Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen
AG Virtuelle Forschungsumgebungen

Momentaufnahme: VREs in 2011/2012

[März 2013]

- Inhalt -

1. Kommentar zu den Interview-Ergebnissen der Allianz Arbeitsgruppe Virtuelle Forschungsumgebungen 2

2. AstroGrid-D – German Astronomy Community Grid 4

3. BW-eLabs - Wissensmanagement in virtuellen und remote Laboren 4

4. C3-Grid - Collaborative Climate Community Data and Processing Grid 5

5. CLARIN - Common Language Resources and Technology Infrastructure 6

6. Meta-Image - Projekt Prometheus an der Universität Köln und Projekt HyperImage, Universität Lüneburg, Computer- und Medienservice der HU Berlin 7

7. Diversity Workbench als VFU 8

8. FuD – Forschungsnetzwerk und Datenbanksystem 8

1. **Zusammenfassender** Kommentar zu den Interview-Ergebnissen der Allianz Arbeitsgruppe Virtuelle Forschungsumgebungen

Die Arbeitsgruppe hat Interviews mit folgenden sieben Projekten, die sich mit dem Aufbau von virtuellen Forschungsumgebungen beschäftigen, geführt:

- AstroGrid-D - German Astronomy Community Grid
- BW-eLabs - Wissensmanagement in virtuellen und remote Laboren
- C3-Grid - Collaborative Climate Community Data and Processing Grid
- CLARIN - Common Language Resources and Technology Infrastructure
- Meta-Image - Projekt Prometheus an der Universität Köln und Projekt HyperImage, Universität Lüneburg, Computer- und Medienservice der HU Berlin
- Diversity Workbench als VFU
- FuD - Forschungsnetzwerk und Datenbanksystem

Basierend auf der Zusammenfassung der jeweiligen Interviews können folgende Schlüsse gezogen werden:

1. Die vorgesehene Reichweite der Projekte ist sehr unterschiedlich: Während es sich bei AstroGrid-D, C3-Grid und CLARIN, um Projekte mit einem (zumindest geplanten) internationalen Zuschnitt handelt, die den Anspruch haben, jeweils eine ganze Wissenschaftscommunity zu adressieren, konzentrieren sich die anderen Projekte auf spezielle Forschungsfragen und werden daher von einer geringeren Anzahl von Forschungsgruppen in Anspruch genommen. In der Mehrzahl der Projekte ist eine nationale oder internationale Nutzung von Daten langfristig vorgesehen.
2. Der in der Definition von Virtuellen Forschungsumgebungen herausgestellte Begriff der kooperativen Forschungstätigkeit wird in den Projekten AstroGrid-D, C3-Grid und CLARIN explizit erwähnt. Hier liegen mit den großen kosmologischen Simulationen, den Klima-Assessment-Reports des IPCC und dem Atlas für bedrohte Sprachen auch Beispiele für umfangreiche internationale Kollaborationen vor. In anderen Projekten liegt der Fokus auf der Bereitstellung von Daten oder anderen Ressourcen (BW-eLabs).
3. In allen Projekten mit der potentiell strukturbedingten Ausnahme des SFB Fremdheit und Armut sind jeweils Wissenschaftler an unterschiedlichen Orten eingebunden.
4. Die meisten Projekte erwähnen explizit die Unterstützung des gesamten Forschungsprozesses oder wesentlicher Teile von ihm, teilweise mit einer Spezifizierung der jeweiligen Komponenten.
5. Die Bereitstellung von Softwarediensten ist Teil aller Projekte.

6. Alle Projekte bieten Zugang zu realen Forschungsressourcen, wobei vielfach der Fokus auf Daten liegt.
7. Viele Projekte legen Wert auf Standardisierung. Schwerpunkt dabei sind Metadaten.
8. Fünf Projekte adressieren explizit Nachhaltigkeitsfragen, wobei bisher nur Lösungsansätze bestehen. Bei den anderen beiden Projekten konzentriert sich die Nachhaltigkeit auf eine langfristige Datenbereitstellung (Langzeitarchivierung).
9. Die virtuellen Forschungsumgebungen werden in allen Projekten nur von wenigen Einrichtungen getragen.

