



Digitale Forschungsdaten bewahren und nutzen - für die Wissenschaft und die Zukunft

nestor Arbeitsgruppe Grid/e-science und Langzeitarchivierung

nestor-bericht



GEFÖRDERT VOM
 Bundesministerium
für Bildung
und Forschung



Digitale Forschungsdaten
bewahren und nutzen
- für die Wissenschaft
und für die Zukunft

nestor Arbeitsgruppe Grid / e-science
und Langzeitarchivierung

nestor - bericht



Herausgegeben von

nestor - Kompetenznetzwerk Langzeitarchivierung und
Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

nestor - Network of Expertise in Long-Term Storage of Digital Resources

<http://www.langzeitarchivierung.de>

Projektpartner:

Bayerische Staatsbibliothek, München

Bundesarchiv

Deutsche Nationalbibliothek (Projektleitung)

FernUniversität in Hagen

Humboldt-Universität zu Berlin - Computer- und Medienservice / Universitätsbibliothek

Institut für Museumsforschung, Berlin

Niedersächsische Staats- und Universitätsbibliothek, Göttingen

Autoren:

Reinhard Altenhöner, Deutsche Nationalbibliothek

Harry Enke, Astrophysikalisches Institut Potsdam

Bernadette Fritsch, Alfred-Wegener-Institut für Polar- und Meeresforschung

Jens Klump, GeoForschungsZentrum Potsdam

Michael Lautenschlager, World Data Center Climate

Jens Ludwig, Niedersächsische Staats- und Universitätsbibliothek Göttingen

Heike Neuroth, Niedersächsische Staats- und Universitätsbibliothek Göttingen

Kontakt für die AG Grid /e-science und Langzeitarchivierung:

c/o Niedersächsische Staats und Universitätsbibliothek Göttingen

Heike Neuroth

Papendiek 14

D-37073 Göttingen

Tel.: +49 (0)551-393866

E-Mail: <neuroth@sub.uni-goettingen.de>

URN: urn:nbn:de:0008-2009071031

<http://nbn-resolving.de/nbn:de:0008-2009071031>

© 2009

nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit
Digitaler Ressourcen für Deutschland

Der Inhalt dieser Veröffentlichung darf vervielfältigt und verbreitet werden, sofern der Name des Rechteinhabers "nestor - Kompetenznetzwerk Langzeitarchivierung" genannt wird. Eine kommerzielle Nutzung ist nur mit Zustimmung des Rechteinhabers zulässig.

Digitale Forschungsdaten bewahren und nutzen – für die Wissenschaft und die Zukunft Was brauchen wir in kollaborativen Forschungsumgebungen?

Präambel

Die DFG Richtlinien zur Sicherung wissenschaftlicher guter Praxis besagen u.a. in der Empfehlung 7, dass Primärdaten als Grundlagen für Veröffentlichungen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden sollen.¹

Die nachhaltige Speicherung und auch Nachnutzung von Forschungsdaten ist eine strategische, organisatorische und technische Aufgabe, die nur gemeinsam mit der Wissenschaft und Forschung gelöst werden kann.

Die Arbeitsgruppe "Grid/eScience und Langzeitarchivierung" des Kompetenznetzwerks Langzeitarchivierung "nestor" sieht einen dringenden Handlungsbedarf für die Sicherung öffentlich geförderter Forschungsdaten. Mit dem Ziel, die Qualität, Produktivität und Wettbewerbsfähigkeit der Wissenschaft zu fördern, sieht sie den folgenden Handlungsbedarf für ein koordiniertes weiteres Vorgehen:

Erklärung

Digitale Forschungsdaten sind sowohl die unverzichtbare Basis als auch das Resultat der modernen wissenschaftlichen Arbeit. Jede Disziplin hat ihre eigenen technischen und inhaltlichen Ansprüche, ihre Datenbestände und Konventionen, sodass Forschungsdaten alle Formate und Medientypen umfassen: von einfachen Tabellen im Textformat für z.B. physikalische Messdaten über semantisch ausgezeichnete Textsammlungen der Literaturwissenschaften bis hin zu interaktiven 3D-Modellen von Ingenieuren. Allen ist aber gemeinsam, dass sie eine wertvolle, jedoch schwierig zu handhabende Ressource darstellen. Forschungsdaten werden aufwändig produziert, im Falle von Beobachtungsdaten wie z.B. Wetterbeobachtungen können sie einmalig und nicht wiederherstellbar sein, sie können riesige Datenmengen umfassen und wie alle digitalen Daten werden sie ohne langfristige wirksame Gegenmaßnahmen technisch veralten, nicht mehr nutzbar sein und damit verloren gehen.

