Symposium Forschungsdaten-Infrastruktur (FDI 2013)

22. Januar 2013, GeoForschungsZentrum Potsdam (GFZ)

Gemeinsam organisiert von den DFG-Projekten Radieschen, re3data.org, KomFor, EWIG und BoKeLa

Symposium - Materialien

doi:10.2312/RADIESCHEN_002

Gefördert von:

DFG Deutsche Forschungsgemeinschaft

Inhalt

Symposium Forschungsdaten-Infrastruktur (FDI 2013)

22. Januar 2013,GeoForschungsZentrum Potsdam (GFZ)

Gemeinsam organisiert von den DFG-Projekten Radieschen, re3data.org, KomFor, EWIG und BoKeLa

Symposium-Bericht

Die Flut an digitalen Daten in Wissenschaft und Forschung wächst rasant. Die dauerhafte Speicherung dieser Daten für zukünftige Generationen von Forscherinnen und Forschern stellt das gesamte Wissenschaftssystem vor Herausforderungen. Jüngst hat die EU-Kommission Empfehlungen an ihre Mitgliedsstaaten zur dauerhaften Zugänglichkeit von Forschungsdaten verabschiedet. Doch noch sind viele Fragen ungelöst. So müssen Aspekte der Finanzierung, Organisation und Technologie der zu schaffenden Forschungsdaten-Infrastrukturen, sowie deren rechtliche und politische Rahmenbedingungen geklärt werden. Diese Themenfelder wurden im Rahmen eines gemeinsam organisierten Symposiums der DFG-Projekte Radieschen, re3data.org, KomFor, EWIG und BoKeLa diskutiert.

Abb. 1: Begrüßung durch den Vorstand des GFZ – Prof. Dr. Dr. h.c. Reinhard F.J.Hüttl

Im Hauptprogramm des Symposiums wurden, mit dem Datenlebenszyklus als Leitlinie, die jeweiligen Herausforderungen und neuesten Erkenntnisse durch Vorträge und daran anschließende Diskussionen erörtert. In begleitenden Workshops wurden zentrale Themenfelder von Expertinnen und Experten bearbeitet und deren Ergebnisse anschließend dem Publikum vorgestellt.

Das Symposium versammelte die Experten-Community zu Forschungsdaten aus dem deutschsprachigen Raum. Kern der Veranstaltung bildeten Vorträge zu verschieden Aspekten der Forschungsdaten-Infrastrukturen, sowie Präsentationen laufender Forschungsprojekte. Darüber hinaus wurden zentrale Herausforderungen der Weiterentwicklung von Forschungsdaten-Infrastrukturen in Workshops diskutiert. Desweiteren bot das Symposium eine Plattform zum Gedankenaustausch und zum Netzwerken.

Das Symposium fand am 22. Januar 2013 am Helmholtz-Zentrum Potsdam, Deutsches GeoForschungsZentrum GFZ statt. Eröffnet wurde die Veranstaltung durch den Vorstand des GeoForschungsZentrums **Prof. Dr. Dr. h.c. Reinhard F.J. Hüttl** und durch ein Geleitwort von **Dr. Stefan Winkler-Nees (DFG)**. Das Vortragsprogramm begann mit einer Keynote von Dr. Torsten Reimer vom Joint Information Systems Committee (JISC).

Keynote

Dr. Thorsten Reimer (JISC) beschrieb in seiner Keynote die Herausforderungen im Aufbau von Forschungsdaten-Infrastrukturen und erläuterte die Vorteile der dauerhaften Zugänglichkeit von Daten für die verschiedenen Stakeholder. Nach einer kurzen Vorstellung von JISC erläuterte Dr. Reimer Visionen für den Entwicklungsstatus von Forschungsdaten - Infrastrukturen mit 10 Jahresperspektive.

Abb. 2: Keynote-Sprecher Dr. Thorsten Reimer, JISC

JISC verfolgt in Großbritannien einen 4-Phasenplan für die Entwicklung von Forschungsdaten-Infrastrukturen und dem Management von Forschungsdaten, an dem sich auch das zugehörige vierstufige Förderprogramm orientiert. Momentan befindet sich das Programm in der Konsolidierungsphase, in der es darum geht, gefundene Lösungen zu verfestigen und institutionelle Kapazitäten weiter auszubauen. In diesem Kontext beschrieb er zwölf damit verbundene Herausforderungen und verschiedene Ansätze, diesen Herausforderungen zu begegnen.

Präsentation der DFG-Projekte Radieschen, re3data.org, KomFor, EWIG und BoKeLa

Frank Scholze, KIT, stellte das Projekt re3data.org – Registry of Research Data Repositories vor. Das Projekt hat das das Ziel Daten-Repositorien in einem web-basierten Verzeichnis zu erschließen und so eine Orientierung über bestehende Datensammlungen zu bieten. Zentrales Anliegen von re3data.org ist es, Wissenschaftlern eine Orientierung in der heterogenen Landschaft der Forschungsdaten-Repositorien zu geben. Dabei wird sowohl die Rolle der Erhebenden, die z. B. durch Förderorganisationen oder wissenschaftliche Zeitschriften aufgefordert sind Daten zugänglich machen, als auch die Rolle der Nutzenden, z. B. Forschende, die Daten von Dritten nachnutzen möchten, fokussiert. Weiter soll infrastrukturellen Dienstleistern, wie Datenzentren, Rechenzentren und Bibliotheken, eine Übersicht über die Landschaft der Forschungsdaten-Repositorien gegeben werden. Darüber hinaus kann der Dienst von Förderorganisationen genutzt werden: Fordern diese Mittelempfänger auf, gewonnene Forschungsdaten offen zugänglich zu machen, kann eine Recherche in re3data.org helfen, potenzielle Repositorien zur Speicherung der Forschungsdaten zu finden. Weiterführende Informationen finden sich auf der Projekt-Website (http://www.re3data.org).

Abb. 3: Präsentation „re3data.org“, Frank Scholze, KIT

Tim Hasler, ZIB, präsentierte das DFG-Projekt EWIG - Entwicklung von Workflowkomponenten für die Langzeitarchivierung von Forschungsdaten in den Geowissenschaften. EWIG zielt auf Probleme speziell bei der Datenübergabe an Langzeitarchive. Dazu werden drei Themen beleuchtet. Zum einen werden auf institutioneller Ebene bei den Projektpartnern GFZ und Institut für Meteorologie der FU Berlin Policies entwickelt, die die Grundlage für eine genaue Übergabedefinition zwischen Datenlieferant und Archiv / Repositorium sind. Am Zuse-Institut Berlin liegt der Fokus auf der technischen Qualitätssicherung gelieferter Daten. Der dritte Schwerpunkt der Arbeiten soll das Thema Forschungsdatenmanagement bereits in der fachwissenschaftlichen Ausbildung etablieren. Dazu werden in EWIG Konzepte für Lehr- und Weiterbildungsveranstaltungen entworfen. http://ewig.gfz-potsdam.de/.

Dr. Jochen Klar, AIP, präsentierte das DFG-Projekt „Rahmenbedingungen einer disziplin-übergreifenden Forschungsdateninfrastruktur“, kurz Radieschen. Ziel des Projekts ist die Erstellung einer Roadmap mit Handlungsempfehlungen für eine disziplinübergreifende Infrastruktur für Forschungsdaten in Deutschland.

Sie identifiziert und behandelt die Anforderungen an generische Komponenten einer Infrastruktur und die Vernetzung mit disziplinspezifischen Komponenten. Die Analyse basiert auf einer Bestandsaufnahme mit bestehenden und neuen Projekten sowie Maßnahmen zum Community Building. Zentrale Dimensionen der Analyse sind Technik, Organisation und Kosten sowie die Untersuchung von Querschnittsthemen. Das Projektkonsortium ist ebenso disziplin- und organisationsübergreifend wie die Analyse. Weiterführende Informationen finden sich auf der Projekt-Website unter http://www.forschungsdaten.org/ .

Dr. Michael Diepenbroek, MARUM, präsentierte das DFG-Projekt KomFor – Kompetenzzentrum für Forschungsdaten aus Erde und Umwelt (http://www.komfor.net). Das Kompetenzzentrum ist als Bindeglied zwischen Forschungseinrichtungen, Verlagen, Bibliotheken und einem bestehenden Archivnetzwerk für Daten aus Erd- und Umweltforschung geplant. Allgemeines Ziel ist die nachhaltige Verbesserung von Datenverfügbarkeit und –qualität. Konkret sollen nachhaltige und verlässliche Wege zur Publikation wissenschaftlicher Daten geschaffen werden die den Qualitätsstandards wissenschaftlichen Publizierens entsprechen. Das Kompetenzzentum soll hierfür wissenschaftliche Projekte, Institute, Forschergruppen aber auch einzelne Wissenschaftler in allen Fragen des Datenmanagements begleiten – von der Planungsphase über die Datenerhebung, Qualitätssicherung, Registrierung und Langzeitarchivierung bis zur Publikation der Daten. Weiterführende Informationen zum Projekt auf http://www.komfor.net/.

Abb. 4: Präsentation ‚KomFor', Dr. Michael Diepenbroek, MARUM

Dr. Hans-Joachim Wallrabe-Adams, MARUM, erläuterte in seinem Vortrag die Ziele und aktuellen Ergebnisse des Projekts BoKeLa - Aufbau des Dateninformationssystems für das GESEP Kern- und Probenlager zur Erfassung und Verwaltung von Bohrkernen und Nachweis der Bestände in einem Internetportal (http://www.gesep.de). Das Projekt BoKeLa wurde vom German Scientific Earth Probing Consortium (GESEP e.V.) initiiert. Dem Konsortium gehören 15 geowissenschaftliche Institute mit einer breiten Expertise im Bereich Forschungsbohrungen an Land, im Meer und im Eis an. BoKeLa wird von drei Projektpartnern getragen: der Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), dem Deutschen GeoForschungsZentrum (GFZ) und dem Zentrum für Marine Umweltwissenschaften (MARUM). Ausgangspunkt der Entwicklung eines Bohrkern- und Probenverwaltungssystems ist das Drilling Information System (DIS) das bereits seit Jahren in diesem Bereich eingesetzt wird, aber für die hier geplante Nutzung umfangreicher Anpassungsarbeiten bedarf. Dieses System unterstützt das Datenmanagement bereits an der Bohrung, dann im Labor und dient später der Verwaltung im Kernlager. Alle Kerne und Proben werden mit einem eindeutigen Identifikator versehen, der International Geo Sample Number (IGSN). Damit wird es möglich sein Kerne und Proben dauerhaft identifizierbar und in der Fachliteratur zitierbar zu machen, sowie wissenschaftliche Daten direkt Proben zuzuordnen. Über ein Webportal werden die Kernlager-Informationen nach dem Open Access Prinzip zugänglich gemacht und mit Datenzentren (z.B. PANGAEA) und Publikationen verknüpft.

Vorträge

Die Vorträge des Hauptprogramms spiegelten die einzelnen Stationen im Datenlebenszyklus wieder. Abb. 5 zeigt das Domänenmodell nach Treloar mit integrierter dauerhafter Domäne.

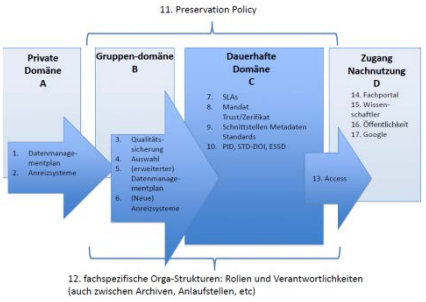


Abb. 5: Domänen-Modell (abgeleitet von Treloar, A. & Harboe-Ree, 2008)

Als ein Beispiel für die Private Domäne erläuterte **Prof. Frederik Tilman (GFZ)** den Datenzyklus in der Wissenschaft aus der Perspektive der Forscher anhand seines Forschungsgebiets Seismologie. Als Herausforderung beschrieb er das Auffinden von früher datierten Datensets, die Verwaltung der Daten, das Vermeiden von Überschreibungen, sowie die Nicht-Reproduzierbarkeit der Daten aufgrund fehlender oder veralteter Software. Eingesetzte Software-Lösungen sind zumeist ‚low-tech solutions'.

Eine Herausforderung sei es, laut Prof. Tilman, die Wissenschaftler dazu zu bewegen, komplexere, aber besser geeignete Lösungen zum informellen Datenaustausch zu verwenden. Prof. Tilmann zeigte persönliche, soziale und technologische Hinderungsgründe auf, die für viele Wissenschaftler dem freien Datenaustausch entgegenstehen. Als mögliche Anreize, um dieses Problem zu beheben, nannte er die gesteigerte Anerkennung von Datenpublikationen als wissenschaftliche Leistungen, die Einführung von Datenpublikationen als Voraussetzung für Publikationen generell, sowie mehrere Anreize, die auf das persönliche Umfeld der Forscher abzielen. Ein wichtiger Punkt sei auch die Nachvollziehbarkeit des zeitlichen Ablaufs sowohl für die Erstellung der Daten als auch für den Einsatz der Software. Er schloss seinen Vortrag mit dem Wunsch nach einfach verfügbaren und leicht auffindbaren Datensets, der Forderung nach wiederverwendbaren und rekonstruierbaren Datenprodukten und der Hoffnung auf bessere Anerkennung für die offene Weitergabe und das Publizieren von Daten.

