannten Grafikprozessoren im Bereich von 32, bei heutigen Vektorrechnern bei 256.

## Prinzip des Vektorrechners

Der entscheidende Unterschied zwischen SIMD-Instruktionen skalarer Architekturen und einem wirklichen Vektorrechner ist das sogenannte Vektorregister, eine Hardware, welche über eine bestimmte Anzahl von Prozessortakten ein Datum, oder je nach Architektur parallel eine gleichbleibende Menge von Daten, pro Takt an die Rechenwerke senden bzw. von diesen empfangen kann. Die Analogie zur Fertigungsstraße in der Automobilindustrie drängt sich einmal mehr auf. Im Vergleich dazu enthalten SIMD-Register skalarer Architekturen eine bestimmte geringe Anzahl von Daten, können diese aber nur zu einem Zeitpunkt aufnehmen oder zur Verfügung stellen. Die Auslastung der Rechenwerke wird durch die Vektorisierung verbessert (Abb. 9).



Abb. 9 Zeit-Segment-Diagramm bei Nutzung eines Vektorregisters

In diesem Beispiel beträgt die Länge des Vektorregisters 32. Die zusätzlich auftretenden, unvermeidbar zu jeder Schleife gehörenden und oben erwähnten zusätzlichen Operationen sind parallel während der Aktivität des Rechenwerkes leicht zu bewältigen. Dies ist ein nicht immer beachteter Vorteil einer Vektorarchitektur: Es gibt hinreichend Zeit bzw. Prozessortakte, bestimmte Dinge zu verstecken, sozusagen die „Administration“ der Schleife, während die gewünschten Fließkommaoperationen ausgeführt werden.

Moderne Vektorsysteme beherrschen natürlich auch das dynamische Umordnen von Operationen abhängig von der Verfügbarkeit der Ressourcen wie Register und Rechenwerke. Diese besondere Eigenschaft wurde in der Vergangenheit besonders vom japanischen Entwickler Tadashi Watanabe in Vektorrechnern implementiert. Ebenso ist es durchaus möglich, schon im Zielregister verfügbare Resultate für weitere Berechnungen an eine andere Recheneinheit weiterzuleiten, bevor dieses Zielregister vollständig von der laufenden Vektorinstruktion gefüllt worden ist, das sogenannte „Chaining“, welches von Steve Chen<sup>3</sup> für die Cray X-MP erstmalig implementiert wurde.

<sup>3</sup> [https://en.wikipedia.org/wiki/Steve\\_Chen\\_\(computer\\_engineer\)](https://en.wikipedia.org/wiki/Steve_Chen_(computer_engineer)).

Wie in dem Zeitschema deutlich zu erkennen ist, führt die „Latenz“ der Pipeline von in diesem Beispiel sechs Takten noch zu ungenutzten Segmenten. Auch dieses kann vermieden werden, indem weitere Vektorbefehle schon abgesetzt werden, zum Beispiel im gegebenen Fall der nächste Schleifendurchgang von 32 Iterationen, bevor die vorherige Vektoroperation beendet ist.

Die beiden Methoden der Hardwareunterstützung für Datenparallelität können kombiniert werden, indem Vektorregister mehr als ein Datum pro Takt zur Verfügung stellen oder aufnehmen können, dieses über eine bestimmte Anzahl von Prozessortakten und durch eine einzige Instruktion gesteuert, und die Rechenwerke entsprechend mehrfach ausgelegt werden, ähnlich wie bei den heutigen skalaren Architekturen.

Um eine datenparallele Schleife auf eine Vektorarchitektur abbilden zu können, wird die Schleife in Stücke der Länge des Vektorregisters VL zerlegt, man nennt das „Stripmining“ (Abb. 10).

```
do i_ = 1, N, VL
  do i = i_, min( i_+VL-1, N )
    y(i) = a * x(i) + b * y(i)
  end do
end do
```

Abb. 10 Schematische Darstellung der vektoriellen Ausführung in FORTRAN

Dabei wird der kursiv notierte Teil des Codes auf einzelne Vektorbefehle abgebildet, die inneren Schleifeniterationen finden nur in der Hardware statt. Während die Vektoroperationen durchgeführt werden, kann die CPU den nächsten Durchgang vorbereiten. Hier wird der Vorteil der Vektorisierung erneut deutlich: Es sind weniger Instruktionen notwendig, und diese können sehr gut zeitlich überlappen. Wenn man auf einer Vektormaschine ein Programm optimiert, dann wird die Rechenleistung gemessen in Fließkommaoperationen pro Sekunde verbessert, gleichzeitig nimmt aber mit zunehmender Vektorisierung die Anzahl der Instruktionen pro Sekunde ab!

### Vergleiche

Dieses Faktum adressiert teilweise den oft gehörten Vorwurf, dass ein Vektorprozessor auf skalarem Code vergleichsweise schwach agiert. Zunächst ist

das richtig, die Steuerung des Prozessors muss zu jedem Zeitpunkt alle Ressourcen überwachen, deren Verfügbarkeit überprüfen und entsprechend umordnen, und das eben nicht nur für die ebenfalls vorhandenen skalaren Register und Pipelines. Der Aufwand dazu ist groß und der „skalare Teil“ eines Vektorprozessors ist kein für skalaren Code optimierter Prozessor. Das Rechenwerk der kommerziell verfügbaren Vektorarchitektur, der NEC SX-Aurora TSUBASA, ist 32-fach parallel. Die für eine vergleichbare Rechenleistung notwendigen Alternativen wären eine skalare CPU mit extrem hohem Takt, welcher mit heutiger Technologie nicht realisiert werden kann, oder ein Grafikprozessor, der ein ähnliches Maß an Parallelität realisiert, dabei aber besondere Programmierparadigmen benötigt.

### Code-Strukturen

Der Nutzer muss also seinen Code vektorisieren. Das sollte er aber heutzutage für eine effiziente Nutzung anderer Architekturen ebenso. Dabei muss die Vektorlänge der Schleifen die Vektorregister der Länge 256 auch zumindest einigermaßen ausnutzen, man braucht also im Vergleich zu den heute vorrangig vertretenen „skalaren Architekturen mit SIMD-Instruktionen“ – ein Oxymoron übrigens – längere Schleifen. Das führt unmittelbar zu der Frage, ob die wissenschaftliche Problemstellung eine entsprechende Dimensionierung erlaubt.

Heutige Problemstellungen im technisch-wissenschaftlichen Bereich sind mehrdimensional und schlicht groß, d. h. die zugrundeliegenden Gitter enthalten viele Gitterpunkte oder es werden viele finite Elemente benutzt. Zunächst ist damit die Voraussetzung für eine effiziente Vektorisierung, die Existenz langer vektorisierbarer Schleifen, strukturell vorhanden. Viele moderne Codes enthalten allerdings kurze innere Schleifen in Schleifennestern, z. B. für Diskretisierungsverfahren hoher Ordnung. Jedoch sind äußere Schleifen, die sich über ganze Gitter oder eine große Anzahl finiter Elemente oder Volumen erstrecken, dann ebenfalls existent.

Die heutigen typischen Codestrukturen spiegeln oft auch eine Cache-Optimierung wider. Die kurzen Schleifen, welche Nachbarschaftsbeziehungen ausdrücken, ermöglichen eine gute Cache-Nutzung. Diese kann man aber auch anders erzielen und zunehmend ist es ohnehin notwendig, zum Zwecke optimaler Leistung für einen Vektorrechner notwen-

dige Umstrukturierungen auch für heutige skalare Architekturen in Betracht zu ziehen.

Konzeptionell ist dies die bekannte Thematik, ob man ein „Array of Structures“ bearbeitet oder den dem hardwarenahen SIMD-Gedanken mehr entsprechenden Weg der „Structure of Arrays“ verfolgt. Selbst in Optimierungskursen für skalare Architekturen wird letzteres empfohlen.

- Skalares Denken: Es gibt einen Gitterpunkt, finites Volumen oder Element, ein Teilchen. Welche Operation ist darauf anzuwenden?
- Vektordenken: Es gibt eine auszuführende Operation. Auf welche Gitterpunkte, auf welches finite Volumen, Element oder Teilchen kann diese Operation unabhängig angewandt werden?

Auf dem Vektordenken basieren letztlich auch zahlreiche neuere Initiativen, „Domain-Specific Languages“ zu entwickeln. Dabei kommt ja gerade zum Ausdruck, dass man z. B. den Gradienten eines physikalischen Feldes ausrechnen will und eben nicht den Gradienten an einem Gitterpunkt. Es sind die Gitter und die auf diesen definierten Variablen, die die diskretisierte Abstraktion der in der mathematischen Beschreibung auftretenden physikalischen Felder darstellen. Der Gitterpunkt selbst

und der auf diesem vorliegende Wert einer Variablen hat in diesem Sinne keine Bedeutung.

In vielen Fällen können moderne Compiler die fast immer vorhandenen, durch „moderne Programmieretechniken“ aber oft nicht leicht erkennbaren äußeren Schleifen für eine Vektorisierung nutzen. Die Codestruktur kann also für eine gute Vektorisierung angepasst werden. Das setzt aber manchmal vertiefte Kenntnisse des Codes und der Verfahren voraus. Anwender haben diese Kenntnisse immer, Computerhersteller selten und nur eingeschränkt.

### Schlussfolgerung

Aufgrund der derzeitigen Stagnation der Leistungsfähigkeit einzelner Rechenkerne gängiger CPU-Architekturen ergibt sich die Notwendigkeit, jegliche in der Problemstellung vorhandene Parallelität effizient zu nutzen, speziell auch die feingranulare Datenparallelität. Vektorisierung als Programmierparadigma und Vektorarchitekturen als eine Hardwarerealisation gewinnen deshalb zunehmend an Bedeutung. Ein Programm durch Veränderung der Strukturen, auch der Datenstrukturen, effizient zu vektorisieren und so eine hervorragende Leistung der einzelnen Prozesse zu erreichen ist vergleichsweise einfach, wenngleich nicht immer ohne Aufwand zu erzielen.

## Gewissensbits – wie würden Sie urteilen?

Christina B. Class,  
Ernst-Abbe-Hochschule Jena  
Carsten Tirinitis,  
Technische Universität München

In den ethischen Leitlinien der GI steht: „Die GI initiiert und fördert interdisziplinäre Diskurse zu ethischen und sozialen Problemen der Informatik.“ Hierzu veröffentlichen Mitglieder der Fachgruppe „Informatik und Ethik“ der GI in dieser Kolumne jeweils einen hypothetischen, aber realistischen Fall zusammen mit einigen Fragen, die zur Diskussion anregen sollen. Die Fälle können jeweils von allen Interessierten im Blog auf der GI-Website <https://gewissensbits.gi.de> diskutiert werden.



### Fallbeispiel: Eine verlockende Perspektive

Erwin ist Professor im Bereich Gesichtserkennung/Bildverarbeitung. Er hat vor einem Jahr einen Ruf an der sehr renommierten University BigU angenommen. In verschiedenen Projekten verwenden er und seine Mitarbeiter Methoden des Deep Learning. Er war daher sehr froh, den Ruf zu erhalten, da seine neue Universität finanziell deutlich besser dasteht und er in den Berufungs-

verhandlungen für die ersten drei Jahre recht erfolgreich einige Mittel für Forschungsprojekte aushandeln konnte. Auch befinden sich zwei Forschungsinstitute in der Nähe der Hochschule, mit denen er bereits in der Vergangenheit erfolgreich zusammengearbeitet hat. Aber auch als neu berufener Professor merkt er den Druck, Drittmittel einzuwerben, da die Universität sich auch für die nächste Runde der Exzellenzinitiative bewerben möchte. Wie lange er seine gute Lehrstuhlausstattung behalten kann, sollte diese Bewerbung nicht erfolgreich sein, vermag Erwin nicht abzuschätzen.