2. AstroGrid-D – German Astronomy Community Grid

http://www.astrogrid-d.org

Das AstroGrid-D ist ein Verbundprojekt von 7 astronomischen Forschungsinstituten und Informatik-Forschungs¬gruppen, (sowie 7 assoziierter Institute aus der Astronomie und Rechenzentren). Das Forschungsvorhaben auf dem Gebiet e-Science und Grid-Middleware zur Unterstützung wissenschaftlichen Arbeitens im Rahmen der deutschen D-Grid-Initiative verfolgte die Entwicklung eines Rahmens samt zugehöriger Standards für das kolla¬borative Management astronomiespezifischer Grid-Ressourcen und einer dafür geeigneten Infrastruktur. Es wurde vom BMBF (PT-IN) 2005-2009 gefördert, Anschluss-Projekte sind VDZ (BMBF PT-DESY) und WissGrid (BMBF PT-IN).

Verwendete Standards und Metadaten-Standards werden einerseits in der Astronomie über die IAU (Internat. Astronomical Union) und die IVOA (international Virtual Observatory Alliance) erarbeitet, die Standards des OGF (Open Grid Forum) für Grid-Dienste werden durch die Verwendung von Globus Toolkit gesichert. Für die zukünftige Weiterentwicklung ist insbesondere die Erarbeitung von integrierten Standards für Metadaten wich¬tig.

Basis-Dienste für die Nutzung von Grid-Infrastruktur werden im Rahmen des VDZ (Virtuelles Datenzentrum) durch das AIP auch nach Projektende unterhalten und von einigen Instituten und Arbeitsgruppen derzeit genutzt. Es haben sich das GAVO-Datenzentrum (German Astrophysical Virtual Observatory) in Heidelberg und das VDZ in Potsdam als wichtige Stützpunkte für VRE-Bestrebungen in der Astronomie gebildet. Derzeit werden v.a. projektbasiert, über GAVO und das VDZ, einzelne Werkzeuge zur Nutzung der Basis-Infrastrukturen weiterentwickelt. Das VDZ hat den Schwerpunkt in der Bereitstellung der Daten von großen kosmologischen Simulationen für internationale Kollaborationen und von wichtigen astronomischen Archiven wie dem SDSS (für Public Access), jeweils in der Größenordnung von >50 Terabyte; das GAVO-Datenzentrum stellt kleinere publizierte Archive gemäß den Standards des Virtual Observatory zur Verfügung.

Die vorhandenen Infrastrukturen (VO AstroGrid, Grid-Datenserver und -Rechner) bilden eine wichtige Grund¬lage für neue internationale Kollaborationen, welche das Datenmanagement von neuen Instrumenten unter Verwendung des Grid aufbauen.

Die Kosten für hierfür werden teils aus Projektmitteln, teils aus den Institutshaushalten aufgebracht. Eine Fi¬nanzierung der nachhaltigen Bereitstellung ist noch nicht in aus-reichen¬dem Masse gesichert. Ein Anteil der Kosten wird zukünftig aus z.B. den Mitteln für das Datenmanagement von Surveys usw. kommen.

3. BW-eLabs - Wissensmanagement in virtuellen und remote Laboren

http://www.bw-elabs.org

Ziel von BW-eLabs ist es, den Zugriff im Umfeld der Materialwissenschaften und Nanotechnologie auf heterogene, experimentelle Ressourcen zu erweitern. Damit hat BW-

eLabs Gemeinsamkeiten mit dem von der Universität Stuttgart koordinierten EU-Projekt "LiLa" (Library of Labs), welches ebenfalls experimentelle Ressourcen, allerdings für Studenten, bereitstellt. Start von BW-eLabs war 08/2009, die Laufzeit betrug 30 Monate. Projektpartner sind die Uni Stuttgart, die Uni Freiburg, die Hochschule der Medien (HdM, Stuttgart) und das FIZ Karlsruhe.

Die Zielgruppe von BW eLabs sind Wissenschaftler aus den Natur- und Ingenieurwissenschaften, die im Rahmen von Technologieprojekten den Zugang zu Laboren und experimentellen Ressourcen nutzen und dabei die Forschungsdaten für Forschungs- bzw. Ausbildungszwecke nachhaltig erschließen möchten. Dafür soll eine Architektur entwickelt werden, die eSciDoc als Forschungsdatenrepository einbezieht.

Verwendete Metadatenformate basieren auf Dublin-Core sowie fachspezifischen Metadaten, die von den jeweiligen wissenschaftlichen Communities definiert wurden. Für die aus den Messungen anfallenden Daten gibt es noch keine allgemein anerkannten Formate, es werden daher eher proprietäre Formate verwendet, die u.a. im Zusammenhang mit den verwendeten Laborsystemen stehen. Es wird aber ein auf dem Core Scientific Metadata Model des Science and Technology Facilities Council (STFC) basierendes Metadatenkonzept für dynamischen Content entwickelt, das den Charakteristika von Experimenten und Versuchsaufbauten/Laboren Rechnung trägt. Eine künftige Standardisierung der Metadatenbeschreibung von Online-Experimenten wird im Global Online Laboratory Consortium (GOLC) diskutiert.