Dr. Peter Bartelheimer (Universität Göttingen) betrachtete in seinem Vortrag die Datenbereitstellung und Virtuelle Forschungsumgebungen (VFU) als zwei Komponenten der IT-Forschungsinfrastruktur. Anhand des Projekts "Virtuelle Forschungsumgebung für die sozio-ökonomische Berichterstattung" beschrieb er Funktionen der VFU in typischen Workflows der Datennutzung. Das Projekt befindet sich derzeit in der Einführungs- und Entwicklungsphase (2012 – 2014). Anwendungsfall für den Einsatz der VFU ist die Erstellung des Dritten Berichts zur sozio-ökonomischen Entwicklung in Deutschland. Die VFU wird projektbegleitend weiterentwickelt und operativ im Verbund der Projektpartner genutzt. Das Ziel ist die Bereitstellung einer nachhaltigen VFU-Plattform.

Dr. Bartelheimer befasste sich mit der Frage des Transfers der Daten in die VFU, sowie der Abläufe hinter der Datenschnittstelle. Besonderes Augenmerk galt hierbei dem Workflow in der VFU, sowie dem Transfer von Syntax- und Metadateien. Als entscheidend für die Akzeptanz einer VFU beschrieb Dr. Bartelheimer zum einen die Kontrollanforderungen an die datenhaltende Einrichtung, zum anderen die intuitive individuelle Anwendung für den Nutzer. Die VFU sollte eine einheitliche Arbeitsumgebung bieten und es erlauben, schrittweise mehr Daten zu integrieren. Die Forschungsdatenzentren sollten bei der Entwicklung von VFUs und der zugehörigen Nutzungsstudien in die technische und soziale Entwicklung integriert werden.

Reiner Mauer bot in seinem Vortrag einen Blick auf die Persistente Domäne und damit der dauerhaften Bereitstellung von Forschungsdaten. Der Vortrag begann mit einem Überblick über Forschungsdaten in den Sozialwissenschaften und den Aufgaben von GESIS als Infrastruktureinrichtung für die Sozialwissenschaften.

Die Langzeitarchivierung von Forschungsdaten ist ein explizites Ziel von GESIS und als solches in der Satzung verankert. Reiner Mauer positionierte die Aufgaben von GESIS im Kontext des Curation Continuum und beschrieb den Arbeitsablauf bei der Archivierung, der Datenaufbereitung und der Datenbereitstellung. Die Langzeitarchivierung dient der Substanzerhaltung, sowie dem Erhalt von Nutzbarkeit und Interpretierbarkeit. Er betonte, dass das Vertrauen der Nutzer in die dauerhaften Domäne von großer Bedeutung sei – Stichwort Zertifizierung. Herr Mauer erläuterte die Definition und die Aufgaben persistenter Identifikatoren und stellte einige Fragen zu deren Implementierung zur Diskussion.

Als Herausforderungen beschrieb Herr Mauer die immer stärker werdende Bedeutung strukturierender und kooperativer Dienste. Die Bedeutung komplexer Forschungsdesigns und neuer Datenformen wachse, so dass die LZA sich vor der Herausforderung sieht, mit immer komplexeren Forschungsdaten umzugehen. Desweiteren nimmt die Verlinkung bzw. Integration von Daten aus unterschiedlichen Datenquellen und Disziplinen zu. Dies resultiert in höhere Anforderungen an Daten und Metadaten. Rechtliche und Lizenzfragen gewinnen an Bedeutung. Bei der Qualität sind insbesondere Aktualisierung, Versionierung, und autoritative Versionen relevant.

Dr. Hans Pfeiffenberger behandelte in seinem Vortrag Portale und Best Practices und bot damit ein Beispiel für Datennutzung im Kontext der Öffentlichen Domäne. Nach einem blick auf Data Sharing in der Wissenschaft und dem geäußerten Wunsch, über einen singulären Zugangspunkt für möglichst alle wissenschaftlichen Datensätze zu verfügen, ging Dr. Pfeiffenberger näher auf die oftmals sehr unterschiedliche Herkunft der Daten und die Diversität der Datenformen ein. In mehreren Beispielen zeigte er die vielfältigen und verstreut existierenden Datenquellen am Beispiel von Forschungsschiffen und Bojen mit Meßfunktion. Weiterhin stellte er ein Modell zur Klassifizierung und Zuordnung der Daten zu Expeditionen, Forschern und Art der Daten vor. Wichtig sei, alle notwendigen Kontext- und Abstammungsinformationen mit den Daten zu erfassen.

Besonders hervor hob Dr. Pfeiffenberger die Notwendigkeit zur Reproduzierbarkeit der Daten. Er beschrieb den Weg von Rohdaten zu qualitätsgesicherten Primärdaten und daraus abgeleiteten Datenprodukten, und betonte die Notwendigkeit Datenakquise-Praktiken zu harmonisieren.

Workshop-Programm

Die Palette der angebotenen Workshops spiegelte die Bandbreite der von Forschungsdaten-Infrastrukturen angesprochenen Themen wieder. Das Spektrum der Themen reichte von Qualifizierung und Policies über eher anwendungsorientierte Themen wie Software-Lösungen und Datenmanagement bis hin zu rechtlichen Rahmenbedingungen. Die Workshops organisierten sich in drei Themenblöcken „Realisierung“, „Daten“ und „Rahmenbedingungen“. Die Workshops waren organisiert als einstündige Diskussionsrunden mit dem Ziel, offensichtliche Lücken (konzeptionell, organisatorisch, technisch) zu identifizieren, vorbildliche Lösungen aufzuzeigen und eine Vision für das Jahr 2020 (Horizon 2020) zu entwerfen.

Abb. 6: Workshop Impression

Workshop WS 1 – Qualifizierung

Der Workshop wurde eröffnet durch ein Impulsreferat von **Maxi Kindling (Humboldt-Universität zu Berlin)**. In ihrem Impulsreferat erläuterte sie die Integration von Datenmanagement als Thema in fachwissenschaftlichen Studiengängen und gab einen Einblick in die Aus- und Weiterbildungsaktivitäten am Institut für Bibliotheks- und Informationswissenschaft (IBI). Die Workshop-Teilnehmer diskutierten anschließend über die Schaffung eigener Studiengänge für das Datenmanagement, sowie die Integration des Datenmanagements in bibliotheks- und informationswissenschaftliche Studiengänge unter Berücksichtigung des Praxisbezugs. Die Teilnehmer kamen überein, dass die Breite des Qualifizierungsangebote insgesamt erweitert werden müsse. Eine Qualifizierung solle auf den Ebenen des Studiums, des weiterführenden Studiums, sowie als Weiterbildung außerhalb der Universitäten und Fachhochschulen angeboten werden. Der Workshop wurde moderiert von **Prof. Dr. Peter Schirmbacher (Humboldt-Universität zu Berlin)**.

Workshop WS 2A – Datenmanagement

Der Workshop begann mit einem Impulsreferat vom Moderator **Dr. Jens Nieschulze (Universität Göttingen)**. Schwerpunkte der anschließenden Diskussionen waren vor allem Hemmnisse und Anreize für ein gutes Datenmanagement sowie die Nachnutzung der Daten.

Als Hemmnisse wurden unter anderem der erhöhte Zeitaufwand bei der Pflege der Daten für Archivierung und Nachnutzung genannt, die Ungewissheit ob die Daten überhaupt nachgenutzt werden und sich der Aufwand deshalb lohnt, die Form der Daten, welche oftmals nur in analogem Format vorlägen (besonders die Metadaten in Form eines Laborbuches), sowie Geheimhaltungswünsche und –pflichten. Als Anreize wurden die Richtlinien der Forschungsförderer sowie ein „Impaktfaktor“ durch die Publikation von Daten genannt. Allerdings wird ein hoher Nutzen für den Mehraufwand erwartet. Auch erfolgreiche Beispiele werden als ein wichtiger Anreiz gesehen. Lösungen zum Datenmanagement sollten sich nahtlos in die Arbeitsumgebungen der Wissenschaftler integrieren und nachhaltig sein, da sie immer weiterentwickelt werden müssen.

Für die Zukunft wurde der Wunsch nach Datenzentren geäußert, die eng mit bestimmten Communities zusammenarbeiten, miteinander vernetzt sind und eine Weiterbildung im Datenmanagement anbieten. Die Universitäten müssen mehr Routine entwickeln, um mit solchen Datenzentren zusammenzuarbeiten - derzeit werden Datenzentren eher als Zusatzlast empfunden.

Workshop WS 2B – Datenmanagement

Wie kann das Datenmanagement in der privaten bzw. ruppen-Domäne (nach Treloar) besser unterstützt werden? Wie können hier Angebote über punktuelle Lösungen hinaus fruchtbar gemacht und besser koordiniert werden? Welche Aspekte gilt es hier über den der technischen Infrastruktur hinaus zu beachten? Diese Fragen standen im Mittelpunkt des von **Dr. Harry Enke (Leibniz-Institut für Astrophysik, AIP)** moderierten Workshops „Datenmanagement“. Kontrovers diskutiert wurde der Ansatz, Meß- und Beobachtungsdaten bereits bei der Entstehung zu speichern, ohne einen korrigierenden oder erläuternden Eingriff durch den Wissenschaftler abzuwarten. Der Moderator betonte, dass Forschungsdatenmanagement die Begleitung des gesamten Prozesses wissenschaftlichen Arbeitens umfassen müsse und in Datenmanagementplänen festgehalten werden müsse. Auch das Formulieren solcher Pläne könne einen kulturellen Wandel in den Disziplinen beschleunigen, die bislang dem Management der Daten noch nicht die nötige Bedeutung zubilligten.

Workshop WS 3 – Policies

Diskutiert wurde die Vielschichtigkeit von Policies - einerseits als Policies praktischer Natur, welche Handlungsempfehlungen darstellen, andererseits als Policies politischer Natur, welche eine Zielsetzung definieren. Betrachtung in diesem Workshop fanden Policies politischer Natur. In vielen Erklärungen wird davon ausgegangen, dass der Wille zur Wiederverwendung von Forschungsdaten vorhanden ist. Ganz eindeutig ist dies jedoch nicht geklärt, da sich die Gemeinschaft der potentiellen Nachnutzer in zwei Gruppen teilt - den eher observations-orientiert arbeitenden Disziplinen, bei denen die Nachnutzung eine besondere Rolle spielt und den eher experimentell arbeitenden Wissenschaften, bei denen die Beschreibung des Verfahrens eine große Bedeutung hat, um Ergebnisse nachvollziehen zu können. Diskutiert wurde die Frage, ob es legitim sei, eine Veröffentlichung zu fordern und ob Daten in größerem Maß fachübergreifend nachgenutzt werden, da es keine wissenschaftlichen Untersuchungen zu dem Problem gäbe. Eine Nachnutzung von Daten bedeutet immer auch deren Aufbereitung und die Kuratierung der zugehörigen Metadaten - ein teurer Prozess, der zur Differenzierung der Relevanz der Datensätze führt. Der Workshop wurde moderiert von **Dr. Christoph Bruch (Helmholtz Open Access Koordinationsbüro)**.

Workshop WS 4 – Software Lösungen

Der von **Dr. Robert Hauser (FIZ-Karlsruhe)** moderierte Workshop thematisierte Software-Lösungen für Forschungsdateninfrastrukturen (FDI). Bei einer Vorstellung der Teilnehmer zeigte sich, dass viele Institutionen sich gerade in der Evaluationsphase für Software für Forschungsdateninfrastrukturen befinden. In der weiteren Diskussion bestätigte sich die Annahme, dass es keine generelle einzelne Softwarelösung gibt, welche die Anforderungen der unterschiedlichen Fachrichtungen erfüllen kann.

Erkenntnisse des Projektes re3data.org zu existierende Forschungsdatenrepositorien zeigen, dass eine Vielzahl von Lösungen im Einsatz sind, meist Eigenentwicklungen oder Weiterentwicklungen existierender Software. Oft wird der Aufwand von Eigenentwicklungen geringer angesehen als die Integration und Anpassung existierender Lösungen. Wichtige Punkte bei der Entwicklung eigener Software-Lösungen sollten die Zusammenarbeit von Wissenschaftlern und Infrastrukturbetreibern, die Abbildung des gesamten Forschungsprozesses und wissenschaftlichen Workflows in der Software, sowie die Nachhaltigkeit sein. Desweiteren wurde die allgemeine Situation der Entwicklung von Softwarelösungen in Deutschland besprochen. Die Teilnehmer schlugen den Aufbau eines nationalen und zentral gesteuerten Netzwerks von Forschungsdatenzentren vor. Es wurde festgestellt, dass die derzeitigen Akteure und Initiativen (DINI, Wissenschaftsorganisationen, etc.) sich zwar in Netzwerken sammeln, jedoch nicht koordiniert agieren.

Bestehende Initiativen in Deutschland sollten gebündelt werden, um durch gemeinsame Veranstaltungen mit Vorträgen und Experten-Workshops den Austausch zu fördern.

Workshop WS 5A – Datenpublikationen

Der Workshop begann mit einer Präsentation von Moderator **Dr. Michael Lautenschlager (Deutsches Klimarechenzentrum)**, welche den Lebenszyklus der Daten in der Langzeitarchivierung, sowie Daten als Teil virtueller Forschungsumgebungen beschrieb. Laut Moderator gehe der Trend zur Anonymisierung von Datenproduzent und Nutzer. Ziel sei ein interdisziplinäres, verteiltes, selbstbeschreibendes, qualitätsgeprüftes Langzeitdatenarchiv, um Daten ohne Kontakt zum Datenautor nachnutzen zu können. Nachfolgend wurde vor allem über die Thematik der Qualitätssicherung von Daten, sowie der Zertifizierung von Archiven und Prozessen diskutiert.