Aktuell ist er mit seiner Gruppe dabei, Verfahren des Deep Learning zu verwenden, um Sentimentanalyse mit Gesichtserkennung zu verbinden. Sie hoffen, dadurch automatisch Krankheitsanzeichen erkennen zu können und so ein automatisches Frühwarnsystem für allein lebende ältere Menschen zu erstellen, um im Notfall rechtzeitig einen Arzt zu rufen. Dies soll analog zu den Sturzerkennungssystemen erfolgen, die jedoch sehr viel später einsetzen und keine Hilferufe absetzen, wenn jemand z. B. im Sitzen oder Liegen das Bewusstsein verliert. Die für diese Arbeit notwendigen Rechenkapazitäten sind ziemlich groß und sie stecken noch in den Anfängen. Aktuell können sie zwar ansatzweise ein Unwohlsein erkennen, allerdings haben sie noch keinen Ansatzpunkt, um zwischen physischem Unwohlsein (Krankheitsanzeichen) und negativen Emotionen (z. B. genervt sein) zu unterscheiden.

Wie andere Hochschulen auch, ist BigU dabei, in Fernost mit der lokalen Universität East University eine gemeinsame Universität mit ausgewählten Studiengängen zu gründen.

445 *Gewissensbits – wie würden Sie urteilen?*

447 *Deep Learning und Rechtsrisiken*

449 *Die Individualisierung des Nutzen*

450 *Rezension*

451 *Zum Titelbild*

<https://doi.org/10.1007/s00287-018-01137-y>

Die deutschen Studienpläne sollen dort umgesetzt werden, begabte Studierende Stipendien für ein Masterstudium in Deutschland erhalten und einige ausgewählte Studierende zum Promovieren nach Deutschland eingeladen werden. Dieses Projekt soll die wirtschaftliche und wissenschaftliche Kooperation stärken und findet daher auch Unterstützung aus der Politik.

Im Bereich der Rechnerarchitektur hat East University einige beachtliche Referenzen aufzuweisen und gemeinsam mit Industriepartnern einen Superrechner in Betrieb, der in der Lage ist, extrem hohe Datenmengen in kürzester Zeit zu verarbeiten. Mit Blick auf seine Projekte und die dafür notwendigen Rechenleistungen interessiert sich Erwin sehr für das Projekt und ist entsprechend engagiert.

Daher ist er auch Teil eines Projektteams, mit dem Ziel, einen gemeinsamen Masterstudienplan in Informatik und eine Strategie für eine Forschungsk Kooperation zu erstellen. Mit einigen Vertretern seiner Hochschule reist er für zwei Wochen in das Land. Sie sind alle begeistert von der East University und den Gesprächen mit Kollegen. Insbesondere der Superrechner beeindruckt Erwin. Als mögliche gemeinsame Projekte im Bereich der Informatik diskutiert werden, finden seine aktuellen Arbeiten großen Anklang. Die lokalen Kollegen sind sehr an einer gemeinsamen Fortführung seiner Arbeit interessiert und stellen ihm für das Projekt praktisch unbegrenzten Zugang zu ihrem Superrechner in Aussicht. Erwin ist begeistert. Damit könnten sie einen großen Sprung nach vorne erreichen.

Nach dem offiziellen Programm hängt er mit zwei Kollegen noch drei Tage an, um die Stadt und die Gegend gemeinsam mit zwei lokalen Kollegen

selber zu erkunden. Zunehmend merkt er dabei, dass dort ganz andere Vorstellungen bez. Privatsphäre herrschen als zuhause in Deutschland. Nun, da sie Zeit haben, sich etwas umzusehen, beobachtet er den umfassenden Einsatz von Überwachungstechnologien im Alltag. So langsam fühlt er sich selber etwas überwacht und merkt auch, wie er etwas vorsichtig(er) wird in seinem Verhalten und mit seinen Meinungsäußerungen anderen gegenüber.

Wieder in Deutschland angekommen, liegt auf dem Mensatisch das Faltblatt „Unsere ethischen Leitlinien“ der Gesellschaft für Informatik. Er steckt es ein und liest es auf dem Heimweg in der U-Bahn.

Erwin ist zunächst alles andere als begeistert von den ethischen Leitlinien. Er hat den Eindruck, dass die Autoren sehr technikkritisch sind und sich irgendwie gegen den Fortschritt sperren. Ob das wirklich Informatiker sind, die daran mitgearbeitet haben? Die hätten bestimmt etwas gegen seine Projekte ... Er schüttelt den Kopf und steckt die Leitlinien, ohne groß weiter darüber nachzudenken, in seine Jackentasche.

Am Abend trifft Erwin einige seiner Freunde im Restaurant. Erwin erzählt begeistert von seiner Reise nach Fernost und den Möglichkeiten, die sich ihm eröffnen. Er kommt richtig ins Schwärmen, als er von dem Besuch im Rechenzentrum berichtet. Hans, ein Jugendfreund, der an einem Gymnasium Deutsch und Politik unterrichtet, fragt sehr interessiert nach und will genaue Schilderungen von Erwins Eindrücken des Alltags. Er ist etwas skeptisch, vor allen Dingen da er viel über Überwachungen gelesen hat. Frank, ein Projektleiter bei einem Energieversorger, wirft ein, er wäre da etwas vorsichtig. Man habe doch genug von Industriespionage gehört, und sie in ihrem Unternehmen

würden daher aus Asien keine Praktikanten einstellen. Erwin berichtet von der offenen Atmosphäre, die er erlebt hat, im Laufe des Gesprächs erinnert er sich aber auch an das etwas unbehagliche Gefühl, das sich bei ihm einschlich, da er sich auf der Straße überwacht fühlte. Marius, ein Kollege, der schon seit 2 Jahren vor Ort lebt, hat ihm sogar erzählt, dass er gelernt habe, immer neutral zu blicken und weder Zustimmung noch Langeweile noch Ablehnung in Besprechungen nonverbal zum Ausdruck zu bringen, da man da doch teilweise schnell darauf angesprochen wird, vor allen Dingen auch von Vorgesetzten ... Erwin bestellt eine neue Runde für alle und sie wechseln das Thema und haben noch einen lustigen Abend.

Als Erwin später im Bett liegt und kurz vor dem Einschlafen ist, kommen ihm einige Gesprächsfetzen wieder in den Sinn. Er beginnt zu grübeln. Was könnte man mit seiner Arbeit alles anstellen? Plötzlich fallen ihm die Leitlinien wieder ein. Gab es da nicht irgendwas zu Kontrolle und Überwachung? Da er jetzt wieder richtig wach geworden ist, steht er auf, geht zu seiner Jacke und liest die Leitlinien noch einmal durch. Er bleibt an Artikel 11 hängen und plötzlich drängen sich ihm Fragen auf ...

Artikel 11 der GI Leitlinien (Version 2018) „Das GI-Mitglied wirkt darauf hin, die von IT-Systemen Betroffenen an der Gestaltung dieser Systeme und deren Nutzungsbedingungen angemessen zu beteiligen. Dies gilt insbesondere für Systeme, die zur Beeinflussung, Kontrolle und Überwachung der Betroffenen verwendet werden können.“ (<https://gi.de/ueber-uns/organisation/unsere-ethischen-leitlinien/>)

### Fragen

- Ein Ziel der Arbeit von Erwins Gruppe ist die Erkennung des phy-

sischen Unwohlseins bei Menschen, um z. B. bei älteren Menschen schneller reagieren und Hilfe holen zu können. In einer alternden Gesellschaft, in der nicht alle Menschen dauerhaft betreut werden können, ist dies eine sehr sinnvolle Anwendung. Doch dieser Nutzen schließt die Beobachtung der Menschen ein. Inwieweit müssen zukünftige Nutzer, also die Patienten in die Entwicklung eingebunden werden? Muss/sollte verlangt werden, dass z. B. Patienten in Heimen der Nutzung solcher Systeme zustimmen? Wie kann eine solche Zustimmung erfolgen? Was ist zu bedenken, wenn Menschen sich der Tragweite ihrer Zustimmung auf Grund mangelnder Sachkenntnis oder Erkrankung (z. B. Demenz) nicht bewusst sind? Was kann man tun, wenn manche Patienten zustimmen und manche nicht? Kann man „zum Wohle“ des Patienten einfach eine Zustimmung voraussetzen? Welche Probleme und Gefahren ergeben sich hierbei?

- Erwin stellt sich die Frage, ob die von seiner Gruppe durchgeführte Forschung in einem anderen Land zu Kontrolle und Überwachung eingesetzt werden kann und möglicherweise wird. Es handelt sich um eine Vermutung, er hat keine Beweise dafür, und es geht um den möglichen Einsatz im Fernen Osten.

Bezieht sich Artikel 11 der GI Leitlinien auch auf einen solchen Fall? Mögliche/zukünftige Betroffene von IT-Systemen in anderen Ländern können schlecht in die Entwicklung von Systemen in Deutschland einbezogen werden. Inwieweit muss man die Betroffenen bei der Entwicklung gedanklich einbeziehen?

- Forschung kostet Geld und viele Hochschulen und Institute sind zunehmend auf den Erwerb von Drittmitteln angewiesen, um Forschungsprojekte durchführen zu können. Wie stark machen sich die Forschenden dann von Interessen außerhalb abhängig? Wie unabhängig kann Forschung noch sein? Können Leitlinien Forschenden den Rücken stärken, um gewisse Projekte nicht weiterzuverfolgen? Welche Unterstützung bräuchten Forschende und Hochschulen, um die wirtschaftliche Freiheit zu haben, fragwürdige Projekte nicht weiterzuverfolgen.
- Forschende stehen unter Druck, Forschungsergebnisse zu veröffentlichen. Die Veröffentlichungen sind weltweit zugänglich. Ist es möglich, den Missbrauch der Ergebnisse zu verhindern? Ist es möglich, die Nutzung der eigenen Arbeit und den Missbrauch, z. B. zur Überwachung und Kontrolle, immer vorherzusehen? Soll/muss sich ein Forschender hierzu Gedanken machen? Oder bezieht sich der entsprechende Artikel

der Leitlinien „nur“ auf diejenigen, die Systeme effektiv entwickeln und nicht auf Forschende?

- Was ist die Rolle von ethischen Leitlinien? Was kann man von ethischen Leitlinien erwarten? Ist es realistisch, dass sie zum Nachdenken anregen? Und können/sollen ethische Leitlinien in Konfliktfällen Handlungsempfehlungen oder gar Handlungsanweisungen geben?
- Wäre Erwin nach dem Gespräch mit den Freunden nicht auch ins Grübeln gekommen, wenn er die Leitlinien nicht zwischen die Finger bekommen hätte?
- Ist es sinnvoll und notwendig, ethische Leitlinien zu haben, die den aktuellen Stand der Technik ansprechen? Wie kann dies sichergestellt werden in Bereichen, die sich sehr stark und schnell ändern? Wie oft sollen/müssen ethische Leitlinien angepasst/überarbeitet werden?
- Sollen und müssen ethische Leitlinien neben Fachexperten, in diesem Fall dann Informatikern, auch Gesellschaftswissenschaftler als Autoren haben?

Die Fachgruppe ist unter <http://www.fg-ie.gi.de/> erreichbar. Unser Buch „Gewissensbisse – Ethische Probleme der Informatik. Biometrie – Datenschutz – geistiges Eigentum“ war im Oktober 2009 im Transkript-Verlag erschienen. Ein neues Werk ist in Arbeit.

## Deep Learning und Rechtsrisiken

Ursula Sury

### Mehrere Beteiligte

Deep Learning ist eine Anwendung oder auch Weiterentwicklung von künstlicher Intelligenz. Es geht darum, dass die Software Daten in-

telligent miteinander verknüpft und daraus Rückschlüsse gezogen werden. Die Software ist so angelegt, dass sie sich selber weiterentwickelt. Dies sowohl durch die Verbreiterung der Datenbasis als auch durch die Varianz der Verknüpfungen. Der

Computer lernt also auch durch Erfahrung.