Der unterstützte Forschungsprozess reicht von der Planung von Experimenten und der hierfür benötigten Ressourcen, der Erfassung der Daten im Labor, dann vor allem über die Weiterverarbeitung der Daten bis hin zur Publikation der Daten. Hierbei kommen zur Verwaltung von Rohdaten das eSciDoc des FIZ Karlsruhe sowie Opus der Universität Stuttgart für das Management der daraus entstandenen Publikationen zum Einsatz. Eine aufzubauende Infrastruktur soll dabei auch eine Beobachtung während der Experimente und Messungen sowie ggf. eine Fernsteuerung der Laborarbeiten (z.B. Chemische Synthese) über ein Portal ermöglichen. Die digitalen Datenressourcen erfordern im Moment einen Speicherplatz von weniger als 100 TB.

Die Finanzierung erfolgt aus Mitteln des MWK Baden-Württemberg sowie Eigenmitteln der beteiligten Hochschulen und FIZ Karlsruhe.

4. C3-Grid - Collaborative Climate Community Data and Processing Grid

http://www. C3-Grid.de

Das Collaborative Climate Community Data and Processing Grid (C3-Grid) Projekt ist ein nationales Projekt im Rahmen der D-Grid-Initiative. Es lief von 2005-2008 und wurde vom BMBF gefördert. C3-Grid stellt eine Grid-basierte kollaborative Infrastruktur für die deutsche Erdsystemforschung zur Verfügung, die das Management und die wissenschaftliche Analyse

hochvolumiger Erdsystemmodell- und -beobachtungsdaten unterstützt. Das Nachfolgeprojekt C3-Grid INAD (Towards an *IN*frastructure for General *A*ccess to Climate *D*ata) läuft von 2010-2013 und wird durch das BMBF im Referat für Umwelt, Kultur und Nachhaltigkeit finanziert. Es wird die prototypische Entwicklung von C3-Grid in ein produktives System überführen und unterstützt als Kollaborationsplattform die Wissenschaftler bei der Erstellung des 5. Assessment Reports des Weltklimarates (IPCC AR5). Für den nachhaltigen Betrieb der C3-Grid-Infrastruktur ist die Umsetzung der Standards im Rahmen des IPCC AR5 eine der wichtigsten technischen Voraussetzungen. Weiterhin wird eine Einbettung in das Climate Model Intercomparison Project (CMIP5) angestrebt. Durch seine 18 nationalen und im wissenschaftlichen Kontext auch international agierenden Partner ist das Konsortium im Rahmen der Standardisierung verschiedener Infrastrukturaspekte an großen europäischen und internationalen Aktivitäten beteiligt. Zum Beispiel wird für das im Rahmen des EU FP7 Projektes Metafor entwickelte Framework für konsistente Metadatenbeschreibung im Bereich der Erdsystemforschung eine Interoperabilität mit C3-Grid angestrebt. Für den internationalen Daten¬zugriff dient das IPCC-Core-Gateway des DKRZ/WDCC als Verbindungsglied zwischen nationaler Infrastruktur und internationalen Aktivitäten. Schließlich geben die Beteiligung im Bereich der internationalen Portalent¬wicklung (GO-ESSP) und die Umsetzung der IPCC-Security Infrastructure die Leitlinien für die Entwicklungen im C3-Grid vor. Für die Metadaten werden die Standards nach ISO und W3C verwendet. Schließlich wird das C3-Grid in die IPCC-Security-Infrastructure, die auf OpenID basiert, eingebettet. Der gesamte Forschungsprozess von Datenakquisition, Datenanalyse, Daten-Kuratierung und -Erhaltung bis hin zur Datenanalyse und Ver¬wendung in Modell und Simulationsläufen wird unterstützt. Insgesamt verwalten die (verteilten) digitalen Datenressourcen der Klimadatenbanken und Klimarechenzentren unter Einbeziehung der IPCC-Daten mehr als 10 PB.