Workshop WS 5B – Datenpublikationen

Diskutiert wurden die vielen Gesichter der Datenpublikationen, das Thema Qualitätssicherung und Visionen für die Zukunft der Datenpublikationen im Jahr 2020. Gefordert wurden Lektorate für Forschungsdaten, da sich die Bewertung und Qualitätssicherung bislang recht schwierig darstellt. Kritisiert wurden bislang eher unbefriedigende, recht technisch ausgelegte Frameworks für Qualitätskriterien. Desweiteren müsse die Nutzbarkeit und Art der Metadaten bedacht werden. Das Publizieren von Daten wurde als kompliziert bewertet. Moderiert wurde der Workshop von **Roland Bertelmann (Deutsches GeoForschungsZentrum GFZ)**.

Workshop WS 6 – Kosten

Fragestellungen dieses Workshops waren die Kostenermittlung für die Langzeitarchivierung von Forschungsdaten und wie Kosten für das Datenmanagement für einen Projektantrag kalkuliert werden sollten. Desweiteren wurde über potentielle Kostenträger, Preismodelle und noch fehlende Richtlinien zum Datenmanagement für die Förderung von Forschungsaufträgen diskutiert. Moderiert wurde der Workshop von **Dr. Henk Harmsen (DANS)**.

Workshop WS 7 – Persistente Identifier

Der Workshop, moderiert von **Dr. Jens Klump (GFZ)**, wurde durch ein Impulsreferat von **Dr. Janna Neumann (TIB Hannover)** eröffnet. Die Teilnehmer widmeten sich zunächst der Terminologie zur Beschreibung von Identifkatoren. Die Definition des Begriffs Persistente Identifikatoren (PID) ist ausschlaggebend für die Wahl der Anwendungsfelder und die Auswahl des darunter liegenden technischen Dienstes. Die nachfolgende Diskussion gliederte sich in drei Hauptthemen: Die Wahl des Anbieters eines PID-Systems, die Anwendungsfälle von Systemen im Kontext von Forschungsdaten, und die Probleme der Interoperabilität zwischen Systemen und Anbietern.

Eine Klärung der Begriffe war notwendig, um die Frage zu diskutieren, wie noch nicht veröffentlichte Objekte eindeutig referenziert werden, welches System dafür geeignet und ob es überhaupt notwendig ist, für verschiedene Phasen des Lebenszyklus von Forschungsdaten unterschiedliche PID-Systeme zu verwenden. Hinter dieser Frage stehen bisher ungeklärte Anforderungen an Qualität, Vertrauenswürdigkeit und Kontrollierbarkeit der Systeme. Aber auch bei praktischen Fragen der Identität, Granularität und Versionierung digitaler Forschungsobjekte wäre eine Vereinheitlichung der Begriffe hilfreich und würde zur konzeptionellen Klarheit beitragen.

Ein aktuelles Beispiel für unterschiedliche Ansätze für den Betrieb eines PID-Systems ist die Identifikation von Autoren. Hier stehen heute mehrere Systeme zur Auswahl, die teilweise miteinander interoperabel sind. Der Hauptunterschied zwischen den verschiedenen Systemen zur Identifikation von Autoren liegt jedoch in ihrer Organisation und in der Rolle der Verlage und Bibliotheken.

Workshop WS 8A – Daten in Virtuellen Forschungsumgebungen

In diesem Workshop, moderiert von **Dr. Heike Neuroth (Niedersächsischen Staats- und Universitätsbibliothek Göttingen)**, wurden zunächst Definitionen von VFUs besprochen und anschließend Probleme zum Thema VFU aus einer vorgestellten Liste zu priorisiert und Lösungswege besprochen. Im Detail diskutiert wurden der Ablauf des Dialogs zwischen (Fach-)Wissenschaftler und VFU-Betreiber, der Gegensatz von „Einfache Nutzung vs. Alleskönner“, sowie Möglichkeiten zur Sicherstellung der Nachhaltigkeit. Für die Zukunft von VFUs wurde die Wunschvorstellung geäußert, alle Arbeitsabläufe der Forschung so zu integrieren, dass der Wechsel zwischen normaler Arbeitsumgebung und VFU für den Nutzer praktisch nicht merkbar ist. Dafür müsste allerdings die Benutzerfreundlichkeit der VFUs erheblich verbessert werden.

Workshop WS 8B – Daten in Virtuellen Forschungsumgebungen

Zu Beginn des Workshops wurde gemeinsam versucht, Begriffsbestimmungen für Forschungsdaten und Virtuelle Forschungsumgebungen (VFU) zu finden. Die Teilnehmerinnen und Teilnehmer waren sich einig, dass für die Kategorisierung von Forschungsdaten vor allem praktische Gesichtspunkte den Ausschlag geben und dass deren Form, Format und Umfang je nach Wissenschaftsdisziplin sehr unterschiedlich sein können.

Virtuelle Forschungsumgebungen stellen komplexe Systeme dar, die disziplinübergreifend oder disziplinspezifisch verschiedene Tätigkeiten des wissenschaftlichen Arbeitsprozesses abbilden. Dabei sind sie meist kein abgeschlossenes Produkt, sondern ein Framework oder Baukastensystem, das zur Unterstützung der Gesamtheit oder eines Teils des Wissenschaftsprozesses dient. In der weiteren Diskussion wurde die Abbildung der Vorlieben und Bedürfnisse der Nutzer als eine besonders wichtige Anforderung an virtuelle Forschungsumgebungen betont. Bei der institutsübergreifenden Nutzung virtueller Forschungsumgebungen ist es wichtig, wer Zugriff auf die enthaltenen Forschungsdaten haben soll, wer dies festlegt und wie der Zugang organisatorisch und technisch geregelt wird. Wenn Daten öffentlich zugänglich gemacht werden sollen, können Sicherheitsaspekte eine Rolle spielen. Virtuelle Forschungsumgebungen sollten als Dienst flexibel und anpassungsfähig sein um eine möglichst einfache Zusammenarbeit bei der Nutzung der Inhalte zu erreichen. Das Vertrauen der Nutzerinnen und Nutzer ist eine wichtige Voraussetzung für die Annahme virtueller Forschungsumgebungen als Arbeitsmittel und damit auch für die Nachhaltigkeit der entstehenden Strukturen. Bei der Finanzierung müssen gerechte Lösungen für die Umlage der Kosten gefunden werden.

Zusammenfassend wurde festgestellt, dass virtuelle Forschungsumgebungen auch in Zukunft als Abbild des bisherigen Wissenschaftsprozesses ihre Berechtigung haben werden, wenn der Mensch als Nutzer im Mittelpunkt steht. Der Workshop wurde moderiert von **Prof. Dr. Peter Schirmbacher (Humboldt-Universität zu Berlin).**

Workshop WS 9 – Rechtliche Rahmenbedingungen

Der Umgang mit digitalen Forschungsdaten wird häufig von rechtlichen Fragestellungen berührt. Vor diesem Hintergrund widmete sich der Workshop „Rechtliche Rahmenbedingungen“ der juristischen Dimension der Forschungsdaten-Infrastruktur. Moderiert wurde der Workshop von **John H. Weitzmann (Creative**

Commons). Im Kern der Diskussion standen Fragen wie „Wem gehören (meine) Daten?“ und angrenzende Themen wie die Haftung für Folgeschäden bei „falschen Daten”. Der spannende Workshop zeigte, dass noch viele Fragen ungeklärt sind. Aus Sicht der Wissenschaft scheint der Wunsch nach rechtsicheren Verfahren im Umgang mit Forschungsdaten höchste Priorität zu haben. Zum Abschluss der Diskussion wurde noch ein Blick auf den gesetzgeberischen Handlungsbedarf geworfen. Hier wurde u.a. festgestellt, dass bei den bisherigen Diskussionen über eine weitergehende Wissenschaftsschranke der Themenkomplex Forschungsdaten noch wenig Berücksichtigung fand.

Abb. 7: Die Teilnehmer des FDI 2013-Symposiums

Fazit

In seinem Schlusswort fasste **Dr. Jens Klump (GFZ)** die Ergebnisse des Symposiums zusammen. Mit etwa 140 Teilnehmern, 10 Vorträgen und 12 Workshops war es den einladenden Projekten gelungen, eine große Zahl an Experten zu einem intensiven und anregenden Symposium zusammenzubringen. In den Workshops und am Rande der Veranstaltung konnten die Teilnehmer neue Kontakte knüpfen und so die Vernetzung der Akteure verbessert werden. Die große Zahl der Teilnehmer zeigt, dass das Thema Forschungsdaten als wichtiges Thema wahrgenommen wird. In den Diskussionen war zu beobachten, dass die Entwicklung von Forschungsdaten-Infrastrukturen zwar nach wie vor heterogen verläuft, und auch viele Fragen auch weiterhin unbeantwortet bleiben. Die Diskussion zeigte jedoch auch, dass das Thema Forschungsdaten-Infrastrukturen eine konzeptionelle Reife erreicht hat, die mit Konzepten in anderen europäischen Staaten, den USA oder Australien vergleichbar ist.

Insbesondere in der technischen und der konzeptionellen Entwicklung ist eine gewisse Konvergenz zu beobachten. Daneben zeigen sich in anderen Bereichen konzeptionelle Unschärfen, bei denen noch Forschungsbedarf besteht, so z.B. bei den Themen „Qualität“, „Vertrauen“ und „Kosten- und Preismodelle“. Lücken bestehen auch noch bei Datenmanagement-Werkzeugen und deren Integration in die Arbeitsabläufe der Wissenschaftler. Die Teilnehmer bemängelten, dass viele Akteure im Datenmanagement nicht weit genug bekannt seien und auch die Vernetzung innerhalb der Community optimiert werden könnte. Auch sei der Bedarf an Angeboten in den Bereichen Qualifizierung und Beratung besonders ausgeprägt.

Auch bei dieser Veranstaltung wurde deutlich, dass eine Verbesserung des Umgangs mit Forschungsdaten nicht nur eine Frage der unterstützenden technischen Infrastrukturen ist, sondern auch einen kulturellen

Wandel in der Wissenschaft erfordert. Der kulturelle Wandel im Hinblick auf Forschungsdaten bewegt sich in Richtung eines offeneren Umgangs mit diesem Teil der wissenschaftlichen Überlieferung, wie es im Bericht *„Science as an Open Enterprise“* der Royal Society (2012) vorgeschlagen wird. Der Grad der Offenheit wird bestimmt durch die Spannung zwischen Vertrauen in die „Peers“ und Kontrolle über das eigene „Werk“. Neben den sozialen Normen muss auch noch der rechtliche Rahmen für Forschungsdaten weiterentwickelt werden.

Für eine Verbesserung der Situation ist es notwendig, die Erstellung von Daten, Software und Infrastrukturen neben den bisher üblichen Literaturveröffentlichungen als Beitrag im wissenschaftlichen Wertesystem zu verankern. Dies bedingt, dass der Nutzen einer Forschungsdateninfrastruktur für Forscher offensichtlich sein muss. Datenpolicies und die Berücksichtigung dieser Leistungen in den institutionellen Bewertungssystemen können diesen Wandel unterstützen.

Positiv bewerteten die Teilnehmer die Möglichkeit, neue und vorbildliche Lösungen für den Umgang mit Forschungsdaten kennenzulernen. Einige Projekte und vielversprechende Ergebnisse wurden vorgestellt. Durch Evaluation der Ergebnisse und Erfahrungsaustausch kann hier eine Entwicklung angestoßen werden, welche über das Stadium des Experimentierens hinausgeht. Wichtig ist, dass der Austausch zwischen den Akteuren weitergeht und auch Praxisvermittlung einschließt, um den Kreis der Akteure zu erweitern.

Sowohl die Organisatoren als auch die Teilnehmer des Symposiums bewerteten die Veranstaltung als äußerst hilfreich für den Austausch von Ideen, der Generierung neuer Impulse und für die Vernetzung der Akteure untereinander. Über Folgeveranstaltungen wird bereits nachgedacht.

Danksagungen

Wir möchten uns herzlich bei unseren Vortragenden, Workshop-Moderatoren und den fast 140 Teilnehmern bedanken, die mit ihren abwechslungsreichen und interessanten Beiträgen unsere Veranstaltung bereichert und neue Impulse für die Weiterentwicklung von Forschungsdaten-Infrastrukturen gegeben haben. Insbesondere danken wir den Protokollanten der Workshops für ihre Mitschriften:

Dr. Torsten Rathmann, Dr. Daniela Koudela, Dr. Beate Rusch, Damian Ulbricht, Paul Vierkant, Vivien Hollad, Dr. Peter Löwe, Britta Dreyer, Markus Schnalke, Dr. Wolfgang Peters-Kottig, Dr. Paul Schulze-Motel und Heinz Pampel.