Wenn man die Entstehung einer Deep Learning Anwendung analysiert, so stellt man fest, dass es mehrere Beteiligte gibt, die dazu beitragen. Zuerst sind es einmal die Softwareentwickler, dann aber auch die jeweiligen Fachexperten,

die die Datenbasis und bekannte Kombinationen/Reaktionsmuster zur Verfügung stellen. Aber auch der (End)Anwender selber trägt zur Weiterentwicklung mit der Art seiner Verwendung bei.

Die Verantwortung für sinnvolle, brauchbare und somit ungefährliche Resultate liegt somit bei mehreren Parteien. Softwareersteller, Experten und Anwender.

Falls das Deep Learning Produkt als Produkt im Sinne des Produkthaftungsgesetzes angeschaut würde (was bei NurSoftware kaum der Fall ist), haben wir also verschiedene Produzenten und bei anderer Form der Haftung, mehrere gemeinsam Haftpflichtige.

## Nachvollziehbarkeit

Wenn solche Deep Learning Produkte in der Praxis eingesetzt werden, zum Beispiel bei der Erkennung von Krebszellen, so ist es für den Anwender nicht nachvollziehbar, warum der Computer auf welche Resultate kommt. Das ist eine gewisse Schwierigkeit, aber nicht ganz neu. Wenn ich den vorliegenden Artikel schreibe, weiss ich auch nicht, was im Hintergrund im Betriebssystem und im Word technisch alles abläuft, aber ich verwende es wie Millionen andere Menschen trotzdem. Dasselbe haben wir beim Einsatz heutiger Haushaltsmaschinen, wie Staubsaugerroboter, Softwaregesteuerte, Backöfen, intelligenten Haussteuerungen etc.

Der Unterschied liegt aber in der Dimension der Nichtnachvollziehbarkeit. Beim Backofen gibt es viele Maschinenbauingenieure, die

genau erklären können, warum es funktioniert, weil das System klarere Grenzen hat und nicht Inputs aus verschiedenen Quellen auf verschiedene Arten kombiniert und in diesen zwei Dimensionen „eigenständig“ neue Kombinationen schafft.

Die Nachvollziehbarkeit der Resultate rückblickend (wie ist denn das gegangen) ist bei Deep Learning Produkten schwierig.

## Arbeitsteilung Mensch-Maschine

Der Einsatz von Deep Learning Produkten verlangt vom Menschen, dass er die entstandenen Resultate trotz der fehlenden Nachvollziehbarkeit aufgrund seiner Expertise und seiner Intuition (genau das fehlt nämlich der Maschine sic!) auf deren Brauchbarkeit einschätzen und überprüfen kann.

Ziel des Einsatzes ist es ja weiter zu kommen, als ohne Maschine, und die Vorteile der Maschine mit den Aktivitäten des Menschen zu ergänzen. Es ist ein grosser Vorteil, dass ich einen Roboter habe, der mir den Küchenboden nass putzt, aber ich kann und muss ihn zu einem gewissen Grad überwachen, denn er könnte die Virtual Wall durchbrechen und auf dem Parkett Schaden anrichten. Zudem muss ich die Ränder noch von Hand putzen.

Wo und wie es sinnvoll ist, die Maschine einzusetzen und wo nicht, ist eine menschliche Entscheidung. Diese muss sorgfältig getroffen werden.

## Test, Riskmanagement und Vollzugsprobleme

Der unsachgemässe Einsatz oder der Einsatz von fehlerhaften Deep

Learning Produkten kann grosse Schäden generieren. Dies muss unter allen Umständen vermieden werden. Deshalb müssen die Anwender entsprechend geschult sein, aber auch die Produkte immer wieder getestet werden. Eine gute Idee ist auch immer ein Framing, d. h. anzugeben, in welchem Kontext und mit welchem Know-how das Produkt angewendet werden kann, und wo eben nicht.

Sollte dann einmal tatsächlich ein Problem auftauchen, dann ist es in so komplexen und vergemeinschafteten Projekten schwierig den Schadenverursacher beweisbar auffindig zu machen. Und selbst wenn dies möglich wäre, ist es in der heutigen internationalen Welt schwierig, die Rechtsansprüche effizient und effektiv durchzusetzen.

Aus diesem Grund wird zum Framing unbedingt ein realistisches Business Continuity Management vorzukehren sein.

## Fazit

- Bei Deep Learning sind immer mehrere Personen beteiligt.
- Die Ergebnisse des Systems sind nur schwer nachvollziehbar.
- Man muss sich der Risiken von Deep Learning bewusst sein und sollte Vorkehrungen treffen, um eine einigermaßen sichere Anwendung garantieren zu können.

*Ursula Sury ist selbständige Rechtsanwältin in Luzern, Zug und Zürich (CH) und Vizedirektorin an der Hochschule Luzern – Informatik. Sie ist zudem Dozentin für Informatikrecht, Datenschutzrecht und Digitalisierungsrecht.*

## **Einsichten eines Informatikers von geringem Verstande**

### **Die Individualisierung des Nutzen**

*Reinhard Wilhelm*

Wir alle haben in den vergangenen Monaten schmerzlich erfahren, welche Mühsal und Pein der Vollzug der EU-Datenschutzverordnung über uns gebracht hat. Unerwartet viele Organisationen und Unternehmen haben sich fürsorglich bei uns gemeldet und uns mitgeteilt, dass sie wie schon bisher gern weiterhin zu unserem Nutzen unsere Daten erfassen und nutzen möchten. Google etwa verarbeitet unsere Daten, um uns nützlichere und personalisierte Inhalte bereitzustellen, um die Qualität seiner Dienste zu verbessern und neue Dienste zu entwickeln, Werbung auf der Grundlage unserer Interessen einzublenden, und die Sicherheit und den Schutz vor Betrug und Missbrauch zu verbessern. Das ist doch alles mehr als reizend von Google! Man stelle sich vor, Google müsste uns, weil es unsere Vorlieben

nicht intensiv genug erforscht hat, ungezielt mit Werbung voll müllen! Unvorstellbar! Der reine Horror! Wir würden wahrscheinlich auf der Stelle werbeblind oder werbeallergisch werden und könnten gar nicht mehr von Werbung profitieren. Da loben wir uns doch die dezente, mengenmäßig überschaubare und vor allen Dingen personalisierte Werbung. Personalisierte Werbung ist doch wie personalisierte Medizin, auf das Individuum zugeschnitten. Hilft einfach viel besser! Auf die Idee, dass wir eventuell gar keine Werbung sehen möchten, ist bei Google sicher noch keiner gekommen.

Die Fluglinien benutzen derzeit schon erfolgreich Algorithmen, um die Ticketpreise individuell zu gestalten. Hat sich ein Kunde beim Portal einer Fluglinie nach Verbindungen erkundigt, aber keine Verbindung gebucht, so bekommt er beim nächsten Besuch des Portals wegen seines früher dokumentierten Interesses einen auf ihn zugeschnittenen, höheren Ticketpreis angeboten.

In anderen Bereichen ist die Individualisierung leider noch

nicht so weit fortgeschritten, wie es der Einsatz von Rechnern ermöglichen würde. Die Tankstellenpreise richten sich bei ihrer Veränderung über den Tag hinweg im Wesentlichen danach, ob der treibstoffhungrige Teil der Bevölkerung gerade arbeitet oder auf dem Wege von der Arbeit nach Hause ist. Der Google-Algorithmus zur Entdeckung von Staus, der die massenhafte Ansammlung von Handys auf Straßen beobachtet, könnte doch auch benutzt werden, den Andrang an oder den Zustrom auf eine Tankstelle festzustellen, um daraufhin die Kraftstoffpreise zu erhöhen. Allerdings wäre das immer noch nicht wirklich individuell. Dieses Ziel wäre erst erreicht, wenn der Autobesitzer, angelockt durch einen günstigen Treibstoffpreis, auf eine Tankstelle zusteuert, sein Näherkommen durch Sensoren, Kameras oder ähnliches entdeckt würde und eine Software daraufhin den Treibstoffpreis ganz individuell für diesen Kunden erhöhen würde.

Wir gehen wahrlich paradiesischen, personalisierten Zeiten entgegen!

## Rezension

### Galileo Galilei kontrovers



Walter Hehl:  
**Galileo Galilei kontrovers – Ein Wissenschaftler zwischen Renaissance-Genie und Despot.**  
Springer Vieweg, Wiesbaden 2017, 302 S., 49,99 €  
ISBN 978-3-658-19294-5

Das Buch des Physikers Walter Hehl bietet dem Leser die Möglichkeit, sich mit Galileis Gesamtwerk und seiner Wirkung zu beschäftigen. Es geht um die Auseinandersetzung zwischen der historisch im 20. Jahrhundert aufgearbeiteten Figur des Galilei – gehandelt und gepriesen als Entdecker neuen Wissens und Vater der modernen Physik oder sogar der modernen wissenschaftlichen Methode – und dem nüchterneren, historisch informierten Blick eines Physikers

auf die Errungenschaften einer seiner Vorreiter.

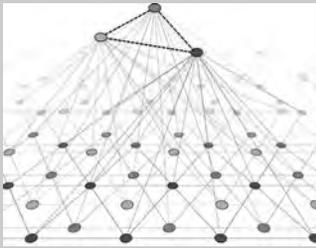
Es handelt sich um eine intellektuelle Biographie Galileis, die die Themen allerdings mehr nach inhaltlichen Kriterien zusammensetzt als nach ihrer Chronologie. Dadurch ist ein neues Buch entstanden, das es schafft, sich von der Tradition der Wissenschaftsgeschichte zu entfernen und gleichzeitig eine interessante Perspektive dafür anzubieten. Walter Hehl untersucht Galileis Mechanik, Optik, Astronomie und Kosmologie, vernachlässigt aber auch nicht die Arbeit des italienischen Physikers in Rahmen der Astrologie. Das Buch schließt mit einer erneuten Diskussion des Verhältnisses zwischen Galilei, seiner Wissenschaft und seinen Kommunikationsstrategien einerseits und der Reaktion der damaligen katholischen Kirche andererseits, die auf die Publikation seines *Dialogo* folgte.

Hehl bietet so oft wie möglich den aktuellen Forschungsstand im klaren Ausdruck und zeigt gleichzeitig eine tiefe historische Sensibilität

wenn es darum geht, die historische Quelle in ihrem Kontext zu analysieren. Besonders brilliant aus dieser Perspektive ist die Rekonstruktion der Forschung Simon Mayrs – des Gegners Galileis – in Bezug auf die Entdeckung der Jupitermonde.

Das Buch von Walter Hehl kann eine schwierige Funktion erfüllen, die viele Bücher in dem *mare magnum* der Literatur zu Galilei nicht haben absolvieren können: die disziplinäre Überbrückung zwischen Geistes- und Naturwissenschaften. Während Historiker sich hier mit dem modernen Blick der Physik bekannt machen können, können Physiker eine Einführung in die historische und, noch viel mehr, historiographische Arbeit erleben. Das Buch scheint in der Tat besonders geeignet für Studierende des Faches Physik, für die eine historische Perspektive auf ihr Fach seit langer Zeit ein Desideratum ist. Vielleicht ist mit diesem Buch der richtige Lesestoff dafür angekommen.

Prof. Dr. Matteo Valleriani



### Simulation von Gitterzellen als Spezialfall eines allgemeinen neuronalen Verarbeitungsprinzips

Die Abbildung zeigt eine schematische Darstellung zweier Ebenen eines Recursive Growing Neural Gas (RGNG), mit dem am Lehrgebiet Mensch-Computer-Interaktion der FernUniversität in Hagen das Verhalten von Gitterzellen (engl. grid cells) simuliert wird. Die drei in der oberen Ebene dargestellten Knoten repräsentieren drei Gitterzellen, während die untere Ebene Prozesse innerhalb der jeweiligen (gleichfarbigen) Gitterzelle repräsentiert.