Forschungsdaten müssen als Ressource begriffen werden, die sowohl für zukünftige Forschungen der eigenen als auch anderer Disziplinen wichtig ist und effizient genutzt werden sollte. Das Potential nachnutzbarer und verknüpfbarer Daten ist bereits in frei verfügbaren Internetanwendungen sichtbar, wo beispielsweise frei zugängliche Geodatenbestände und Dienste für unvorhergesehene Zwecke wie die wissenschaftliche Untersuchung von Epidemien eingesetzt werden.² Um die Investitionen in die Erzeugung von Forschungsdaten voll auszuschöpfen, reicht es nicht, eine Langzeitarchivierung zu verfolgen, die als einmalige Aktivität ein statisches Er-

¹ DFG: Vorschläge zur Sicherung guter wissenschaftlicher Praxis, Bonn 1998, http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

Neben den Anforderungen an die gute wissenschaftliche Praxis der DFG von 1998 sind 2009 auch Empfehlungen für den Umgang mit Forschungsdaten erschienen: DFG: Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, Bonn 2009, http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/veroeffentlichungen/dokumentationen/download/ua_inf_empfehlungen_200901.pdf

² Z.B. <http://www.nature.com/avianflu/google-earth/index.html> oder auch Declan Butler, Virtual globes: The web-wide world Nature 439, 776-778 (16 February 2006), doi:10.1038/439776a

gebnis konserviert und wegschließt. Stattdessen muss eine Pflege der Daten über ihren gesamten Lebenszyklus betrieben werden, wie sie im Englischen unter "Data Curation" verstanden wird³ und die ein breites Spektrum von Aktivitäten wie die angemessene inhaltliche Erschließung, Veröffentlichung, sichere Speicherung und die inhaltliche und technische Versionierung der Daten an den jeweiligen Stand der Wissenschaft und Technik umfasst.

Um eine Nachnutzung von Forschungsdaten in Gegenwart, Zukunft und über Disziplingrenzen hinweg zu ermöglichen, ist Interoperabilität eine der zentralen Anforderungen. Eine wesentliche Voraussetzung dafür ist eine technische und organisatorische Basisinfrastruktur für die Bereitstellung und das Management von Forschungsdaten. Und diese muss sich mit ihren Diensten bruchlos in die Arbeitsumgebungen ganz verschiedener Fachdisziplinen integrieren. Neue wissenschaftliche Fragestellungen motivieren die Verwendung neuartiger Methoden und Instrumente, die z.B. durch virtuelle Forschungsumgebungen und Grid-Infrastrukturen unterstützt werden. Die Dateninfrastruktur muss damit schritthalten. Die dafür nötigen Aufgaben werden hier vorgestellt.⁴

Konzeptionelle Aufgaben

Langzeitarchivierung, Verfügbarkeit und Nachnutzung von Forschungsdaten setzen fachlich fundierte und in den Fachdisziplinen akzeptierte Vorgehenskonzepte voraus. Solche Konzepte beinhalten technisch-operative Aspekte, planerische Schritte, sowie Richtlinien, Regeln und Prinzipien, die helfen, den richtigen Ansatz für jede Fachdisziplin zu finden. Dabei muss jede Fachdisziplin geeignete Strategien, Methoden und Arbeitsabläufe entwickeln, die je nach Anwendungsfall variieren.