Lizenz

Zitationsvorschlag

Schäfer, Leonie; Pampel, Heinz; Klump, Jens; Häsler, Tim (2013), Bericht Symposium „Forschungsdaten-Infrastrukturen (FDI 2013)“, Workshop Report, Helmholtz-Zentrum Potsdam Deutsches GeoForschungs-Zentrum, Potsdam, Germany. [online] Available from: http://dx.doi.org/10.2312/RADIESCHEN_003

Agenda Symposium Forschungsdaten-Infrastruktur (FDI 2013) - Dienstag, 22.01.2013

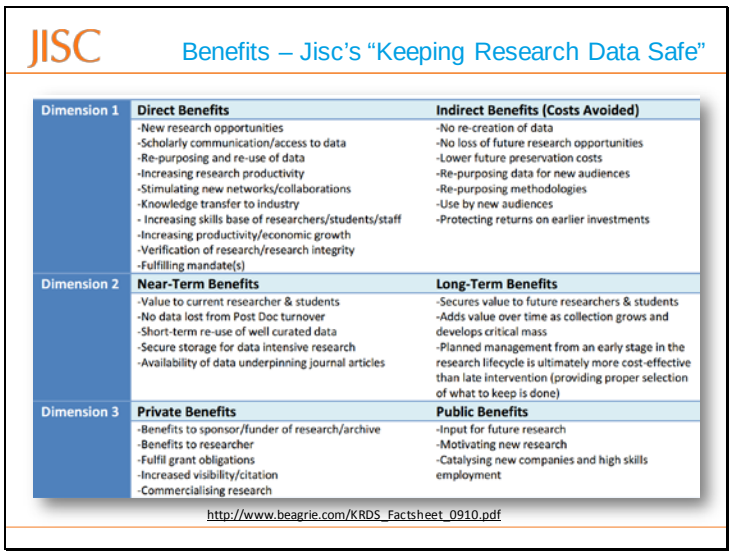
	Hörsaal			
9:30 - 10:00	Registration & Kaffee			
10:00 - 10:30	Eröffnung durch den Vorstand Prof. Dr. Dr. h.c. Reinhard F. J. Hüttl des GFZ - Geleitwort Dr. Winkler-Nees (DFG) - Vorstellung des Tagesablaufs			
10:30 - 11:00	Keynote "Policy" Dr. Torsten Reimer (JISC)			
	Hörsaal	**Tagungsraum 1**	**Tagungsraum 2**	**Tagungsraum 3**
11:00-12:00	**Session 1** Projektpräsentationen "EWIG" Tim Hasler "Radieschen" Jochen Klar(AIP) "KOMFOR" Dr. Michael Diepenbroek "re3data.org" Frank Scholze (KIT) und "BoKeLa"Dr. H.-J. Wallrabe-Adams	**Workshop 1** "Qualifizierung" Prof. Michael Seadle (HUB)	**Workshop 2** "Datenmanagement inkl. Datenpublikationen" ⇨**WS 2A**: Dr. Jens Nieschulze (Uni Göttingen) ⇨**WS 2B**: Dr. Harry Enke (AIP)*	**Workshop 3** "Policies" Dr. Christoph Bruch (Helmholtz OA Büro)
12:00 - 12:15	Kaffee			
12:15 - 13:15	**Session 2** **Vortrag 1:** "Private Domäne 'Die Sicht des Forschers'" Prof. Frederik Tilmann (GFZ) **Vortrag 2:** "Gruppendomäne 'VFUs' " Dr. Peter Bartelheimer (Uni Göttingen)	**Workshop 4** "Software Lösungen" Dr. Robert Hauser (FIZ)	**Workshop 5** "Daten-Publikationen + QS" ⇨**WS 5A:** Dr. Michael Lautenschlager (DKRZ) ⇨**WS 5B:** Roland Bertelmann (GFZ)*	**Workshop 6** "Kosten" Dr. Henk Harmsen (DANS)
13:15 - 14:15	Mittagessen - Kantine			
14:15 - 15:15	**Session 3** Persistente Domäne **Vortrag 1:** "PID, DLZA, Zertifikate" R.Mauer, GESIS **Vortrag 2:** "Zugangsdomäne 'Portale, Best Practices' " Dr. Hans Pfeiffenberger (AWI)	**Workshop 7** "Persistente Identifier" Dr. Jens Klump (GFZ) / Dr. Janna Neumann (TIB)	**Workshop 8** "Daten in VFUs" ⇨**WS 8A**:Dr. Heike Neuroth (SUB) ⇨**WS 8B**: Prof. Peter Schirmbacher*	**Workshop 9** "Rechtliche Rahmenbedingungen" J.H. Weitzmann (Creative Commons)
15:15 - 15:30	Kaffee			
	Hörsaal			
15:30-16:30	Zusammenfassende Berichte aus den Workshops			
16:30 - 17:00	Diskussion und Zusammenfassung - Dr. Jens Klump (GFZ)			

* Raum A 17/10.27 Treffpunkt am Empfang 11:00/12:10/14:10 Uhr

Vortragsprogramm (Materialien)

Keynote – Dr. Torsten Reimer (JISC)

Folie 1



Folie 2

JISC

Who are we?

- We drive innovation in UK education and research, and have been doing so for more than 15 years.
- We are a registered charity and work on behalf of UK higher education, further education and skills to champion the use of digital technologies.
- Jisc offerings and services include:
  - Janet, the UK's high speed network for education and research
  - Jisc Collections, providing digital content for education and research
  - (Advisory) services on the use of digital technology
  - An innovation programme to help our customers make the most of new technologies

Folie 3

Folie 4

JISC

Ten years from now

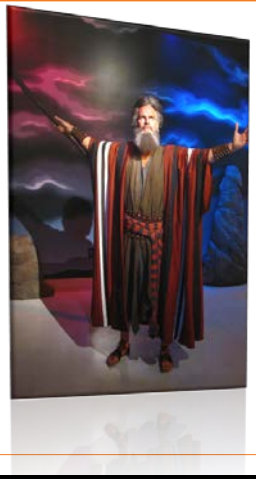
- "In ten years, it is expected that..."
- Leonardo Candela (2011): *Virtual Research Environments.* In: Technological & Organisational Aspects of a Global Research Data Infrastructure - A View from Experts.

http://www.grdi2020.eu/Pages/SelectedDocument.aspx?id_documento=eb0e8fea-c496-45b7-a0c5-831b90fe0045

Folie 5

JISC

"In ten years, it is expected that..."



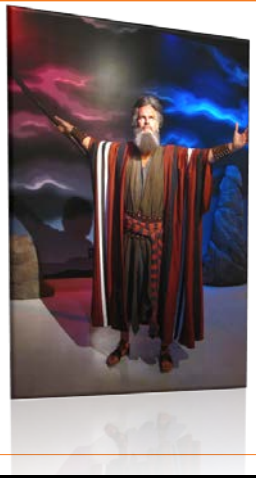
In ten years, it is expected that the trend characterizing science and scientific collaborations discussed above continues, thus becoming the "default" approach for scientific investigations as well as for any societal collaboration-based activity.

*Virtual Research Environments will be **integrated into standard practices** and tools used by communities of practice.*

Folie 6

JISC

"In ten years, it is expected that..."



*The creation and management of Virtual Research Environments **will be a very straightforward process** that relies on specific services – VRE Management Services – **built atop a "global virtual infrastructure"**.*

*The VRE definition phase will guide an authorised actor of an application domain in **characterizing the expected VRE** service in very abstract terms.*

*The VRE **deployment phase will be almost automatic**.*

*The VRE **monitoring / maintenance phase will require little direct human control**.*

*The resulting Virtual Research Environment will be **very flexible and customizable**.*

Folie 7

JISC

"In ten years, it is expected that..."

*Thus Virtual Research Environments creation and management **will become a societal and organisational process rather than a technological one**.*

*Fundamentally, the most important point to have emerged from our study is that VREs need to be conceptualised as **community building projects rather than technology projects**. […] By far the most important challenge faced by VREs is **sustainability**.*

Carusi / Reimer: VRE Landscape Study
http://www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf

Folie 8

JISC

Managing Research Data: Holistic Approach

Folie 9

JISC

Managing Research Data at Jisc

- Understanding the problem (pre-2007-2009)
- Prototyping solutions (2009-11)
- Hardening solutions and building institutional capacity (2011-13)
- Developing elements of national infrastructure (2013+)

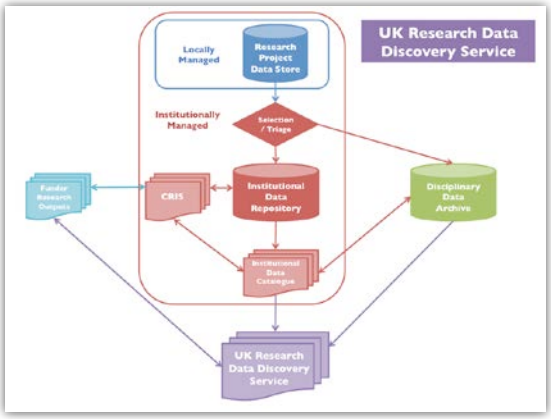
- MRD Phase 1, 2009-2011
  - Infrastructure, Planning, Support, Citation and Publishing, Training
- MRD Phase 2, 2011-2013
  - institutional projects infrastructures and policies;
  - disciplinary best practice, implementation of data management plans
  - customised data management planning tools for institutional use
  - innovative data publication
  - disciplinary training materials

slide 10

Folie 10

JISC

Jisc UK Research Data Discover Service pilot



Folie 11

JISC

Context: RCUK Policies

Full Coverage ◐ Partial Coverage ○ No Coverage

	Policy Coverage		Policy Stipulations					Support Provided			
Research Funders	Published outputs	Data	Time limits	Data plan	Access/ sharing	Long-term curation	Monitoring	Guidance	Repository	Data centre	Costs
AHRC	●	●	●	●	●	◐	○	●	○	◐	○
BBSRC	●	●	●	●	●	●	●	●	●	◐	●
CRUK	●	●	●	●	●	●	●	◐	●	○	○
EPSRC	●	●	●	◐	●	●	●	◐	○	○	●
ESRC	●	●	●	●	●	●	●	●	●	●	◐
MRC	●	●	●	●	●	●	○	◐	●	○	◐
NERC	●	●	●	●	●	●	●	●	●	●	◐
STFC	●	●	●	●	●	●	●	◐	●	◐	○
Wellcome Trust	●	●	●	●	●	●	●	●	●	◐	●

http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

Folie 12

JISC

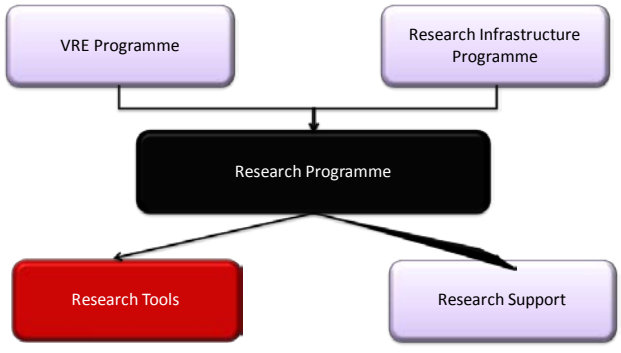
Jisc Virtual Research Environments programme

VRE1	VRE2	VRE3	VRE RI
2004-2007	2007-2009	2009-2011	2010
15 Projects	4 Demonstrators	10 Projects	14 Projects
Technology Focused	User and Research Practice focused	Broadening Use, across institutions and disciplines	Rapid Innovation throughout Research Lifecycle
Experimental	Developmental	Embedding	Rapid Innovation
Diverse design and developmental approached	Unified design and development models	Diverse design – community and challenge driven	'Scratching itches', solution driven
Standalone solutions	Integrated pilots	Tools, frameworks and interoperability	Technical Solutions, Business Community Engagement
Collaboration			
Large and small scale research			
Single and Multi-disciplinary Research			

Folie 13

JISC

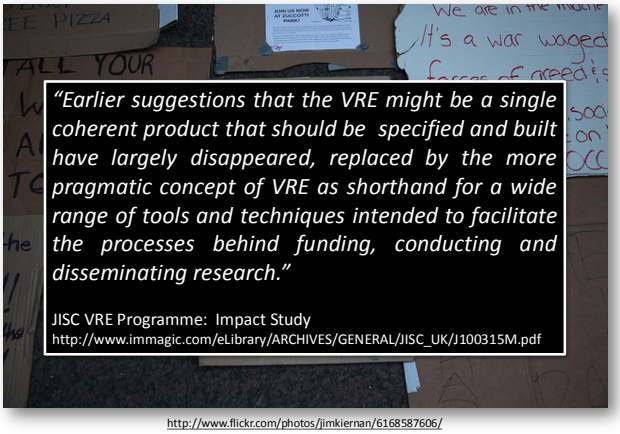
Jisc Research Programme, 2011-



Folie 14

JISC

VRE to Research Tools



http://www.flickr.com/photos/jimkiernan/6168587606/

Folie 15

JISC

VRE Evaluation

VREs still in operation?	
Uptake of code beyond the project?	
Interoperable with other developments?	
New collaborations?	
New research facilitated?	
Sustainable technology/code?	
Sustainable business model?	
Skills/capacity gained?	
Lessons learned?	