Im Jahr 2014 erhielten May-Britt und Edvard Moser den Nobelpreis für Physiologie für ihre Entdeckung sogenannter Gitterzellen im entorhinalen Kortex von Säugetieren. Gitterzellen sind Nervenzellen, deren Aktivität mit der absoluten Position des Tieres in seiner Umgebung korreliert und hierbei ein regelmäßiges, hexagonales Muster ausbildet. Aufgrund dieses Verhaltens werden sie als wesentlicher Bestandteil eines Orientierungs- und Navigationssystems angesehen. In der Öffentlichkeit werden sie daher häufig als das „GPS“ unseres Gehirns beschrieben. Neuere

Beobachtungen von Gitterzellen deuten jedoch darauf hin, dass sie möglicherweise eine weiter gefasste Funktion besitzen. Dieser Möglichkeit geht ein neues Berechnungsmodell zur Beschreibung von Gitterzellen nach [1]. Im Gegensatz zu bereits existierenden Modellen basiert es nicht auf der Annahme, dass das Verhalten von Gitterzellen explizit auf eine spezifische Aufgabe ausgerichtet ist. Es geht vielmehr davon aus, dass die Aktivität von Gitterzellen einen Spezialfall eines allgemeinen neuronalen Prinzips zur Verarbeitung von Informationen auf einem hohen Abstraktionsniveau im Kortex darstellt.

Grundlage dieses vorgeschlagenen Verarbeitungsprinzips ist die Hypothese, dass Neuronen in den tiefen Verarbeitungsschichten des Kortex bestrebt sind, ein prototypbasiertes Modell ihres Eingaberaums zu erlernen, während sie gleichzeitig in Konkurrenz mit benachbarten Neuronen stehen. Hierbei lernt jedes Neuron, eine begrenzte Anzahl an prototypischen Eingabemustern zu erkennen, die möglichst vollständig den Raum möglicher Eingabemuster abdecken. Im Falle der Gitterzellen spiegeln deren räumlich periodischen Aktivitätsfelder diese Form der Modellbildung wider. Durch Anwendung des vorgeschlagenen Modells auf echte Bewegungsdaten von Ratten, die von der Arbeitsgruppe des Ehepaars Moser aufgezeichnet wurden, war es möglich, die typischen Aktivierungsmuster natürlicher Gitterzellen zu reproduzieren [2].

Da das vorgeschlagene Berechnungsmodell mit dem Anspruch

entwickelt wurde, nicht nur navigationsbezogene Prozesse erklären zu können, ist die Entdeckung von Neuronen im entorhinalen Kortex von Affen [6] interessant, die in Abhängigkeit von sakkadischen Augenbewegungen des Tieres ebenfalls ein gitterzellartiges Antwortverhalten aufweisen. Auch hier ist das vorgeschlagene Berechnungsmodell, jetzt angewendet auf den Eingaberaum der Motoneuronen-Signale zur Steuerung von sakkadischen Augenbewegungen, in der Lage, das gitterzellartige Antwortmuster zu reproduzieren [3].

Ein weiterer Aspekt des vorgeschlagenen Modells besteht in seiner intrinsischen Robustheit gegenüber Störungen in den Eingabemustern der Zellen, wie sie in einem neurobiologischen System zu erwarten sind [4]. Nachfolgend wurde untersucht, wie eine Störungskompensation in das Berechnungsmodell integriert werden kann. Dabei wurden drei Parameter variiert: die Intensität der Störungen im Eingangssignal, die Kapazität jeder Zelle, sich an die Stärke der eigenen Aktivierung in Reaktion auf frühere Eingaben erinnern zu können sowie die Dauer, bis eine derartige Erinnerung verblasst. Die Simulationsläufe ergaben hierbei, dass die Erinnerungsdauer den stärksten Einfluss auf die Störungskompensation hat, während eine Erhöhung der Erinnerungskapazität nur zu graduellen Verbesserungen führt [5].

Die Arbeiten dieses Themengebiets wurden mehrfach auf internationalen Konferenzen ausge-

zeichnet. Für seine Dissertation erhielt Jochen Kerdels 2017 den Dissertationspreis der Fakultät für Mathematik und Informatik der FernUniversität.

Jochen Kerdels und Gabriele Peters

### Literatur

1. Kerdels J (2016) A Computational Model of Grid Cells based on a Recursive Growing Neural Gas, Disser-

tation, Research Report 1/2016, FernUniversität in Hagen, Hagen, Germany

2. Kerdels J, Peters G (2013) A Computational Model of Grid Cells based on Dendritic Self-Organized Learning. In: Proceedings of the 5th International Conference on Neural Computation Theory and Applications (NCTA 2013). SciTePress, pp 420–429
3. Kerdels J, Peters G (2016) Modelling the Grid-like Encoding of Visual Space in Primates. In: Proceedings of the 8th International Conference on Neural Computation Theory and Applications (NCTA 2016). SciTePress, pp 33–41
4. Kerdels J, Peters G (2016) Noise Resilience of an RGNNG-based Grid Cell Model. In: Proceedings of the 8th International Conference on Neural Computation

Theory and Applications (NCTA 2016). SciTePress, pp 42–49

5. Kerdels J, Peters G (2019) A Noise Compensation Mechanism for an RGNNG-based Grid Cell Model. In: Computational Intelligence. Springer International Publishing, pp 263–276
6. Killian NJ, Jutras MJ, Buffalo EA (2012) A Map of Visual Space in the Primate Entorhinal Cortex. Nature 491(7426):761–764

Vorschläge für Titelbilder  
bitte an Prof. Deussen  
([Oliver.Deussen@uni-konstanz.de](mailto:Oliver.Deussen@uni-konstanz.de))



## Aus Vorstand und Präsidium

### Protokoll der Ordentlichen Mitgliederversammlung vom 25. September 2018 in Berlin

#### Tagesordnung

1. Begrüßung
2. Bericht des Vorstandes über das abgelaufene Geschäftsjahr
  - 2.1 Bericht des Präsidenten
  - 2.2 Jahresabrechnung
3. Bericht der Rechnungsprüfer für das Jahr 2017 und Entlastung von Vorstand, erweitertem Vorstand, Präsidium und Geschäftsführung
4. Wahl der Rechnungsprüfungskommission für das Jahr 2018
5. Entgegennahme des Haushaltsplanes 2019
6. Bestätigung der Mitgliedsbeiträge 2019
7. Bericht der Kandidatenfindungskommission und Feststellung der endgültigen Kandidatenliste für die Wahl 2018 (Präsidium)
8. Bestimmung der Kandidatenfindungskommission für die Wahl 2019 (Präsidiumsämter)
9. Wahl des Wahlausschusses für die Wahl 2018
10. Festlegung von Ort und Zeit für die Ordentliche Mitgliederversammlung 2020
11. Stellungnahme zu Anträgen auf Satzungsänderung
12. Genehmigung des Beschlussprotokolls der OMV 2018
13. Berichte
14. Verschiedenes

#### TOP 2.1 Bericht des Präsidenten

Herr Federrath berichtet über die Arbeit des vergangenen Jahres.

#### TOP 2.2 Jahresabrechnung

Herr Goedicke trägt die Jahresabrechnung vor.

#### TOP 3 Bericht der Rechnungsprüfer

Herr Meier trägt den Rechnungsprüfungsbericht für das Jahr 2017 vor und beantragt die Entlastung von Vorstand, erweitertem Vorstand, Präsidium und Geschäftsführung.

*Beschluss: Die Ordentliche Mitgliederversammlung entlastet Vorstand, erweiterten Vorstand, Präsidium und Geschäftsführung.*

#### TOP 4 Wahl der Rechnungsprüfungskommission für das Jahr 2018

Der Vorstand schlägt als Rechnungsprüfer für das Jahr 2018 vor: Prof. Dr. Michael Meier, Universität Bonn und Prof. Dr. Ali Sunyayev, KIT Karlsruhe

*Beschluss: Die Ordentliche Mitgliederversammlung setzt Prof. Meier und Prof. Sunyayev als Rechnungsprüfer für das Jahr 2018 ein.*

#### TOP 5 Entgegennahme des Haushaltsplanes 2019

Herr Goedicke trägt den Haushaltsplan 2019 vor.

#### TOP 6 Bestätigung der Mitgliedsbeiträge 2019

GI-Beitragsstruktur und Beitragshöhe 2019 (Präsidiumsbeschluss vom 29. Juni 2018)

#### Ordentliche Mitglieder

Vollmitglied	105,00 €
Doppelmitglied	79,00 €
Mitglieder assoziierter Organisationen	61,00 €
Vollmitglied im Ruhestand	52,50 €
Mitglieder mit rabattiertem Beitrag	21,50 €
Studierende im Bachelor/Master/Azubis	00,00 €

#### Korporative Mitglieder

Unternehmen	nach Umsatz
Studentische Gruppierungen	60,00 €

Der Sitzungsleiter bittet die OMV 2018, die vom Präsidium laut § 8.6.8 der Satzung beschlossene Beitragsstruktur und Beitragshöhe zu bestätigen.

*Beschluss: Die Ordentliche Mitgliederversammlung bestätigt die vom Präsidium beschlossene Beitragsstruktur und Beitragshöhe.*

#### TOP 7 Bericht der Kandidatenfindungskommission

Die von der OMV 2017 eingesetzte Kandidatenfindungskommission hat die folgende Liste für die Präsidiumswahl 2018 erstellt. Es sind drei Plätze zu vergeben.

- Prof. Dr. Dominik Herrmann
- Rolf Kickuth
- Karsten Merschjan
- Prof. Dr. Daniela Nicklas
- Dr. Jan Sürmeli
- Rainer Typke

Nach der Vorstellung der Kandidatinnen und Kandidaten wird die Liste nochmals eröffnet. Die OMV hat das Recht, weitere Personen vorzuschlagen.

*Beschluss: Die Ordentliche Mitgliederversammlung schlägt keine weiteren Personen vor. Damit ist obige Liste für die Präsidentswahl 2018 beschlossen.*

### **TOP 8 Bestimmung der Findungskommission für die Wahl 2019**

Das Präsidium benennt folgende Mitglieder der Kommission zur Findung von Präsidentskandidatinnen und -kandidaten:

Sprecher: Prof. Dr. Daniel Abawi (Sprecher)  
Beisitzer/ Prof. Dr. Christina Claß  
innen: Prof. Dr. Erhard Rahm  
Prof. Dr. Stefan Strecker

Die Ordentliche Mitgliederversammlung hat das Recht, zwei weitere Personen vorzuschlagen.

*Beschluss: Die Ordentliche Mitgliederversammlung schlägt keine weiteren Personen vor. Damit ist die obige Liste für die Kandidatenfindungskommission für die Präsidentswahl 2019 beschlossen.*

### **TOP 9 Wahl des Wahlausschusses für die Wahl 2018**

Der Vorstand schlägt folgende Damen und Herren vor:

Sprecher: Prof. Dr. Alexander Rossnagel, Kassel  
Stv. Prof. Dr. Rüdiger Grimm,  
Sprecher: Koblenz  
Beisitzer: Karl-Heinz Künkel, Bonn  
Dr. Michael Nüsken, Bonn  
Prof. Dr. Gerd Stumme, Kassel  
Viktor Schröder, Bonn  
Cornelia Winter, Bonn

*Beschluss: Die Ordentliche Mitgliederversammlung stimmt der Liste zu.*

### **TOP 10 Festlegung von Zeit und Ort für die OMV 2020**

Der Vorstand schlägt vor, die Ordentliche Mitgliederversammlung 2020 im Herbst in Karlsruhe abzuhalten.