5. CLARIN - Common Language Resources and Technology Infrastructure

http://www.clarin.eu

CLARIN ist ein gesamteuropäisches, von der EU im Rahmen des ESFRI-Prozesses gefördertes Projekt. Es hat zum Ziel, eine gesamteuropäische Infrastruktur für Sprachressourcen aufzubauen. Das Konsortium umfasst 32 Partner aus 22 Ländern. Bis einschließlich 2010 befand sich CLARIN in der sog. ESFRI Präparationsphase, in der Rahmenbedingungen für den Aufbau einer Infrastruktur untersucht und erprobt werden, die digitale Textressourcen und Werkzeuge zu deren wissenschaftlichen Untersuchung verfügbar machen soll. In der ESFRI Konstruktionsphase (seit 2011) geht es vorranging um den nachhaltigen Aufbau eines paneuropäischen Verbundes von Zentren mit digitalen Textressourcen, sowie Entwicklung und Bereitstellung von digitalen Werkzeugen. CLARIN hat als Zielgruppe alle Wissenschaftsbereiche - speziell aber die Geistes- und Sozial-wissenschaften - die Zugang zu Sprachressourcen und Sprachverarbeitungswerkzeuge

nutzen möchten und unterstützt damit das interdisziplinäre und kollaborative Arbeiten mit diesen Ressourcen.

Metadatenstandards werden z.B. nach ISO, TEI und W3C verwendet. Es kommen deskriptive Metadaten, Herkunfts-Metadaten sowie technische Metadaten (für Daten-Ressourcen und Tools) zur Anwendung.

Die unterstützten Forschungsprozesse umfassen den gesamten Workflow von Datenakquisition, Datenanalyse, Datenkuratierung und Daten-Preservation. Die (verteilten) digitalen Datenressourcen umfassen mehr als 100 TB.

Nach der EU-finanzierten Vorbereitungsphase wird die Konstruktionsphase mit nationalen Mitteln gefördert, in Deutschland mit Mitteln des BMBF und von Forschungsministerien der Länder.

Als technische Voraussetzungen hinsichtlich einer Nachhaltigkeit wird auf die Konformität mit Standards und die Interoperabilität von Services, sowie auf persistente Identifizierung von Objekten und langfristig nutzbare Metadaten geachtet. Beteiligte Datenzentren verpflichten sich langfristig zur Datenbereitstellung.

Inzwischen hat CLARIN in der Konstruktionsphase den rechtlichen Status eines ERIC (European Research Infrastructure Consortium) erhalten.

6. Meta-Image - Projekt Prometheus an der Universität Köln und Projekt HyperImage, Universität Lüneburg, Computer- und Medienservice der HU Berlin

http://www.meta-image.de

In dem Web 2.0 Projekt Meta-Image ist es gelungen, das in anderen Projektzusammenhängen weitgehend ausentwickelte System "HyperImage" (www.hyperimage.eu) zur kollaborativen Bilddetailannotation in den Kontext einer großen, digital gestützten Kunstbildsammlung zu integrieren (prometheus). Diese Kunstbildsammlung ist institutionell in den deutschsprachigen Ländern, individuell weltweit verbreitet.

Im Zuge des Projekts ist "HyperImage" in der Bedienung so weit vereinfacht worden, dass es problemlos von allen Nutzern der Bildsammlung angewendet werden kann. Standards, Daten und Techniken orientieren sich an der Industrienorm. Verwendet werden open source-Produkte wie Glassfish Application Server und Java, darüber hinaus open source abgeleitete Eigenentwicklungen wie HyperImage Server, Editor und Reader. Die Daten sind jpg-kodierte Bilddaten.

Die Langzeitarchivierung wird vom Rechenzentrum der Universität Köln übernommen. Die Nachhaltigkeit ist durch das Lizenzmodell von prometheus abgesichert.

7. Diversity Workbench als VFU

http://www.diversityworkbench.net | http://www.diversitymobile.net

Das Projekt *Diversity Workbench* (DWB) wurde in 2000 gestartet und verwaltet georeferenzierte quantitative und qualitative Beobachtungs- und Messdaten aus den Bereichen Biodiversitätsforschung und Ökologie. Dazu wurde ein modularisiertes Komponenten-Framework zur Datengenerierung, -erhaltung und -prozessierung aufgebaut (12 eigenständige, interoperable Datenbanken, jeweils mit Rich-Clients zum Datenmanagement sowie browserbasierten Anwender-Schnittstellen, s. www.diversityworkbench.net/Portal/Diversity_Workbench_Performance).

Beteiligt sind derzeit 5 Arbeitsgruppen bzw. Lehrstühle aus Deutschland – zwei von der Universität Bayreuth (www.daneco.uni-bayreuth.de/daneco/) und (www.ai4.uni-bayreuth.de/de/index.html), eine von der Universität Regensburg, eine vom Julius-Kühn Institut Berlin (http://www.jki.bund.de) und eine vom IT-Zentrum der Staatl. Naturwissenschaftlichen Sammlungen Bayerns (SNSB), München, bei welcher auch die meisten Daten verwaltet und archiviert werden (www.snsb.info).