Folie 16

JISC

12 Challenges

Business Models	Service vs Research	Cost	Staff Roles, Careers
Support	Skills	Uptake	Institutional Policies
Interoperability	Law	Students	Usability

Folie 17

JISC

Service vs Research / Uptake

Folie 18

JISC

Staff Roles, Careers

Folie 19

JISC

Law

- Jisc study into *The Value and Benefits of Text Mining*
- "the market intervention of copyright, originally intended to protect creative producers, is itself becoming a barrier to new creative production and may be inhibiting new knowledge discovery and innovation"

http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx

Folie 20

JISC

Interoperability

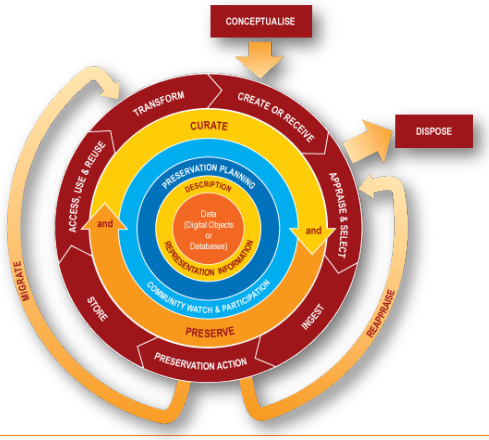
Folie 21

JISC

Skills / Support

The Digital Curation Centre (DCC) works on building capacity, capability and skills for research data management across the UK's higher education research community.

www.dcc.ac.uk



Folie 22

JISC

Students – Researchers of Tomorrow

- "This apparent and **striking dependence on published research resources** implies that, as the basis for their own analytical and original research, **relatively few** doctoral students in social sciences and arts and humanities are using **'primary' materials** such as newspapers, archival material and social data. In sciences, few may be drawing on **large datasets**."
- "Current institutional engagement with open web and **Web 2.0 technologies** does not convince the majority of Generation Y doctoral students of the credibility of using such applications in a research setting, and reinforces their feeling that actively using, for example, social media and online forums in research **lacks legitimacy**."

http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow.aspx

Folie 23

JISC

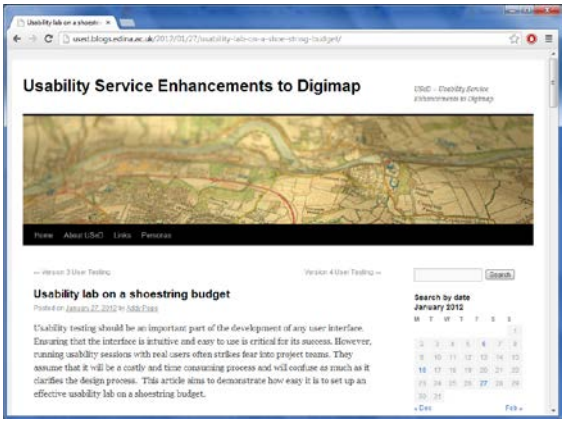
Institutional Policies

http://www.flickr.com/photos/rpenalozan/5367289745/ CC BY NC SA 2.0

Folie 24

JISC

Usability



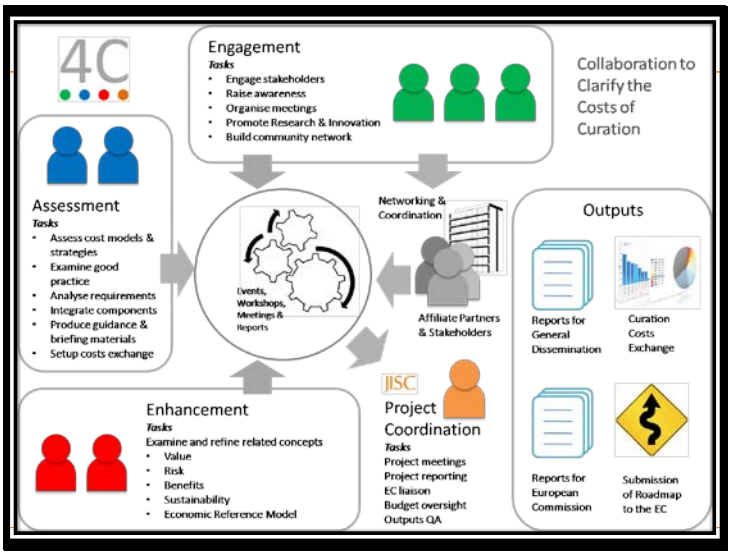
http://used.blogs.edina.ac.uk/2012/01/27/usability-lab-on-a-shoe-string-budget/

Folie 25

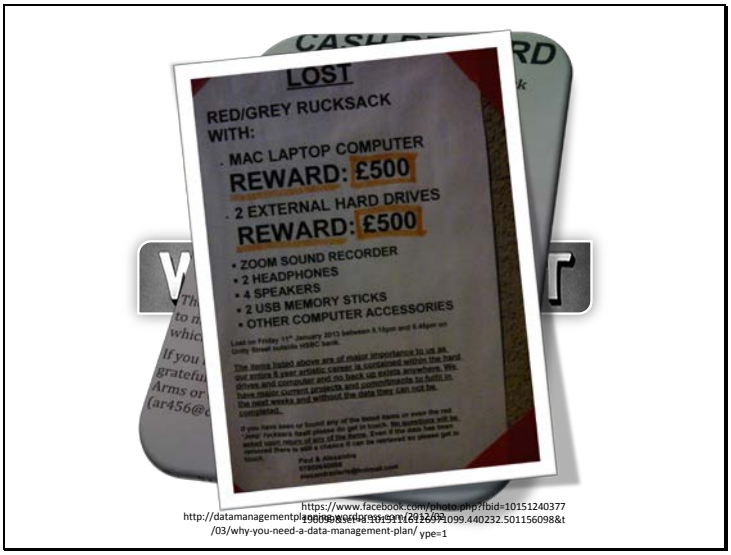
JISC

Business Models

Folie 26



Folie 27



http://datamanagementplan.wordpress.com/2012/02/03/why-you-need-a-data-management-plan/

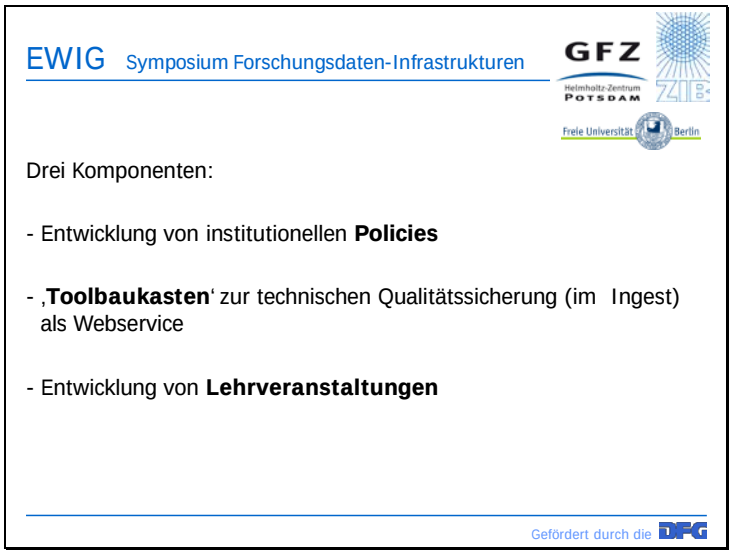
https://www.facebook.com/photo.php?fbid=10151240377100099&set=a.10151240371099.440232.501156098&type=1

Folie 28

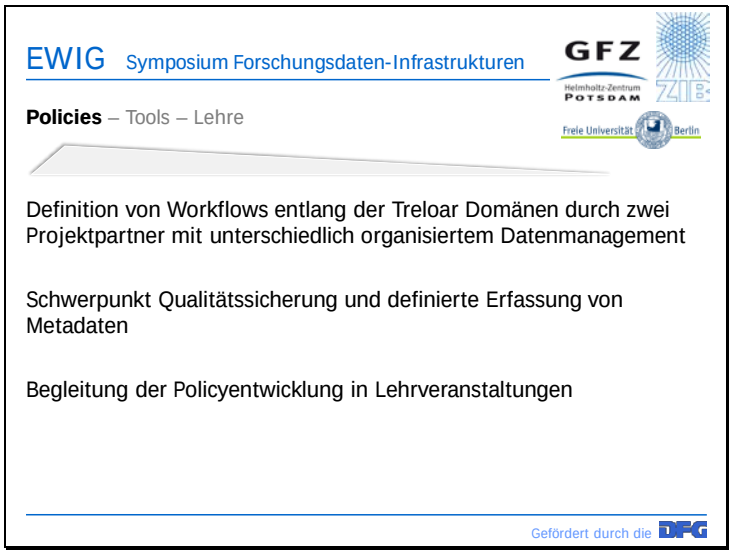
Session 1 - Projektpräsentationen

DFG-Projekt EWIG – Tim Hasler (ZIB)

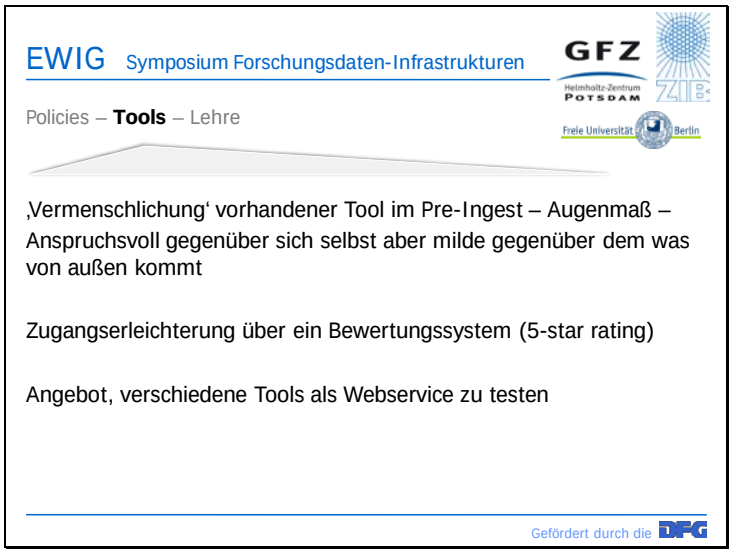
Folie 1



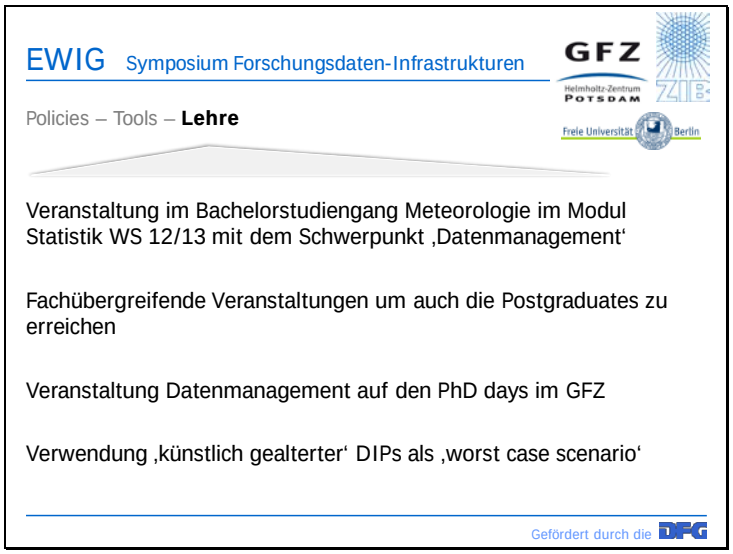
Folie 2



Folie 3



Folie 4



Folie 5

Folie 6

DFG Projekt Radieschen – Dr. Jochen Klar (AIP)

Folie 1

Hintergrund

- Es besteht Konsens über die herausragende Bedeutung des Umgang mit Forschungsdaten
- Diverse Initiativen resultierten in einer Vielzahl von verschiedensten Projekten
- Anforderungen zum Teil projekt- oder disziplinspezifisch, aber auch disziplin-übergreifend

Folie 2

Projektziel

- Bestandsaufname und Analyse der bestehenden und geplanten Projekte im Bereich der Forschungsdateninfrastruktur
- Erstellung einer Roadmap mit Handlungsempfehlungen für eine disziplinübergreifende Forschungsdateninfrastruktur in Deutschland.

Folie 3

Förderung & Projektpartner

- 2 Jahre Laufzeit, bis April 2013
- Förderung durch DFG (LIS)
- 6 Partner aus unterschiedlichen Disziplinen und Wissenschaftsorganisationen

ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen

MAX-PLANCK-INSTITUT FÜR PSYCHOLINGUISTIK

SUB Niedersächsische Staats- und Universitätsbibliothek Göttingen

AIP Leibniz-Institut für Astrophysik Potsdam

GFZ Helmholtz-Zentrum Potsdam

DKRZ Deutsches Klimarechenzentrum

Folie 4

1. **Bestandsaufname**
   - Konzeption und Durchführung von 9 Experteninterviews
2. **Technik**
   - Anforderungen und Entwicklungsbedarf
   - Dokumentation vorbildlicher, disziplinspezifischer Lösungen
3. **Organisation**
   - Akteure und Strukturen in den einzelnen Disziplinen
   - Einbettung des Forschungsdatenmanagements
4. **Kosten**
   - Kosten des Forschungsdatenmanagements
   - Ableitung generische Faktoren und Risikoabschätzungen
5. **Community**
   - Website www.forschungsdaten.org
   - Workshop 17. 04. 2012, FDI 2013
6. **Synthese**
   - Zusammenfassung der Ergebnisse
   - Entwicklung von Handlungsempfehlungen

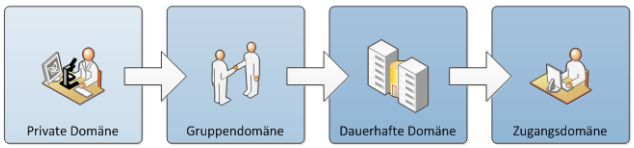
Folie 5

Bestandsaufname

- Interviews mit verschiedenen Akteuren Bereich Forschungsdateninfrastruktur.
- Fokus auf:
  - Projekte der Ausschreibung „Informationsinfrastrukturen für Forschungsdaten“
  - INF Projekte in den SFB und TRR der DFG
- Literaturrecherche

Folie 6

Domänenmodell



- **Private Domäne**
  - Datenerzeugung und Datenverarbeitung in einer Hand.
- **Gruppendomäne**
  - Eine Kollaboration arbeitet auf einem gemeinsamen Satz von Daten.
- **Dauerhafte Domäne**
  - Langzeitarchivierung ohne direkten Bezug zum erzeugenden Akteur.
- **Zugangsdomäne**
  - Veröffentlichung der Daten für eine Gruppe oder die Öffentlichkeit.