*Beschluss: Die Ordentliche Mitgliederversammlung beschließt, die Ordentliche Mitgliederversammlung 2020 im Herbst in Karlsruhe abzuhalten.*

### **TOP 11 Stellungnahme zu Anträgen auf Satzungsänderung**

Es liegen keine Anträge auf Satzungsänderung vor.

### **TOP 12 Genehmigung des Beschlussprotokolls der OMV 2018**

Die Mitgliederversammlung wird gebeten, das Beschlussprotokoll der Sitzung vom 25. September 2018 in Berlin zu genehmigen.

*Beschluss: Die Mitgliederversammlung genehmigt das vorgelegte Beschlussprotokoll der OMV 2018.*

### **TOP 13 Berichte**

#### **TOP 14 Verschiedenes**

Berlin, im September 2018,  
Prof. Dr. Hannes Federrath (Sitzungsleitung),  
Cornelia Winter (Protokoll)

#### **Jahresbericht 2017–2018**

Der Jahresbericht des Präsidenten steht unter <https://gi.de/service/infomaterial/> im Web zum Herunterladen bereit.

#### **Zur finanziellen Lage der GI**

##### **1. September 2018**

#### **Zusammenfassung**

Das Jahr 2017 wurde mit einem Überschuss in Höhe von 301 T€ abgeschlossen, wobei etwa 100 T€ durch die Korrektur einer Fehlbuchung

aus dem Jahr 2016 verursacht sind. Dieses positive Ergebnis von ca. 200 T€ im Jahr 2017 ist für die GI das erste Jahr mit einem deutlich positiven Abschluss nach fünf Verlustjahren. Dieser Überschuss ist das Ergebnis verschiedener Maßnahmen. Insbesondere die Erhöhung der Mitgliedsbeiträge hat für eine deutliche Verbesserung der Einnahmenseite gesorgt. Die Mitgliederzahl ist 2017 wiederum gesunken, die Maßnahmen der vergangenen Jahre zur Mitgliedergewinnung zeigen weiterhin nur geringe Wirkung. Trotz der im Moment positiven finanziellen Lage der GI müssen auch weiterhin Maßnahmen und Investitionen zur Steigerung der Attraktivität der GI sowie zur Gewinnung neuer Mitglieder durchgeführt werden.

#### **Jahresabschluss 2017**

Im Haushaltsjahr 2017 haben die Gesamterträge von 3.055 T€ im Vorjahr um 349 T€ auf 3.404 T€ zugenommen. Gleichzeitig sind die Aufwendungen von 3.226 T€ im Vorjahr um 123 T€ auf 3.102 T€ zurückgegangen. Insgesamt ergibt sich somit für 2017 ein Überschuss von 301 T€ (nach Defiziten 2016: –171 T€, 2015: –235 T€, 2014: –327 T€ und 2013: –386 T€).

Die Zunahme der Erträge im Vergleich zu 2016 ist ganz wesentlich der moderaten Erhöhung der Mitgliedsbeiträge und weiteren positiven Effekten im Bereich der Tagungseinnahmen und Einsparungen im Betrieb zu verdanken. Die Einnahmen aus den Mitgliedsbeiträgen lagen um 199 T€ höher als 2016, Tagungserträge um 141 T€.

Abgenommen haben auf der Ertragsseite u. a. die Projektzuschüsse (25 T€). Wie bereits in den Jahren seit 2012 erfolgte auch 2017 keine Gewinnausschüttung durch die DLGI (nach einer letztmaligen Ausschüttung in Höhe von 203 T€ in 2011).

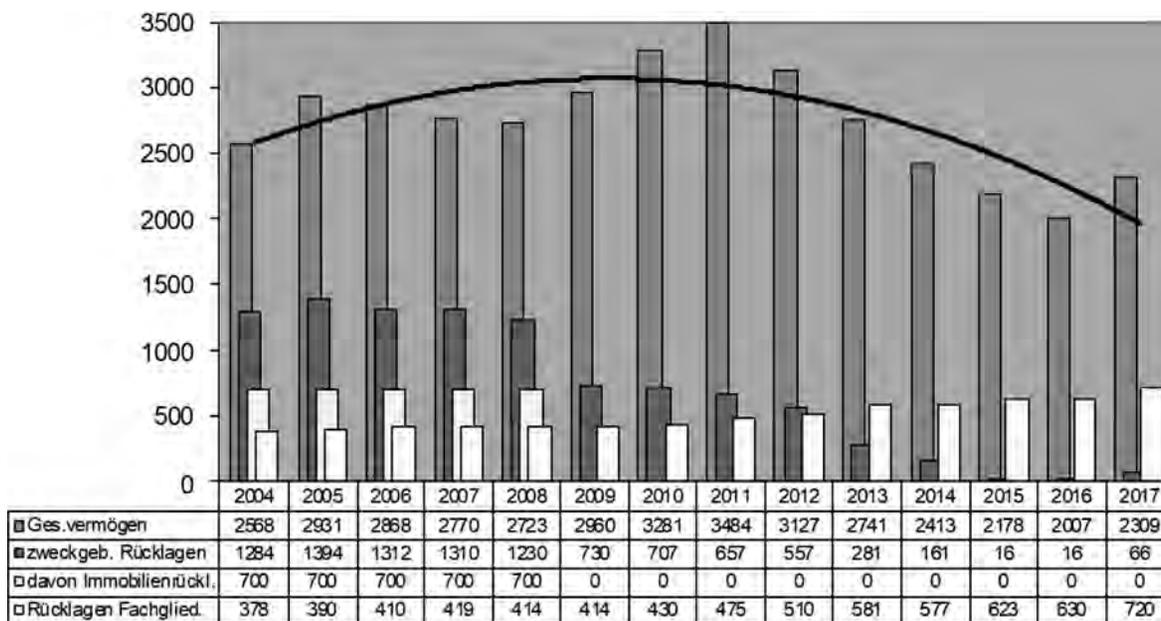


Abb. 1 Vermögensentwicklung 2004–2017

Auf der Aufwandsseite haben 2017 die Aufwendungen für Tagungen (83 T€) und die Personalkosten (81 T€) zugenommen. Die Reisekosten sind 2017 leicht (2,2 T€) gestiegen. Zurückgegangen sind die Aufwendungen für Publikationen (-53 T€), die Verwaltungskosten und sonstige Kosten (u. a. Portokosten, Druckkosten, IT -49 T€ bzw. -188 T€).

Das per Saldo im Jahr 2017 erzielte positive Tagungsergebnis in Höhe von 58 T€ liegt leicht höher als 2016 und ist durch erfolgreiche große Tagungen (u. a. Jahrestagung, Mensch und Computer) bewirkt, die hohe Überschüsse erwirtschaften konnten, jedoch auch größere Ausgaben mit sich brachten. Es gilt weiterhin, die Bemühungen zu verstärken, attraktive Tagungen für unsere Mitglieder auszurichten.

Die Mitgliederzahl ist 2017 von 16.453 auf 15.701 persönliche Mitglieder gesunken, dabei sind 182 beitragsfreie Mitgliedschaften mitgezählt. Die Zahl der korporativen Mitglieder ist um 12 auf 228 zurückgegangen. Die Steigerung der Attraktivität der GI, die Gewinnung von Neumitgliedern sowie die

Mitgliederbindung bleiben daher auch in Zukunft zentrale Herausforderungen. Das positive Ergebnis bei den Mitgliedsbeiträgen darf über diesen nach wie vor anhaltenden Trend nicht hinwegtäuschen. Dazu kommt, dass 2018 beobachtet werden muss, wie die weiteren Änderungen (Informatik Spektrum nur noch in elektronischer Form) sich auf die Mitgliederzahl auswirken. Weitere Maßnahmen (Steigerung der studentischen Mitglieder) sind in Planung bzw. Umsetzung begriffen.

### Vermögensentwicklung

Das Vereinsvermögen der GI ist durch den Gewinn 2017 gegenüber dem Vorjahr von 2.007 T€ auf 2.309 T€ gestiegen. Die Entwicklung des GI-Vermögens in den Jahren 2004 bis 2017 wird in Abb. 1 dargestellt.

Im Vereinsvermögen enthalten sind interne Rücklagen der Fachgliederungen, die auch 2017 leicht gestiegen sind. Es muss darauf geachtet werden, dass aus den Rücklagen auch Entnahmen für zeitnahe Aktivitäten der Fachgliederungen erfolgen.

Die zweckgebundenen Rücklagen wurden leicht erhöht (66 T€), um erwartete Arbeiten im Bereich der IT abzudecken, und enthalten 2017 unverändert 16 T€, die als Rücklage für den ICSI-Beirat bestimmt sind. Es wird jedoch erwartet, dass die Aktivitäten des ICSI-Beirats eingestellt werden, und diese Rücklage aufgelöst werden kann.

### Rechnungsprüfung 2018

Die Jahresrechnung und die zugrundeliegende Buchführung der Gesellschaft für Informatik für das Geschäftsjahr vom 1. Januar 2017 bis zum 31. Dezember 2017 wurden von Freudenhammer Maas & Partner mbB im Hinblick auf die Einhaltung der Rechenschaftslegungsgrundsätze für Vereine sowie satzungsmäßiger Regelungen geprüft und am 6. April 2018 bescheinigt.

Die Prüfung durch die von der Mitgliederversammlung bestellten internen Rechnungsprüfer Prof. Michael Meier und Prof. Ali Sunyaev fand am 18. Mai 2018 in der GI-Geschäftsstelle in Bonn statt. In ihrem Bericht bestätigen die Rechnungsprüfer die Ordnungsmä-





die Gruppe der Studierenden (s. o.) angenommen.

Das Präsidium hat in seiner Sitzung am 29. Juni 2018 in Bonn dem in Abb. 3 dargestellten Haushaltsentwurf 2019 zugestimmt. Er wird der Mitgliederversammlung 2018 zur Genehmigung vorgelegt. Als Ergebnis für das Jahr 2019 wird ein Verlust in Höhe von 50 T€ erwartet. Dabei ist ein geplanter Abbau von Vermögen der Fachgliederungen in Höhe von 38 T€ zur Finanzierung von Vereinsaktivitäten bereits berücksichtigt.

Abschließen möchte ich diesen Bericht zur finanziellen Lage der GI mit einem herzlichen Dank an die in der Geschäftsstelle für die Finanzen zuständigen Mitarbeiterinnen für ihre Unterstützung und die stets sehr angenehme Zusammenarbeit.

Essen, im September 2018



Prof. Dr. Michael Goedicke  
Universität Duisburg-Essen, Institut für Informatik und Wirtschaftsinformatik und  
paluno, The Ruhr Institute for Software Technology

## Personalia

### Frieder Nake mit Klaus-Tschira-Medaille ausgezeichnet

2018 haben die GI und die Klaus Tschira Stiftung zum ersten Mal im Gedenken an Klaus Tschira die Klaus-Tschira-Medaille verliehen. Mit dieser Medaille werden Personen geehrt, die durch ihre Arbeiten in Anwendungsgebieten unterschiedlicher Art (Wirtschaft, Gesellschaft, Politik,



Abb. 4 Frieder Nake (links) und GI-Präsident Hannes Federrath (rechts). (Foto: Cornelia Winter)

Kunst usw.) Anregungen zur Weiterentwicklung der Informatik und ihrer Methoden gegeben haben.

Mit Frieder Nake ehren die Gesellschaft für Informatik (GI) und die Klaus Tschira Stiftung (KTS) einen herausragenden Wissenschaftler, der die Wechselwirkungen zwischen der Informatik einerseits und künstlerischen und gesellschaftspolitischen Prozessen andererseits frühzeitig und nachhaltig erforscht und dokumentiert hat.