Zentrale Maßnahmen sind:

- Definition akzeptierter Kriterien für die Bewertung der Vertrauenswürdigkeit von Langzeitarchiven
- Definition von Service Levels für verschiedene Anwendungsfälle in denen Auswahlkriterien, Erhaltungsmaßnahmen, Aufbewahrungszeiträume, Bereitstellung und Verfügbarkeit der archivierten Daten beschrieben werden. Entwicklung von Service Level Agreements, die als Mustervorlagen dienen können.
- Aufstellung eines Kriterienkatalogs zur Bewertung der Archivfähigkeit von Dateiformaten
- Entwicklung von Metadatenschemata und angemessenen Fach-Ontologien für Forschungsdaten, um langfristig Austauschbarkeit, Auffindbarkeit und Nachnutzbarkeit zu fördern
- Sammlung von Use Cases und Best Practices für fachspezifische Langzeitarchivierung

³ "Digital curation [...] is about maintaining and adding value to a trusted body of digital information for both current and future use: [...] it is the active management and appraisal of digital information over its entire life cycle." Pennock, Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information, 2007, http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf

⁴ Viele dieser Aufgaben werden in den folgenden drei, von der nestor AG betreuten Expertisen genauer begründet und vorgestellt: Jens Klump: Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten, nestor 2008, urn:nbn:de:0008-2008040103; Uwe M. Borghoff, Peter Rödiger: Standards und Standardisierung im Kontext von Grid/eScience und Langzeitarchivierung, nestor 2009, urn:nbn:de:0008-2009012103; Udo Hönig, Wolfram Schiffmann: Synergiepotenziale zwischen Grid- und eScience-Technologien für die digitale Langzeitarchivierung, nestor 2009, urn:nbn:de:0008-2009012115

Organisatorische Aufgaben

Die Rahmenbedingungen müssen definiert werden, unter denen die Langzeitarchivierung in den einzelnen Fachdisziplinen umgesetzt wird. Dies ist neben den organisatorischen Anforderungen vor allem die Finanzierung der Langzeitarchivierung. Es sind der Aufwand abzuschätzen und Modelle zu entwickeln, wie dieser in der wissenschaftlichen Fachdisziplin erbracht werden kann. Nicht alle wissenschaftlichen Einrichtungen können derzeit eine Langzeitarchivierung für die wissenschaftlichen Daten ihrer Mitarbeiter bereitstellen. Daher ist über die Etablierung von fachspezifischen Zentren für die Langzeitarchivierung nachzudenken, wobei diese Zentren fachübergreifend zusammenarbeiten sollten, um Synergien zu entwickeln und zu nutzen.

Zentrale Maßnahmen sind:

- Ausarbeitung von Organisations- und Kostenmodellen für Langzeitarchivierung/Data Curation
- Etablierung von (weiteren) fachspezifischen Zentren für Data Curation (moderne Gedächtnisinstitutionen)
- Zusammenarbeit zwischen Data Curation Zentren zur Identifikation von Synergien für Schnittstellen, Dienste, Workflows und Service Levels, um Zugänglichkeit, Interoperabilität und Ausfallsicherheit der Langzeitarchive zu gewährleisten
- Förderung von Aus-, Fort- und Weiterbildung im wissenschaftlichen Datenmanagement

Technologische Aufgaben

Geeignete Technologien zur Durchführung und Unterstützung der Langzeitarchivierung müssen entwickelt werden. Hierzu zählen neben den Basisinfrastrukturen (Netzwerke, Hardware, Sicherheit (PKI)) insbesondere Repositorien- und Archiv-Technologie mit zugehörigen Werkzeugen und Diensten zur effizienten Nutzung. Diese müssen auf technischen und disziplinspezifischen Standards beruhen, die noch weiter zu entwickeln sind.

Fachspezifische Maßnahmen:

- Entwicklung von fachspezifischen Systemen/Repositorien basierend auf Best Practices
- Entwicklung von Diensten und Werkzeugen zur Integration der Repositorien in fachspezifische Arbeitsabläufe und Forschungsumgebungen
- Infrastruktur zur Dokumentation von Forschungsdaten (Metadaten, Ontologien)

Übergreifende Maßnahmen:

- Infrastruktur für verteilte Rechteverwaltung
- Infrastruktur zur persistenten Identifikation, Versionierung und Lokalisierung von Datenobjekten und ihren Teilobjekten
- Infrastruktur für Metadaten für förderbare Kataloge
- Infrastruktur für Ontologien wie Verzeichnisdienste und Mappingdienste
- (fachspezifische) Entwicklung von Diensten zur Migration und Emulation
- (fachspezifische) Entwicklung von Diensten zur Validierung, Charakterisierung/Metadatenextraktion und Qualitätsmanagement
- zentrale Dienste und Verzeichnisse für Informationen über Dateiformate
- Durchführung von Pilotentwicklungen, Testbeds, etc.