Folie 7

Technik

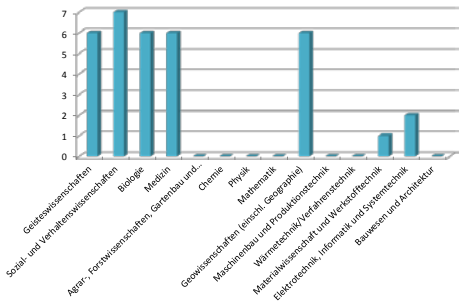
- Höchst unterschiedliche Anforderungen, selbst innerhalb von Disziplinen. Workflows sind stark Disziplinabhängig.
- Manuelle Arbeitsschritte (z.B. zur Qualitätskontrollen) noch verbreitet.
- Der Großteil der verwendeten Software sind Eigenentwicklungen, vor allem im Bereich Datenanalyse.
- Disziplinübergreifender Bedarf ergibt sich hauptsächlich bei Fremdsoftware (z.B. im Bereich der Virtualisierung).

Folie 8

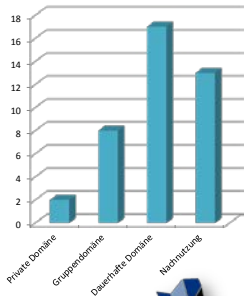
Organisation

DFG Projekte „Informationsinfrastrukturen für Forschungsdaten"

Verteilung nach Fachgebieten:



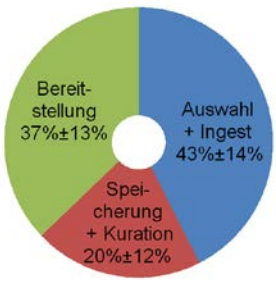
Verteilung nach Domänen:



Folie 9

Kosten

- Belastbare Zahlen sind schwer zu erhalten und zu vergleichen.
- Kosten von Archivierung
  - Personalaufwand für die einzelnen Schritte der Langzeitarchivierung
- Risiken der Nicht-Archivierung
- Risiken der Archivierung



Folie 10

Synthese

- Überblick über den aktuellen Status in der Entwicklung der Forschungsdateninfrastruktur
- Identifikation von Defiziten und Entwicklungsbedarf
- Entwicklung von Zukunftsszenarien
- Erarbeitung von Handlungsempfehlungen

Folie 11

www.forschungsdaten.org

Folie 12

KomFor

Centre of Competence for Research Data
in the Earth & Environmental Sciences

Michael Diepenbroek
PANGAEA® / MARUM, University Bremen

Potsdam 2013 www.komfor.net

Folie 1

Partner

- World Data Center for Remote Sensing of the Atmosphere (WDC-RSAT)
- World Data Center on Climate (WDC-C)
- Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum (GFZ)
- PANGAEA® - Data Publisher for Earth & Environmental Science
- Technische Informationsbibliothek (TIB)

Potsdam 2013 www.komfor.net

Folie 2

Ziele

- Aufbau eines Kompetenzzentrum
  - Bindeglied zwischen wissenschaftlichen Einrichtungen, Verlagen, Bibliotheken & Datenarchiven
  - Vergleichbar ANDS oder DCC
- Publikation wissenschaftlicher Daten

Potsdam 2013 www.komfor.net

Folie 3

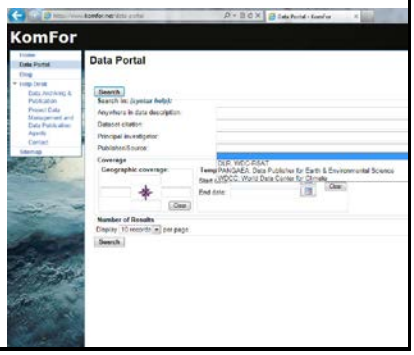
Arbeitspakete

Potsdam 2013 www.komfor.net

Folie 4

Mandantenfähige Service Plattform

- Helpdesk, Redaktionssysteme und Wissensbasis
- Datenportal



Potsdam 2013

Folie 5

Datenpublikation

- Organisatorische Voraussetzungen und Zusammenhänge (Workflows)
- Modelle für ein Peer-Review wissenschaftlicher Daten

Potsdam 2013 www.komfor.net

Folie 6

Katalog- und Registrierungsdienste

- Datenkatalogdienste (cross-linking)
- Bibliometrische Dienste
- Zertifizierung und Akkreditierung neuer Datenarchive / Publikationsagenten

Folie 7

Geschäftsmodelle

- Kostenmodelle für Langzeitarchivierung & Datenpublikation
  - Open Access
- Service Modell
  - Projektdatenmanagement

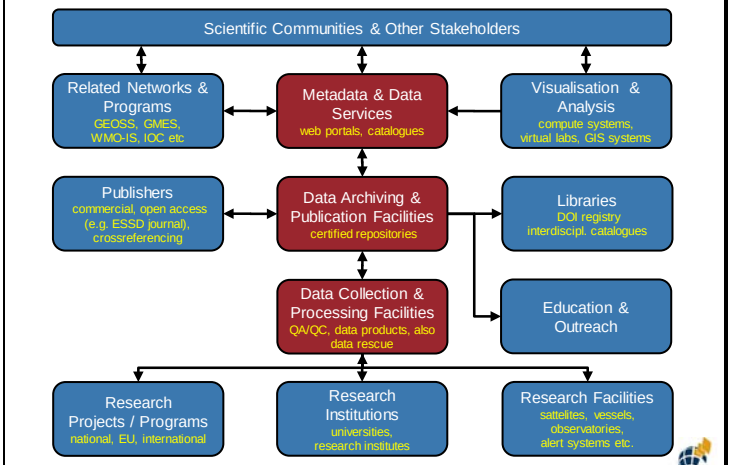
Folie 8

Data publishing - building blocks

- DataCite -> DOI Registry for scientific data ->
- Collaborations with science publishers (Elsevie Viley, Thompson Reuters etc.)
  - ✓with science journals
  - ✓linking & crossreferencing data & articles
- Thomson Reuters Data Citation Index (2012) ->
- ORCID registry for researchers (2012) ->
- ICSU World Data System (WDS) for long-term data stewardship and publication

Folie 9

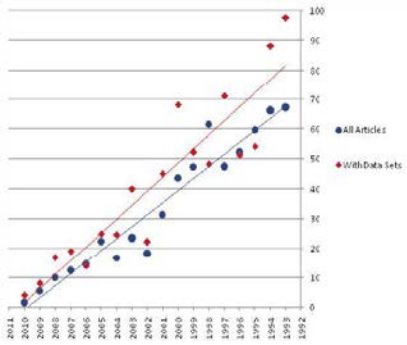
ICSU WDS - roles & relations in a federated system



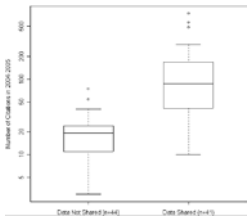
Folie 10

Bibliometrics

35% to 69% more citations



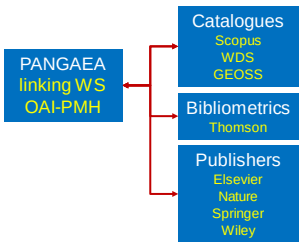
courtesy of Jon Sears (AGU)



Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

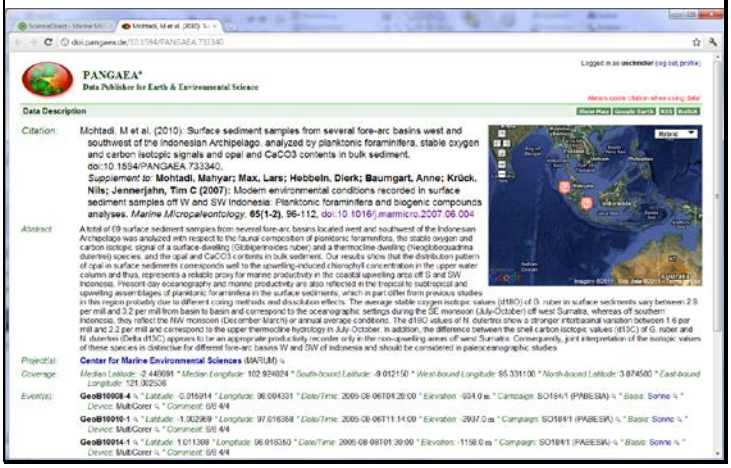
Folie 11

Linking infrastructure



Folie 12

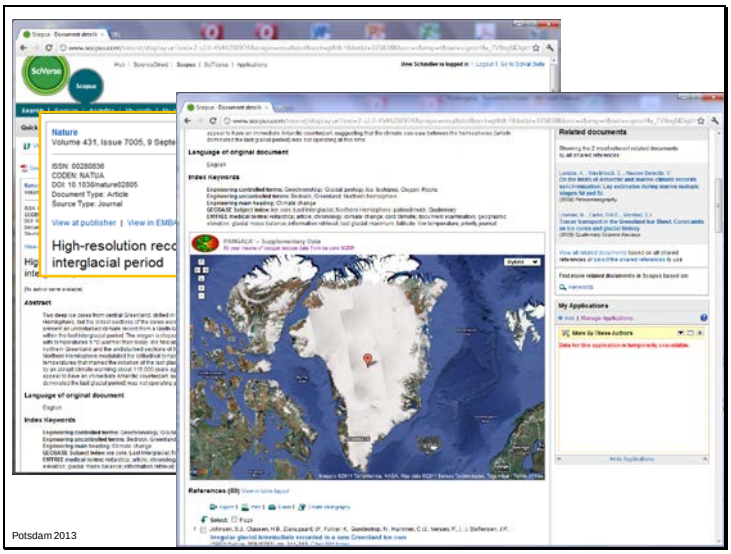
Data Publishing – Cross-referencing



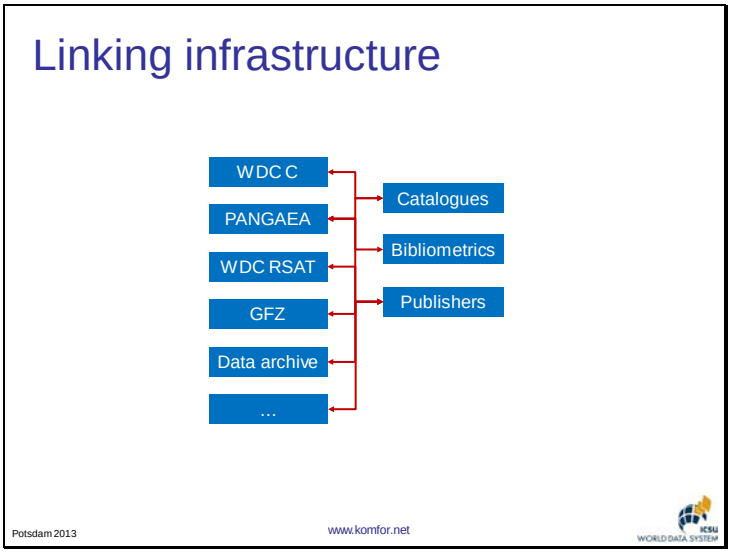
Folie 13

Data Publishing – Cross-referencing

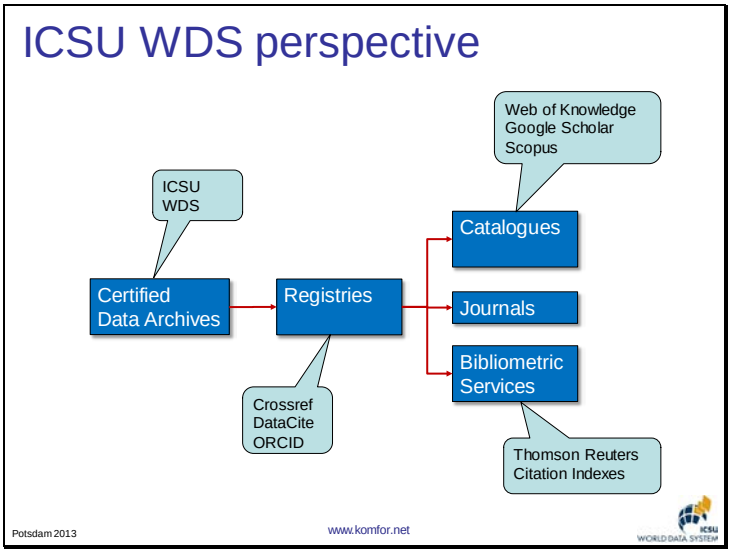
Folie 14



Folie 15



Folie 16



Folie 17

DFG-Projekt re3data.org – Frank Scholze (KIT)

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

re3data.org

Registry of Research Data Repositories

Frank Scholze, Hans-Jürgen Goebelbecker, KIT-Bibliothek
Heinz Pampel, GFZ Potsdam

Symposium „Forschungsdaten-Infrastrukturen“, Potsdam, 22.01.2013

GFZ Helmholtz-Zentrum Potsdam · KIT Karlsruhe Institute of Technology · DINI · DFG Deutsche Forschungsgemeinschaft

Folie 1

Forschungsdaten-Repositorien

- EC: ICT infrastructures for e-science

  „The landscape of data repositories across Europe is fairly heterogeneous, but there is a solid basis to develop a coherent strategy to overcome the fragmentation and enable research communities to better manage, use, share and preserve data.“


EN EN

European Commission. (2009). ICT infrastructures for e-science. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. COM(2009) 108 final.
Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0108:FIN:EN:PDF

GFZ · KIT · re3data.org

Folie 2

Ziele

- Aufbau eines “Registry of Research Data Repositories”
- Weltweites, disziplinübergreifendes, web-basiertes Verzeichnis von Forschungsdaten-Repositorien
- Struktur- und Inhaltsanalyse der Angebote
- Orientierung für Wissenschaftler, Förderer, Verlage, Infrastruktur
- Vernetzung und Standardisierung von Forschungsdaten-Repositorien
- Förderung der “culture of sharing”

GFZ · KIT · Symposium „Forschungsdaten-Infrastrukturen“ Potsdam, 22.01.2013 · re3data.org

Folie 3

Aktueller Stand

- Erste Version von re3data.org seit Dezember 2012 online.
  - ca. 120 voll erschlossene Forschungsdaten-Repositorien
  - ca. 210 Forschungsdaten-Repositorien in Bearbeitung
    -> Basisdatensatz sichtbar
  - Die einfache Suche kann durch diverse Filtermöglichkeiten verfeinert werden
- Die Dateneingabe und –haltung erfolgen im XML-Format.
  - Dazu wurde ein eigenes XML-Schema entwickelt.