Prof. Dr. Hannes Federrath, Präsident der Gesellschaft für Informatik: „Frieder Nake wird weltweit als einer der Pioniere der Computerkunst geschätzt. Noch während seiner Studienzeit unternahm er die ersten Versuche zur Erstellung künstlerischer Zeichnungen mit Hilfe des von Konrad Zuse entwickelten computergesteuerten Zeichentisches ‚Graphomat‘. Zusammen mit anderen Pionieren dieses Gebietes organisierte er schon 1965 eine der weltweit ersten Ausstellungen der Computerkunst.“

Schon früh beklagte Frieder Nake, dass die Diskussion über die Nutzung des Computers in der Kunst in die falsche Richtung gehe. Bereits seit den frühen 1970er-Jahren beschäftigt er sich mit der Bedeutung des zentralen Konzeptes der Informatik, des Algorithmus, für den Umgang mit Information und deren Rückwirkungen auf die verschiedenen Aspekte des wissenschaftlichen und gesellschaftlichen Lebens. Frieder Nake ist ein Mittler zwischen den verschiedenen Wissenschaftsgebieten mit kritischer Haltung gegenüber dem unreflektierten Umgang mit der Informationstechnik.

Beate Spiegel, Geschäftsführerin der Klaus Tschira Stiftung: „Klaus Tschira war selbst Physiker, aber die Informatik war eine seiner großen Leidenschaften. Er hat sich ganz besonders für die Schnittstellen der unterschiedlichen Disziplinen interessiert. Die Verbindung von Informatik und Kunst, wie sie Frieder Nake betreibt, hätte Klaus Tschira sicher fasziniert.“

Weitere Informationen zur Klaus-Tschira-Medaille finden Sie unter: <https://gi.de/ueber-uns/personen/klaus-tschira-medaille/>.

### **Gesellschaft für Informatik ernennt Debora Weber-Wulff und Ralf Steinmetz zu Fellows**

Auf der INFORMATIK 2018 hat die GI Prof. Dr. Debora Weber-Wulff und Prof. Dr. Ralf Steinmetz zu Fellows ernannt. In seiner Laudatio würdigte GI-Präsident Prof. Dr. Hannes Federath die herausragenden Leistungen der beiden engagierten Mitglieder:

„Mit Debora Weber-Wulff ehrt die GI eine Hochschullehrerin, die zur internationalen Medieninformatik lehrt und forscht und sich mit großem Engagement für die Förderung von Frauen in der Informatik einsetzt. Darüber hinaus hat sie sich auf dem Gebiet der Plagiatserkennung und der Ethik in der Informatik sowohl in der Wissenschaft als auch in der breiten Öffentlichkeit einen Namen gemacht. In der GI hat Debora Weber-Wulff die Fachgruppe Ethik geleitet und die Neufassung der Ethischen Leitlinien aktiv mitgestaltet. Mit der Kolumne „Gewissensbits“ befördert sie darüber hinaus den Diskurs zu ethischen Fragen in der Informatik und hat dazu zahlreiche Praxisbeispiele für Lehre und Anwendung formuliert.

Prof. Dr. Ralf Steinmetz hat sich als Wissenschaftler und Hochschullehrer auf dem Gebiet der Rechnernetze, multimedialen Systeme und Anwendungen mit seiner Forschung zur Synchronisation multimedialer Daten und adaptiver Kommunikationssysteme national wie international verdient gemacht. Dabei hat er sich bei seiner Arbeit stets für den Brückenschlag zwischen der Informatik und anderen Disziplinen sowie den Technologietransfer aus der Wissenschaft in die Wirtschaft eingesetzt. In der GI hat er

sich darüber hinaus in verschiedenen Fachgruppen, bei Tagungen, der Herausgabe von Publikationen und in der Nachwuchsförderung engagiert.“

Weitere Informationen zu dem GI-Fellowship finden Sie unter: <https://gi.de/ueber-uns/personen/fellows/>.

### **Daniel Gruss und Amin Baumeler mit Dissertationspreis der Gesellschaft für Informatik ausgezeichnet**

Auf der INFORMATIK 2018 haben die Gesellschaft für Informatik e. V. (GI), die Österreichische Computer Gesellschaft (OCG) sowie die Schweizer Informatikgesellschaft (SI) den GI-Dissertationspreis an Dr. techn. Daniel Gruss (TU Graz) für seine Arbeit „Software-based Microarchitectural Attacks“ und Dr. Amin Baumeler (Österreichische Akademie der Wissenschaften) für seine Arbeit „Causal Loops: Logically Consistent Correlations, Time Travel, and Computation“ verliehen.

Dr. Ronald Bieber, Generalsekretär der OCG: „Daniel Gruss hat eine ingenieurwissenschaftliche Arbeit im Bereich der Mikroarchitekturangriffe verfasst, in der er zeigt, dass Angriffe vollständig automatisiert werden können, es neue Seitenkanäle gibt und Angriffe auch in stark eingeschränkten Umgebungen und auf jedem Computersystem durchgeführt werden können. Mit diesem Preis wird eine herausragende Arbeit mit hoher praktischer Relevanz auf dem Gebiet der Sicherheit moderner Computersysteme ausgezeichnet.“

Prof. Dr. Hannes Federrath, Präsident der GI: „Herr Baumeler hat eine theoretische Arbeit auf höchstem Niveau im Spannungsfeld zwischen Physik und Informatik vorgelegt. Die Dissertation beschäftigt sich etwa mit quantenmechanischen und relati-

vistischen Effekten und mit Fragen der Berechenbarkeit in selbstreferenzierenden Modellen. Gerade vor dem Hintergrund neuartiger Rechnerarchitekturen wie etwa dem Quantencomputer zeigt die Arbeit aber auch die Grenzen der Berechenbarkeit und leistet so einen wichtigen Beitrag zum Verständnis zukünftiger Informatiksysteme. Mit dem Preis wird eine herausragende theoretische Leistung gewürdigt, die viele unterschiedliche Gebiete miteinander verknüpft, ohne an der Oberfläche zu bleiben und ohne die traditionellen Grenzen der wissenschaftlichen Klassifikation zu beachten.“

Der Dissertationspreis zeichnet hervorragende Arbeiten aus, die einen Fortschritt für die Informatik oder andere Disziplinen bedeuten und die die Wechselwirkungen zwischen Informatik und Gesellschaft untersuchen.

### **GI-Junior-Fellows 2018 ernannt**

Die GI hat auf der INFORMATIK 2018 Prof. Dr. Simon Nestler (HS Hamm-Lippstadt), Alexander Steen (FU Berlin) und Prof. Dr. Andreas Vogel-sang (TU Berlin) zu GI Junior-Fellows ernannt.

Mit dem Junior-Fellowship-Programm fördert die Gesellschaft für Informatik e. V. (GI) seit 2013 jedes Jahr herausragende Jungtalente, die sich um die Informatik in Wissenschaft und Gesellschaft in besonderem Maße verdient gemacht haben. Prof. Dr. Hannes Federrath, Präsident der Gesellschaft für Informatik e. V. erklärt dazu: „Mit den Junior-Fellowships wollen wir exzellente Informatikerinnen und Informatiker ermutigen, zukunftsweisende Ideen für die Gestaltung der Informatik in allen gesellschaftlichen Bereichen zu entwickeln und umzusetzen. Aber auch wir als größte

Fachgesellschaft für Informatik im deutschsprachigen Raum profitieren von den neuen Impulsen, welche die Junior-Fellows durch den engen Austausch mit Vorstand und Präsidium setzen. Ich freue mich auf die Zusammenarbeit mit den drei nun ernannten Junior-Fellows.“

Prof. Dr. Andreas Vogelsang ist Juniorprofessor für Automotive Software Engineering an der Technischen Universität Berlin und leitet die Software Engineering Gruppe am Daimler Center for Automotive IT Innovations (DCAITI). Im Rahmen des Fellowships will er sich dafür einsetzen, Software besser erklärbar zu machen. Sein Ziel: Auch Menschen ohne Informatikhintergrund sollen besser verstehen, warum Softwaresysteme sich so verhalten, wie sie es tun.

Prof. Dr. Simon Nestler ist Professor für Mensch-Computer-Interaktion an der Hochschule Hamm-Lippstadt und forscht im Kontext sicherheitskritischer Systeme. Daneben engagiert er sich unter anderem beim Lippstädter Science Slam, dem HSHL Hackathon, dem mobilen Science-Labor zdi-Truck, der BMBF-Forschungsbörse und dem Gedankenblitz Wettbewerb sowie Jugend forscht. Als GI Junior-Fellow will er einen Beitrag dazu leisten, dass die Bandbreite und Vielfältigkeit der Informatik in der Gesellschaft noch stärker sichtbar wird und sich Schülerinnen und Schüler durch neue Formate für das vielfältige Themenspektrum der Informatik begeistern.

Alexander Steen schloss kürzlich seine Promotion an der Freien Universität Berlin im Bereich der formalen Logik ab. Nun arbeitet er an der Universität Luxemburg und entwickelt Methoden für die computergestützte Untersuchung von ethisch-normativen Modellen für intelligente autonome Systeme. 2016



Abb. 5 V. l. n. r. Christoph Rensing (Sprecher FG E-Learning), Preisträger Majd Edriss, Johannes Konert (stellv. Sprecher und Beauftragter für Nachwuchsförderung FG E-Learning). (Foto: Linda Rustemeier, Goethe-Universität FFM)

wurde er zusammen mit weiteren Kollegen für eine interdisziplinäre Lehrveranstaltung ausgezeichnet, die auf innovative Art und Weise Computersysteme für die formale Analyse von Argumenten der Metaphysik nutzt. Als GI Junior-Fellow möchte sich Alexander Steen verstärkt für die interdisziplinäre Verknüpfung der Informatik mit weiteren gesellschaftsrelevanten Anwendungsgebieten einsetzen.

## Aus den GI-Gliederungen

### Ausgezeichnete Unterstützung: Chatbot unterstützt Erstellung von Übungsaufgaben (B. Sc.) und Machine Learning Algorithmen helfen beim Programmieren lernen (M. Sc.)

Im Rahmen der DeLFI 2018 an der Goethe-Universität in Frank-

furt/Main wurden am 11. September 2018 die besten Abschlussarbeiten zum Thema E-Learning des Jahres 2017 ausgezeichnet. Die Auszeichnung für die beste Bachelorarbeit ging an Herrn Majd Edriss von der Beuth Hochschule Berlin für seine Arbeit mit dem Titel „Entwicklung eines Chatbot-Prototypen für die Erstellung von Übungsaufgaben im E-Learning“.

Die Unterstützung bei der Erstellung von passenden Lernressourcen ist insbesondere mit dem Thema *Open Educational Resources* (OER) wieder stark in den Fokus gerückt. Wie können Personen mit Fachwissen, bspw. im Handwerk, bei der Erstellung unterstützt werden? Herr Edriss hat dazu einen Chatbot entwickelt, welcher Anwendende eines Aufgabeneditors durch die Schritte bis zur fertigen Übungsaufgabe begleitet. Die Lösung basiert auf einem Drehbuch als Zielvorgabe für den Chatbot, welche Informationen im Dialog abzufragen sind. Besonders charmant ist die Vor-



**Abb. 6 V. l. n. r. Christoph Rensing (Sprecher FG E-Learning), Preisträger Alisan Öztürk, Johannes Konert (stellv. Sprecher und Beauftragter für Nachwuchsförderung FG E-Learning). (Foto: Linda Rustemeier, Goethe-Universität FFM)**

schau der Übungsaufgabe, welche der Chatbot automatisch aktualisiert und befüllt. Die Lösung wurde in den Online-Editor des Lernmanagementsystems „Smart Learning im Handwerk“ integriert und evaluiert. Als besondere Stärken der Arbeit nennen die Gutachten die detaillierte und gewissenhafte Definition der Anforderungen und die gelungene Überleitung in Konzept und Implementierung. Außerdem wird der komplette Zyklus einer wissenschaftlichen Arbeit von der Motivation über die Anforderungen bis zur Evaluation der Implementierung durchlaufen. Damit ist die Arbeit im Vergleich zu anderen sehr umfangreich. Insgesamt entstand so eine qualitativ herausragende Arbeit, die das Forschungsfeld der Dialogsysteme weiter voranbringt.