Digitale Erfassung von Daten aus der biologischen Feldforschung ermöglicht es, über längere Zeiträume biologische bzw. ökologische Veränderungen in der Natur zu erkennen und sie auch im Kontext von Klimaveränderungen zu bewerten. Damit ist ein wichtiger Schritt zur Nachhaltigkeit gegeben, der maßgeblich durch die Möglichkeit einer Langzeitarchivierung der erfassten und standardisierten Daten unterstützt wird. Nach Analyse der Daten durch die beteiligten Wissenschaftler und Veröffentlichung der daraus resultierenden Ergebnisse sind diese Daten dann der Wissenschaftsgemeinschaft über die Anbindung an internationale Netzwerke frei zugänglich. Allgemeine Basisdaten sind auch unmittelbar über den GBIF-D Knoten (www.gbif.de bzw. www.gbif-mycology.de) zugänglich.

Darüber hinaus werden Datensammlungen zu Sammlungsobjekten sowie strukturierte Beschreibungen und Taxonomien von 28 Instituten weltweit (u. a. aus Deutschland, Brasilien, Großbritannien, Russland, Spanien und USA) in den DWB-Komponenten an den SNSB verwaltet. An mehreren Forschungsinstitutionen in Deutschland wird zurzeit die DWB als eigenständige VFU für Umweltdaten eingerichtet.

8. FuD – Forschungsnetzwerk und Datenbanksystem

http://www.fud.uni-trier.de

Das Forschungsnetzwerk und Datenbanksystem (FuD) wird seit 2005 im Rahmen des Sonderforschungsbereichs 600 „Fremdheit und Armut“ an der Universität Trier entwickelt (ab 2009 im Rahmen eines INF-Projektes). Das Software-System unterstützt die Koordinierung und Organisation von Forschungsprojekten sowie die Vernetzung der Forschungsmaterialien. An der Konzeption und Programmierung sind das Kompetenzzentrum für elektronische

Erschließungs- und Publikationsverfahren in den Geisteswissenschaften sowie das Zentrum für Informations-, Medien- und Kommunikationstechnologie (ZIMK) der Universität Trier beteiligt.

Die Open-Source-Software (u.a. Client-Server-Architektur in Tcl/Tk, Browseranwendung mit CMS-Plone) ist als integrierte Arbeits-, Publikations- und Archivumgebung für die Geisteswissenschaften konzipiert. Sie ermöglicht die Zusammenarbeit in räumlich verteilten Arbeitsgruppen während der verschiedenen Phasen des Forschungsprozesses von der Inventarisierung und Erfassung der Primärdaten über ihre Erschließung und Analyse bis hin zur Ergebnispublikation und Datenarchivierung.

Die Forschungsprimärdaten, die die Wissenschaftler hauptsächlich in Archiven und Bibliotheken zusammengetragen bzw. durch eigene Untersuchungen erhoben haben, werden über standardisierte Eingabemasken im FuD erfasst und in einer MySQL-Datenbank gespeichert.

FuD entwickelt Metadatenstandards für Forschungsdaten (einheitliche Beschreibung von Dokumenttypen), die in allen Forschungsprojekten verbindlich eingesetzt werden. Die für den SFB entwickelten Beschreibungsstandards dienen dabei als Grundlage für weitere FuD-Anwendungen.

Neben den Metadaten der Dokumente können auch Volltexte und Digitalisate erfasst werden. Zusätzlich bietet die Software unterschiedliche Werkzeuge zur Dokument- und Rechteverwaltung sowie zur Textanalyse. Außerdem unterstützt FuD die Vorbereitung von Online- und Buchpublikationen. Darüber hinaus werden die Daten für die Übergabe in das im Aufbau befindliche Primärdatenarchiv der Universität Trier in standardisierte XML-Formate (MODS, TEI) umgewandelt und um Langzeitarchivierungsmetadaten ergänzt.

FuD bietet flexible Lösungen für unterschiedliche Forschungsanwendungen in den Geisteswissenschaften und wird inzwischen außerhalb des SFB in verschiedenen Projekten an Universitäten, außeruniversitären Instituten sowie Akademien und Bibliotheken eingesetzt; hierzu gehören u.a. Forschungsvorhaben der Akademie der Wissenschaften und der Literatur Mainz, der Deutschen Historischen Institute in London und Paris, der Universitäten Bochum, Mainz, Marburg, Tübingen und Wuppertal sowie der Sächsischen Landesbibliothek - Staats- und Universitätsbibliothek Dresden.