Folie 4

Vokabular

- Vierkant, P. et al. (2012). Vocabulary for the Registration and Description of Research Data Repositories. Version 2.0. doi:10.2312/re3.002

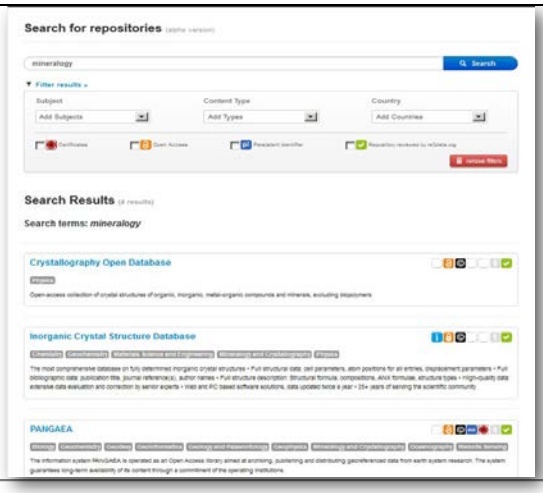
Folie 5

Erschließung

- Erschließung durch Autopsie (review)
  - 37 Hauptkriterien
  - 22 weitere Unterkriterien
  - 25 kontrollierte Vokabulare
- Icons als einfache Außendarstellung
  - Mehrwert
    für Repositorien-Betreiber
    und wissenschaftliche Nutzer

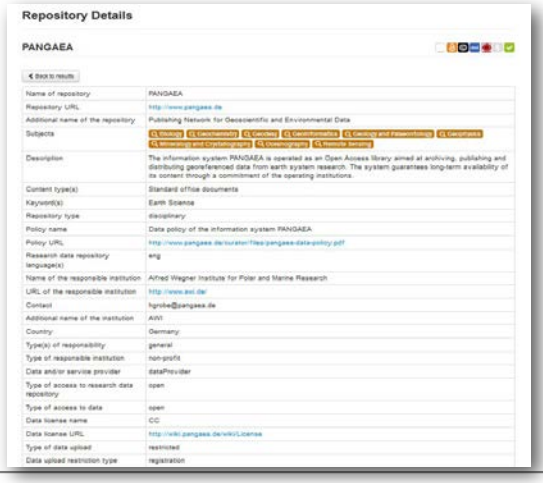
Folie 6

Screenshot: Ergebnisliste



Folie 7

Screenshot: Vollanzeige



Folie 8

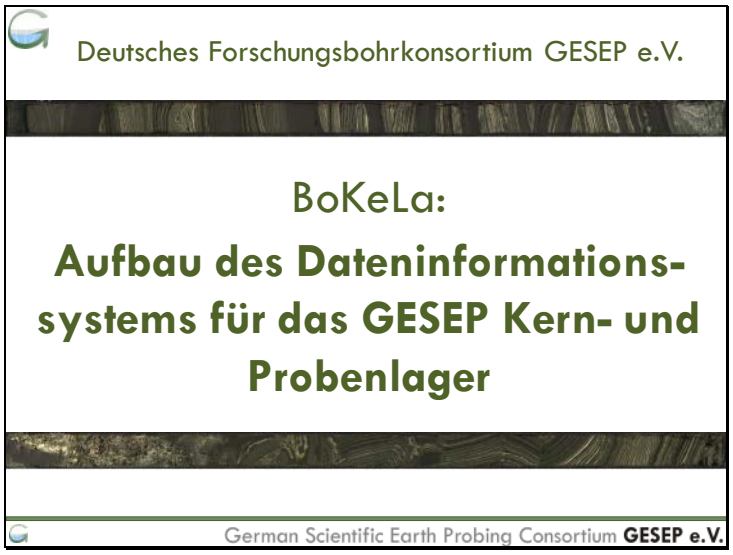
Nächste Schritte

- Entwicklung der Webformulare
- Aufbau eines Redaktions-Workflows
- Einrichtung von Import- und Export-Schnittstellen (XML, Dublin Core, Statistiken)
- Öffentlichkeitsarbeit und Vernetzung
  - Dialog mit diversen Akteuren
  - Daten werden nachnutzbar sein
  - DINI-Wiki zum Thema im Aufbau
  - Mailingliste
  - Twitter

Folie 9

Folie 10

DFG-Projekt BoKeLa – Dr. H.-J.Wallrabe-Adams



Folie 1

„Aufbau des Dateninformationssystems für das GESEP Kern- und Probenlager“

DFG-Vorhaben im Rahmen **LIS*** Förderprogramm:
„Informationsmanagement“

Projektstart August 2011 für die Dauer von drei Jahren
initiiert durch GESEP

*Wissenschaftliche Literaturversorgungs- und Informationssysteme (*LIS*)

Folie 2

Was ist GESEP?

Kooperation 15 geowissenschaftlicher Institute, die

- im Bereich Forschungsbohrungen (Meer, Land, Eis) stark engagiert sind
- breite Expertise und Infrastruktur vernetzen
- gemeinsam neue Verfahren und Modelle entwickeln

Folie 3

Drei Projektpartner – drei Aufgabenfelder

BGR (Bundesanstalt für Geowissenschaften und Rohstoffe):

Aufbau und Betrieb eines Bohrkernlagers am Standort Berlin-Spandau für kontinentale Kerne ohne Kühlung (Festgesteinskerne)

GFZ (Deutsches GeoForschungsZentrum):

Entwicklung eines Bohrkern- und Probenverwaltungssystems

MARUM (Zentrum für Marine Umweltwissenschaften, Bremen):

Betrieb eines Bohrkernlagers in Bremen für Kerne, die der Kühlung bedürfen (+4°C) (See-Sedimente)

Entwicklung eines Web-Portals mit entsprechenden Web-Diensten für die Integration der beiden Standorte Berlin-Spandau und Bremen sowie weiterer wissenschaftlicher Bohrkernlager im Umfeld von GESEP

Folie 4

Kernarchiv für kontinentale Kerne

2009 Aufruf zur Einrichtung des Kernlagers

Einreichen eines gemeinsamen Konzepts von MARUM und BGR

2010 Start des Kernlagers
erste „Sampling Party“ am MARUM durch das Lake Van-Projekt, Türkei

09.2012 Eröffnung des nationalen Bohrkernlagers am BGR-Standort in Berlin-Spandau

Folie 5

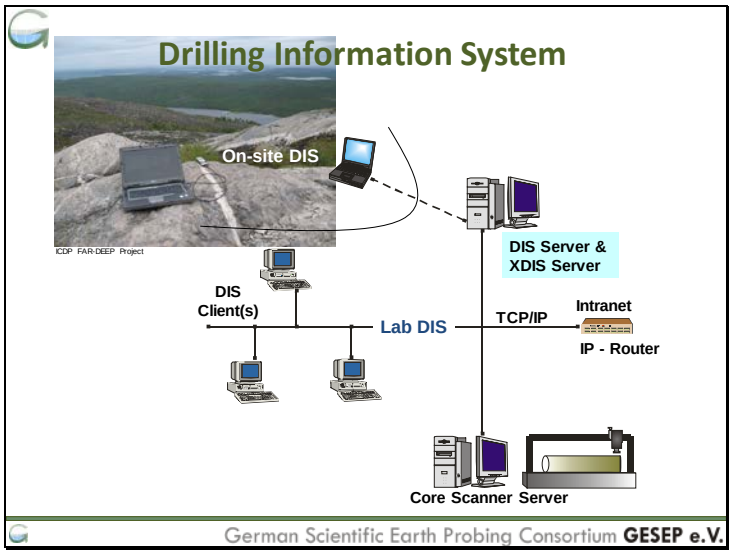
GFZ: Entwicklung eines Bohrkern- und Probenverwaltungssystems

➡ **Drilling Information System (DIS)**

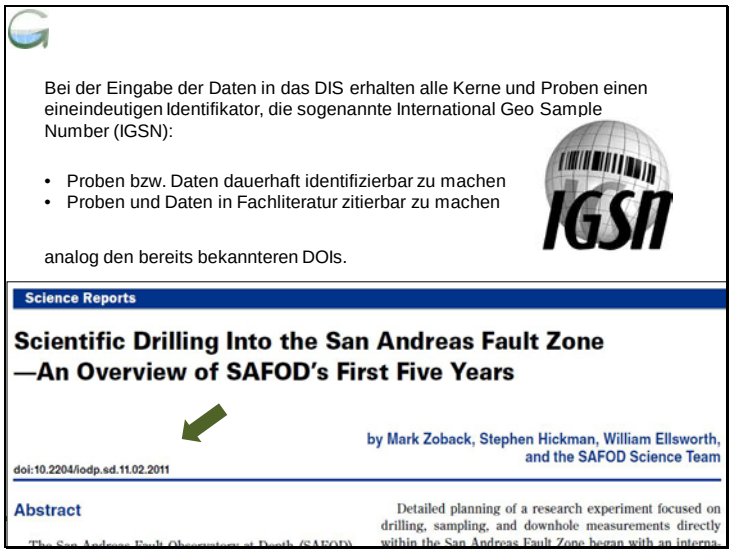
- zur Unterstützung des Datenmanagements an der Bohrung, im Labor und später zur Verwaltung im Kernlager
- dient der Dokumentation und Erfassung von
  - **Basis Daten** (Expedition, Lokation,…)
  - **Messungen und Berichten** (lithologische Beschreibung,…)
  - **Probenverwaltung**
- zur Bereitstellung einer
  - **gemeinsamen Referenz für alle beteiligten Wissenschaftler des Projektes**
- so früh wie möglich einsetzen – bereits an der Bohrung
- auch um nicht-synchronisierte und nicht-authorisierte Datensätze zu vermeiden

Folie 6

Folie 7

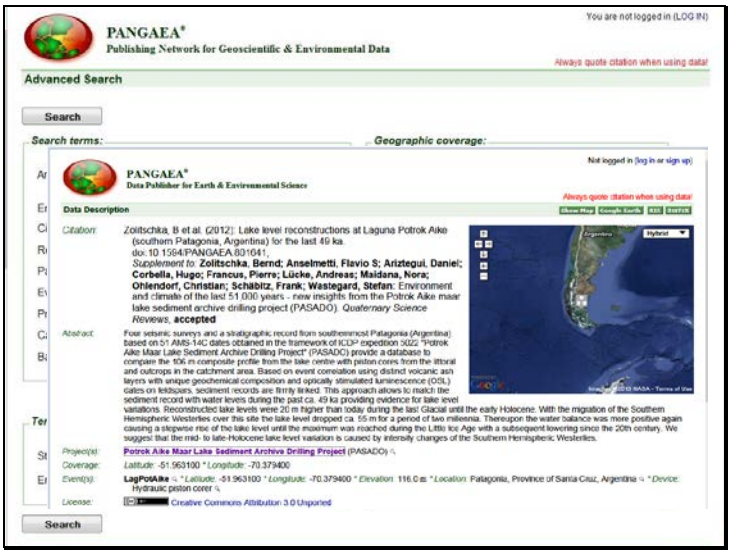


Folie 8



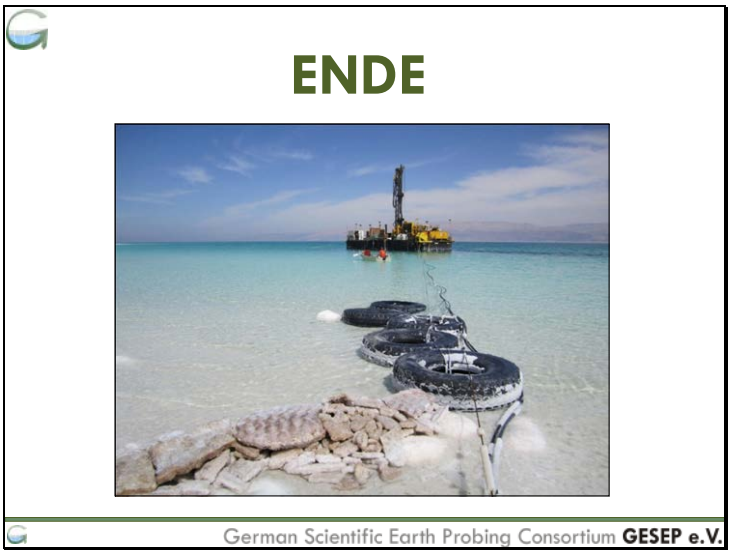
Folie 9

Folie 10



Folie 11

Folie 12



Folie 13

Session 2 – Data Curation Continuum – Teil 1

Die Vorträge des Hauptprogramms spiegelten die einzelnen Stationen im Datenlebenszyklus wieder. Abb. 8 zeigt das Domänenmodell nach Treloar[1] mit integrierter dauerhafter Domäne.