Die Auszeichnung für die beste Masterarbeit ging an Herrn Alisan Öztürk von der Helmut-Schmidt-Universität der Bundeswehr Hamburg für seine Arbeit mit

dem Titel „A Data-Driven Approach to Improve the Teaching of Programming“.

Programmierkenntnisse sind in praktisch allen Berufsfeldern hilfreich, in manchen auch notwendig. Ähnlich wie beim Erlernen von Fremdsprachen tun sich Lernende besonders mit den ersten Schritten der Grammatik schwer. Herr Öztürk hat für die Unterstützung der Programmierausbildung einen webbasierten Code-Editor weiterentwickelt und die kryptischen Compiler-Fehlermeldungen in (Zitat): „friendly errors“ überführt. Als Basis diente die Datenanalyse vieler tausend Bedienungen des Editors durch Studierende und die Erkenntnis der häufigsten Compilerfehlermeldungen. Herr Öztürk hat darüber hinaus unter Verwendung von Maschine Learning ein Bewertungs- und Vorhersagemodell trainiert, um Lernenden und Lehrenden mittels Punktesystem und Diagrammen zu visualisieren, wie

der aktuelle Lernfortschritt ist. So lassen sich Studierende mit Lerndefiziten besser unterstützen. Besonders positiv bewerten die Gutachten die systematische Aufarbeitung des aktuellen Standes der Forschung und den sehr guten Anschluss daran, der wesentliche wichtige Erkenntnisse über gute Metriken hinzufügt. Herrn Öztürks Datenaufbereitung, Analyse und Visualisierung sind umfangreich und unermüdlich akribisch und zeichnen diese Arbeit aus. Insgesamt erstellte Herr Öztürk eine herausragende wissenschaftliche Arbeit zur Anwendung von Maschine Learning in der Unterstützung der Programmierausbildung. Seine Erkenntnisse über gute Vorhersagemetriken bringen das Forschungsfeld Educational Data Mining maßgeblich voran.

Beide Abschlussarbeiten können Sie als PDF auf den Internetseiten der Fachgruppe E-Learning herunterladen (<https://fg-elearning.gi.de/deutsch/nachwuchsfoerderung/beste-abschlussarbeit/2017.html>). Im Namen der Ausrichtenden, der Jury und der gesamten Fachgruppe E-Learning der Gesellschaft für Informatik gratulieren wir sehr herzlich. Die Ausschreibung für die besten Arbeiten des Jahres findet sich unter <https://fg-elearning.gi.de/deutsch/nachwuchsfoerderung/beste-abschlussarbeit/ausschreibung-2018.html>.

## Tagungs- ankündigungen

### 40 Jahre GI-Fachgruppe EMISA: Von den Roots in die Zukunft

Digital Ecosystems of the Future: Methods, Techniques and Applications: Vor 40 Jahren war die GI gerade mal zehn Jahre alt und hatte rund

2000 Mitglieder, die sich größtenteils sehr aktiv in die Arbeit ihrer Fachgesellschaft einbrachten: Mitgliederversammlungen waren damals riesige Ereignisse mit Hunderten von Teilnehmern. Vor allem die Fachausschüsse (Fachbereiche wurden erst später eingeführt) konnten die vielfältigen Interessen und Aktivitäten ihrer Mitglieder organisatorisch nicht mehr ohne zusätzliche Strukturen bewältigen. Dies führte zur Einrichtung von Fachgruppen, die sich jeweils einem engeren Themenbereich widmen sollten.

Es begann mit den Fachgruppen „Softwaretechnik“ und „Datenbanksysteme“, denen die EMISA (damals noch unter der Bezeichnung „Formale Modelle für Informationssysteme“) im Mai 1979 folgte. Ihre Mission war und ist, Modellierung und Methodik in den Vordergrund zu stellen und dabei auch nicht-technische, organisatorische und gestalterische Fragestellungen im Kontext des Entwurfs und der Realisierung von Informationssystemen zu behandeln – für „gestandene praktische Informatiker“ damals noch ein eher schillerndes Gebiet.

Zur Gründungsveranstaltung in Tutzing am Starnberger See, damals noch im Tanzsaal eines alten Gasthofes mit knarrenden Dielen und urigem Ambiente, kamen über hundert höchst diskussionsfreudige Teilnehmer. Im Folgejahr wurde in die Evangelische Akademie mit ihrem runden Auditorium gewechselt, das Diskussionen geradezu automatisch entstehen ließ, konnte doch jeder jedem ins Auge sehen. Unvergesslich ein Abend, an dem die Teilnehmer – nach einem ausgiebigen bayrischen Buffet – bis weit nach Mitternacht um Carl Adam Petri saßen, begeistert von seinen feinsinnigen und in jedem Wort überlegten Ausführungen. Diese Veranstaltungen in der

Evangelischen Akademie wurden in der Folge fast zur Tradition – zum 40. Geburtstag soll sie wieder aufleben genauso wie die EMISA-Tradition Wissenschaftler und Praktiker zu intensiven Diskussionen zusammen zu bringen. Denn immer wenn diese Integration von Theorie und Praxis besonders gut gelungen war, war auch der Mitgliederzuspruch am höchsten.

Die Fachtagung „40 Years EMISA – Digital Ecosystems of the Future: Methods, Techniques and Applications“ findet also in der Evangelischen Akademie Tutzing statt, vom 15.–17. Mai 2019. Das Format ist dem Anlass angepasst und besteht zum einen aus eingereichten Beiträgen, begutachtet von Mitgliedern eines internationalen Programmkomitees. Zum anderen werden Wegbegleiter und Vortragende früherer Veranstaltungen über ihre Erfahrungen, Einschätzungen und Prognosen zur Entwicklung im Themengebiet der EMISA berichten. Darüber hinaus sind alle einschlägigen Forschungsgruppen eingeladen, ihre aktuellen Projekte im Rahmen einer „EMISA 2019 Madness“ und eines World Cafés vorzustellen. Fazit: da muss man einfach dabei sein!

Tagungsleitung: Heinrich C. Mayr, Stefanie Rinderle-Ma und Stefan Strecker  
Deadline für Einreichungen: 15. Dezember 2018  
Webseite: <https://ae-ainf.aau.at/EMISA2019>

## **INFORMATIK 2019 in Kassel „50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft“**

Die Jahrestagung INFORMATIK 2019 findet anlässlich der GI-Gründung im Jahr 1969 vom 24.–26. September 2019 an der Universität Kassel statt. Dort wird der 50. Geburtstag mit ei-

nem attraktiven Programm für und mit „Jung und Alt“ aus Wissenschaft und Praxis gewürdigt. Das Leitthema der Tagung ist „50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft“. Es reflektiert zum einen ein zentrales Anliegen der GI, zum anderen einen Schwerpunkt der Informatik an der Universität Kassel. Gegenüber früheren Jahrestagungen ändert sich die Tagungsstruktur: Themen, die besondere Aufmerksamkeit in Wissenschaft, Praxis und Gesellschaft erfahren, werden im Rahmen vorgegebener thematischer „Tracks“ behandelt, die jeweils von ausgewiesenen Fachleuten aus der Informatik-Community gestaltet werden. Die besten Beiträge der Tracks werden zur Veröffentlichung in passenden Zeitschriften weitergeleitet. Weitere Informationen finden Sie unter [www.informatik2019.de](http://www.informatik2019.de).

## **Aus den assoziierten Gesellschaften**

### **50 Jahre German Chapter of the ACM: Mensch-Sein mit Algorithmen**

Die so genannte digitale Transformation ist allgegenwärtig und unumkehrbar. Sie betrifft alle Bevölkerungsschichten und Lebensbereiche. Der zentrale Begriff „Algorithmus“ wird dabei in letzter Zeit oftmals reduziert auf eine negative konnotierte Interaktion zwischen Mensch und Maschine. Eigentlich ist ein Algorithmus aber etwas neutrales, eine Berechnungsvorschrift wie sie Menschen seit vielen Jahrhunderten anwenden.

Zur Förderung des Dialogs zwischen den Algorithmen-Schaffenden – oftmals Informatikern – und An-

wendern schuf das German Chapter of the ACM anlässlich seines 50jährigen Bestehens unter Federführung von Marc-Oliver Pahl und Gerhard Schimpf das Forum Mensch-Sein mit Algorithmen (<https://menschsein-mit-algorithmen.org/>). Neben Fachartikeln kommen dort vor allem Menschen mit ihrer persönlichen Sicht auf die digitale Transformation zur Sprache.

Kulminationspunkt der Aktivitäten war das gleichnamige Symposium in Heidelberg am 20. und 21.9.2018, an dem unter vielen anderen die Turing-Award-Gewinner Vint Cerf, Martin Hellman, Whitfield Diffie und Bob Tarjan teilnahmen.

Den Auftakt machte am Donnerstag in der alten Aula der Ruprecht-Karls-Universität Heidelberg Turing-Award-Gewinner Vint Cerf. Der Mitbegründer des heuti-

gen Internet stellte ausgehend von der Bedeutung des Internet für unsere Gesellschaft die Notwendigkeit heraus, noch stärker an Sicherheit, Stabilität, Offenheit und Nachvollziehbarkeit zu arbeiten. Er appellierte dabei an die Politik, notwendige Anreize zu schaffen.

Der Symposiumsfreitag im Studio der Villa Bosch der Klaus Tschira Stiftung begann mit einem Vortrag von Martin Arend (BMW) über die Möglichkeiten und Veränderungen, die autonomes Fahren und personalisierte Assistenten mit sich bringen. Ihm nachfolgend sprach Katharine Jarmul (KI-Project) über die Unterschiede zwischen KI und menschlicher Intelligenz. Sie hob hierbei vor allem die Persönlichkeit hervor, die Maschinen nicht haben. Als Dritter sprach Martin Hellmann (Stanford). Der Turing-

Award-Gewinner erläuterte die Notwendigkeit zu ethischem Handeln in allen Lebensbereichen. Ausgehend von einem frei verfügbaren, kürzlich mit seiner Frau Dorothee veröffentlichten, Buch beschrieb er, wie sich ethisches Handeln durch tägliche Praxis auch im Kleinen lernen und verinnerlichen lässt.

Als nächster beleuchtete Peter Weibel, Direktor des ZKM Karlsruhe, die künstlerische Perspektive. Er stellte heraus, wie Computeralgorithmen vor allem neue Ausdrucksmöglichkeiten in der Medienkunst schaffen und das Medium verändern. Sein Mitarbeiter Bernd Lintermann beleuchtete die ZKM Satellitenausstellung OpenCodes, von der drei wunderbare Exponate in der Villa Bosch zu sehen waren.



Abb. 7 Gruppenfoto der Teilnehmer im Garten der Villa Bosch. (Quelle: Florian Freundt)



**Abb. 8** Marc-Oliver Pahl und Vint Cerf interagieren mit „Drei Phasen der Digitalisierung“ von Bernd Lintermann und Nikolaus Völzow. (Quelle: Florian Freundt)

Den Nachmittag eröffnete Michael Strube (HITS), der vor allem über die Risiken von Natural Language Processing sprach. Er zeichnete die Gratwanderung nach, auf der sich Forschung in diesem Bereich befindet, da viele Entwicklungen sich auch missbrauchen lassen – beispielsweise zur gezielten Meinungssteuerung. Anschließend besprach Alexander Filipović, Medienethiker der Hochschule für Philosophie München, die möglichen Rollen von Philosophie und Ethik im Dialog zur digitalen Transformation insbesondere durch Aufzeigen von Handlungsräumen zur Förderung informierter Entscheidungen. Den Abschluss machte Franziska Böhm (KIT), die die rechtliche Perspektive erläuterte. Sie ging vor allem darauf ein, dass Regeln notwendig sind und Algorithmen für juristische Prozesse nicht der Allheilsbringer sind.

Anschließend diskutierten Vint Cerf, Martin Hellmann, Whitfield Diffie, Katharine Jarmul und Jonas Bedford-Strohm (BR), moderiert von Thomas Matzner, über

die Auswirkungen der digitalen Transformation. Ein wichtiges Ergebnis des Panels ist in Bezug auf die Arbeitswelt, dass ihre Arbeit den Menschen Würde gibt. Gerade bei der zunehmenden Digitalisierung der Arbeitswelt und damit einem potenziellen Wegfall von Tätigkeiten benötigen Menschen auch in Zukunft würdige Jobs, um glücklich sein zu können. Neben weiteren spannenden Diskussionen zwischen den Teilnehmern wurde das Publikum Zeuge eines „Diffie-Hellman Seat Exchange“ in Anlehnung an den Diffie-Hellmann Key Exchange, der bei verschlüsselten Webseitenaufrufen jeden Tag vielfach zur Anwendung kommt.

Die vielen spannenden Diskussionen vom Symposium werden auf der Website <https://menschsein-mit-algorithmen.org/> auch in Zukunft fortgesetzt. Neben weiteren Interviews unter anderem mit weiteren Turing-Award-Gewinnern ist das Erscheinen von spannenden Vortragsmitschnitten für die nächsten Monate dort geplant.

## **Bundesweit Informatiknachwuchs fördern**

### **Medaillen für Informatiknachwuchs bei Informatikolympiade 2018 in Japan**

Bei der Abschlussfeier der 30. Internationalen Informatikolympiade (IOI) im japanischen Tsukuba konnten sich in diesem Jahr drei der vier Mitglieder des deutschen Teams über eine besondere Auszeichnung freuen. Tobias Schindler aus Kelheim sicherte sich eine Silbermedaille, dank starker Leistung gab es Bronze für Erik Sünderhauf aus Werdau und Florian Jüngermann aus Mayen.

Die IOI ist die jährlich stattfindende Weltmeisterschaft für den Informatiknachwuchs. Über 80 Länder beteiligen sich daran. In zwei Prüfungen müssen die Schülerinnen und Schüler innerhalb kurzer Zeit anspruchsvolle algorithmische Probleme bearbeiten und ihre Lösungen in fehlerfreie Programme umsetzen. In zwei der sechs Aufgaben ging es zum Beispiel darum, Steuerungseinheiten für die in Japan traditionsreichen mechanischen Puppen effizient zu entwerfen, bzw. mit möglichst wenig Versuchen eine geheime Tastenkombination in einem Videospiel zu ermitteln.

Tobias Schindler erreichte bei beiden Aufgaben hervorragende Punktzahlen und konnte weitere Aufgaben zumindest teilweise lösen. Mit dieser Leistung landete er im oberen Viertel des mit 341 Jugendlichen aus 87 Nationen besetzten Teilnehmerfeldes und gewann eine Silbermedaille. Für Deutschland vor Ort, in der „Stadt der Wissenschaft“ Tsukuba City war auch Janine Lohse aus Dresden, die nur knapp die Medaillentränge verpasste.



**Abb. 9** Das stolze Team: v. l. n. r. Erik Sünderhauf, Janine Lohse, Tobias Schindler und Florian Jüngermann. (Quelle: BWINF)

Gesamtsieger der IOI 2018 wurde der US-Amerikaner Benjamin Qi mit 499 von möglichen 600 Punkten. Damit lag er 30 Punkte vor dem Zweitplatzierten Maolong Yang aus China. Insgesamt schnitt das Team aus China am besten ab, das als einziges vier Goldmedaillen mit nach Hause nehmen konnten. Jeweils drei Goldmedaillen gingen an die Teams aus Korea und USA.

Im nächsten Jahr wird die IOI in Singapur stattfinden. Von den Mitgliedern des diesjährigen deutschen IOI-Teams kann sich aus Altersgründen nur Erik Sünderhauf erneut qualifizieren. Diese Chance haben ebenfalls erfolgreiche Teilnehmerinnen und Teilnehmer des 36. Bundeswettbewerbs Informatik. Das deutsche Auswahlverfahren zur Internationalen Informatikolympiade wird seit 1989 von den Bundesweiten Informatikwettbewerben (BWINF) durchgeführt. BWINF wird vom Bundesministerium für

Bildung und Forschung gefördert. Träger sind die Gesellschaft für Informatik e. V., der Fraunhofer-Verbund IUK-Technologie und das Max-Planck-Institut für Informatik. Die Bundesweiten Informatikwettbewerbe sind von der Kultusministerkonferenz empfohlene Schülerwettbewerbe und stehen unter der Schirmherrschaft des Bundespräsidenten.

### **Deutschlands talentiertester Informatiknachwuchs ausgezeichnet – Finale des 36. Bundeswettbewerbs Informatik am Heinz Nixdorf Institut**

Für 25 junge Informatiktalente aus ganz Deutschland war es ein großer Tag, rund ein Jahr nach dem Start des 36. Bundeswettbewerbs Informatik. Sie waren die Besten der ersten beiden Wettbewerbsrunden, an denen insgesamt rund 1500 Schülerinnen und Schüler teilgenommen hatten. Damit hatten sie sich für

die Endrunde des von den Bundesweiten Informatikwettbewerben (BWINF) jährlich durchgeführten Talentwettbewerbs qualifiziert. Das Finale wurde vom Heinz Nixdorf Institut und dem Institut für Informatik der Universität Paderborn ausgerichtet. Unterstützt wurde die Endrunde außerdem von den Firmen Atos und dSPACE sowie dem Sonderforschungsbereich 901 „On-the-fly Computing“ der Universität Paderborn.

### **Fünf Nachwuchstalente wurden als Bundessieger ausgezeichnet:**

Kamal Abdellatif, 18 Jahre, aus Jena (Carl-Zeiss-Gymnasium, Jena)

Lorenzo Conti, 16, aus Bonn (Carl-von-Ossietzky-Gymnasium, Bonn)

Tim Gerlach, 18, aus Gotha (Carl-Zeiss-Gymnasium, Jena)

Janine Lohse, 18, aus Dresden (Gymnasium Bürgerwiese, Dresden)

Paul Weiß, 17, aus Sinzing (Goethe-Gymnasium, Regensburg)

Sie erhalten jeweils ein Preisgeld in Höhe von 750 € und werden in die Studienstiftung des deutschen Volkes aufgenommen. Wegen ihrer herausragenden Gesamtleistung wurde Janine Lohse außerdem mit dem Ingo-Wegener-Preis ausgezeichnet.

### **Weitere Preisträger sind:**

Martin Bartram, 18, aus Cottbus (Max-Steenbeck-Gymnasium, Cottbus)

Jonathan Baumann, 17, aus Scheibhardt (Europa-Gymnasium, Wörth am Rhein)

Vincent de Bakker, 17, aus Willich (Priv. St. Bernhard-Gymnasium, Willich)

Gabriel Dengler, 18, aus Reichertshofen (Gnadenenthal-Gymnasium Ingolstadt)

Simon Döring, 19, aus Großen Buseck



**Abb. 10 Die Bundessieger des 36. BwInf – eingerahmt von BWINF-Geschäftsführer Dr. Wolfgang Pohl (links) und dem Vorsitzenden der Jury Prof. Dr. Till Tantau (rechts). (Quelle: Heinz Nixdorf Institut)**



**Abb. 11 Die Preisträger des 36. BWINF – eingerahmt von BWINF-Geschäftsführer Dr. Wolfgang Pohl (links) und dem Vorsitzenden der Jury Prof. Dr. Till Tantau (rechts). (Quelle: Heinz Nixdorf Institut)**

(Gesamtschule Gießen Ost, Gießen)  
Christian Hagemeier, 18, aus Düsseldorf (St. Ursula-Gymnasium, Düsseldorf)

Sie erhalten jeweils ein Preisgeld in Höhe von 500 €.

In den ersten beiden Runden des bundesweiten Informatikwettbewerbs hatten die Jugendlichen Zimmerverteilungen für Klassenfahrten berechnet und bauliche Maßnahmen gegen Wildschwein-

plagen optimiert. In der Endrunde führten sie Gespräche mit hochkarätigen Experten und bearbeiteten unter Zeitdruck zwei Informatikprobleme im Team. Eine Aufgabe beschäftigte sich mit der Modellierung praktischer Probleme durch logische Formeln, die von Computern effizient verarbeitet werden können; die andere mit der Unterstützung blinder Menschen durch Smartphone-Apps.

Die Leistungen der jungen Leute überzeugten Jury und Gastgeber. „Die Preisträger haben mit ihrem breiten Wissen überzeugt, das bereits das Niveau des ersten Studienjahres Informatik erreicht hat. Zugleich haben sie bewiesen, dass sie komplexe Probleme auch als Teamplayer effizient und kreativ bearbeiten können“, erklärt der Vorsitzende der Jury und des BWINF-Beirats, Prof. Dr. Till Tantau (Universität Lübeck).

## GI-Veranstaltungskalender

**24.01.2019 – Potsdam**

Industrie 4.0-Konferenz  
<https://hpi.de/veranstaltungen/hpi-veranstaltungen/2019/industrie-40-konferenz-2019.html>

**18.02.–22.02.2019 – Stuttgart**

Software Engineering 2019  
**SE 2019**  
<https://se19.uni-stuttgart.de/>

**24.02.–27.02.2019 – Siegen**

14. Internationale Tagung Wirtschaftsinformatik  
**WI 2019**  
<http://wi2019.de/>

**04.03.–08.03.2019 – Rostock**

18. Fachtagung „Datenbanksysteme für Business, Technologie und Web“  
**BTW 2019**  
<http://www.btw2019.de>

**18.03.–20.03.2019 – Potsdam**

10. Konferenz Professionelles Wissensmanagement: Wissensmanagement in digitalen Arbeitswelten – Aktuelle Ansätze und Perspektiven  
**WM2019**  
<http://wm2019.fh-potsdam.de>

**25.03.–29.03.2019 – Hamburg**

IEEE International Conference on Software Architecture  
**ICSA 2019**  
<http://icsa-conferences.org/2019/>

**23.05.–24.05.2019 – Potsdam**

Potsdamer Konferenz für Nationale CyberSicherheit  
<https://www.potsdamer-sicherheitskonferenz.de/konferenz.html>

**18.06.–21.06.2019 – Rennes/France**

CARS 2019 Computer Assisted Radiology and Surgery – 33rd International Congress and Exhibition  
**CARS 2019**  
<https://www.cars-int.org>

**08.09.–11.09.2019 – Hamburg**

Tagung Mensch und Computer 2019  
**MuC 2019**  
<http://muc2019.mensch-und-computer.de/>

**16.09.–18.09.2019 – Dortmund**

Informatik und Schule  
**INFOS 2019**  
<https://infos2019.cs.tu-dortmund.de/>

**18.09.–20.09.2019 – Chemnitz**

Simulation in Produktion und Logistik  
**ASIM 2019**  
<http://www.asim-fachtagung-spl.de/>

**23.09.–26.09.2019 – Kassel**

INFORMATIK 2019: Jahrestagung der Gesellschaft für Informatik  
**INFORMATIK 2019**  
<http://www.informatik2019.de>

**06.09.–09.09.2020 – Magdeburg**

Tagung Mensch und Computer 2020  
**MuC2020**  
<https://muc2020.mensch-und-computer.de>

**29.09.–01.10.2020 – Karlsruhe**

INFORMATIK 2020: Jahrestagung der Gesellschaft für Informatik  
**INFORMATIK 2020**  
[gs@gi.de](mailto:gs@gi.de)