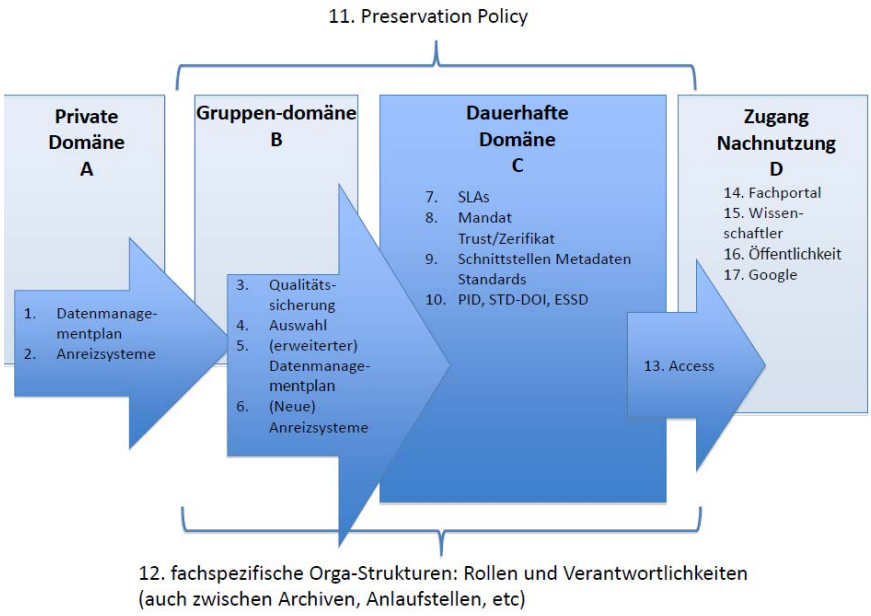
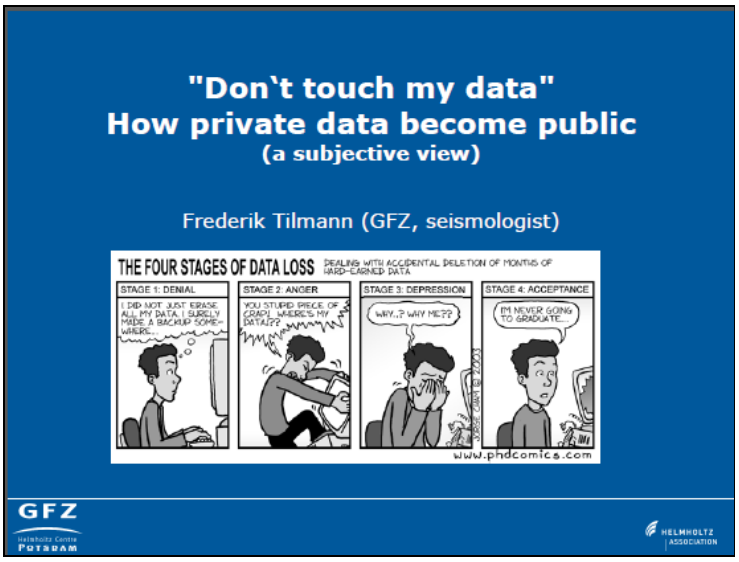


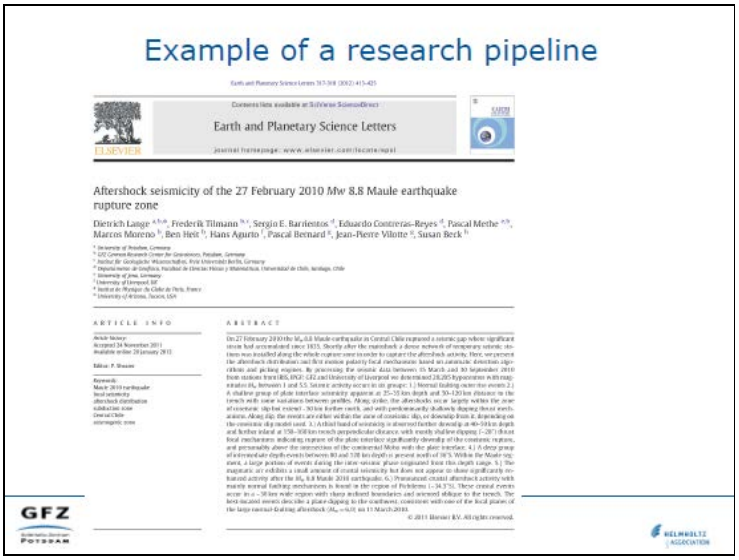
Abb. 8: Domänen-Modell (abgeleitet von Treloar, A. & Harboe-Ree, 2008)

[1] Treloar, A. & Harboe-Ree, C., 2008. Data management and the curation continuum. How the Monash experience is informing repository relationships. VALA2008 14th Biennial Conference. Melbourne, Australia 5.-7. Feb. 2008

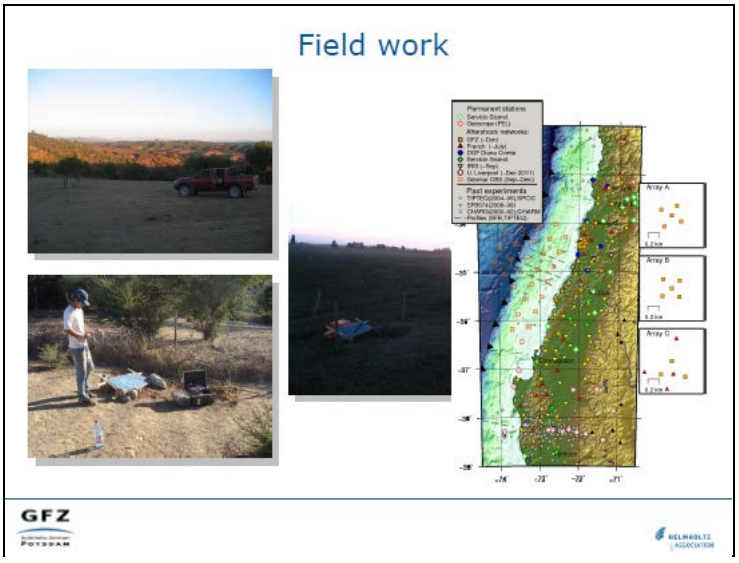
„Private Domäne ‚Die Sicht des Forschers'" – Prof. Frederik Tilman (GFZ)



Folie 1

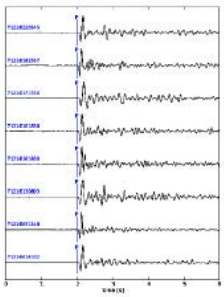


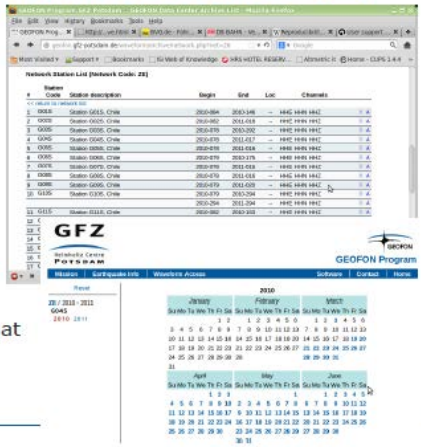
Folie 2



Folie 3

Initial Processing and metadata setting





Archival in standard format with established access tools (GEOFON)

GFZ

Folie 4

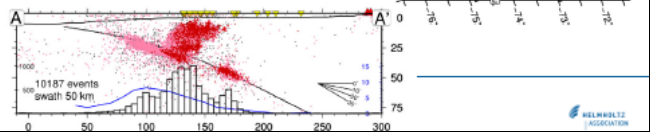
Earthquake bulletin

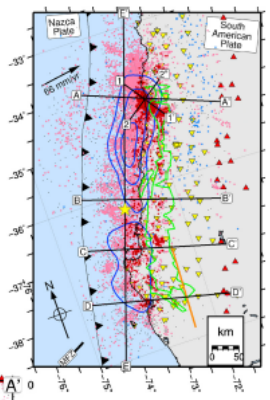
- Travel time picking
- Location in 1D model
- Earthquake bulletin

```
2010  315 0042 50.0 L -34.367 -72.124 24.8   XXX 16 0.4 4.2LKEX 4.0WEXX        1
GAP=191         0.71      2.8     7.4  4.1  -0.1312E+02  0.1399E+02 -0.2639E+01E
ACTION:SPL 11-10-14 20:00 OP:CD   STATUS:              ID:20100315004122      I
STAT SP IPHASW D HHMM SECON CODA AMPLIT PERI AZIMU VELO AIN AR TRES W  DIS CAZ7
QO3B HZ   P      0MU 0042 54.68                          152    -0.2630 10.6 112
QP16 HZ   P      0MU 0042 56.81                          125     0.2630 26.2  58
QP17 HZ   P      0MU 0042 57.99                          113     0.0710 37.2 146
QP12 HZ   P      0MU 0042 58.10                          113    -0.0330 37.9  88
Q27B HZ   P      2MU 0043 02.25                          100     0.42 5 64.0  50
QC12 HZ   P      0MU 0043 05.20                           78    -0.0410 87.4  77
QP14 HZ   P      0MU 0043 06.35                           78     0.1410 94.1 171
QP59 HZ   P      2MU 0043 19.12                           73    -0.71 5  185  34
QP04 HZ   P      4MU 0043 20.47                           73     0.11 0  191  45
QP03 HZ   P      2MU 0043 22.90                           73     0.21 5  208  18
QP11 HZ   P      1MU 0043 25.31                           73    +0.21 5  228 186
QP21 HZ   P      0MU 0043 25.48                           73    -0.0410  228 186
QP01 HZ   P      4MU 0043 27.22                           73     1.36 0  229  29
QC01 HZ   P      4MU 0043 29.77                           73    -0.50 0  260  26

2010  315 0158 12.7 L -35.120 -72.675 37.9   XXX  8 1.0 4.3LKEX 4.0WEXX        1
GAP=211         3.15      16.1   103.3 90.8 -0.3868E+03  0.8751E+04 -0.4130E+02E
```

→ Write first paper focussing on seismotectonics!



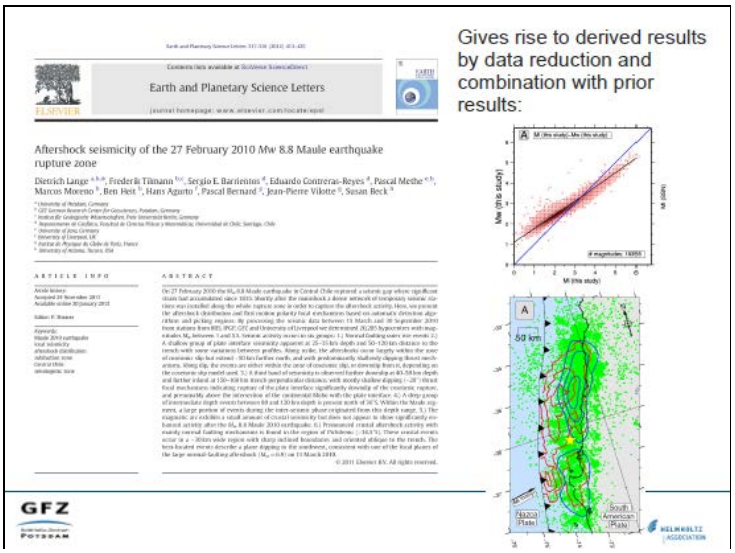


GFZ

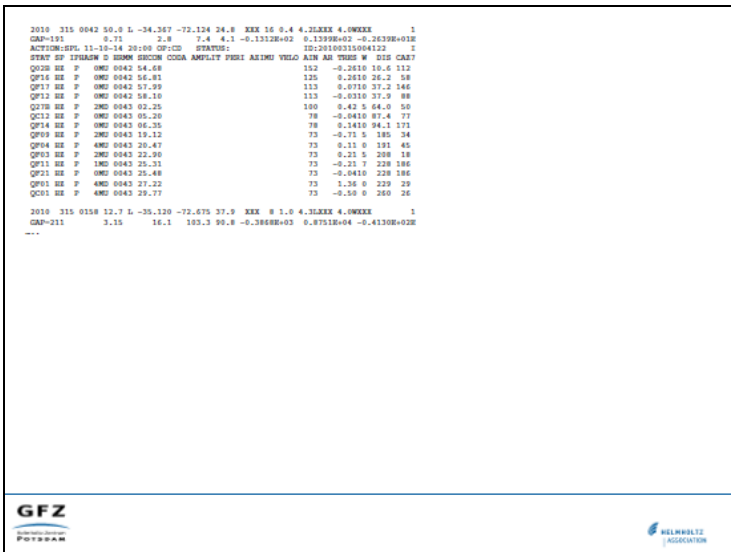
Folie 5

GFZ

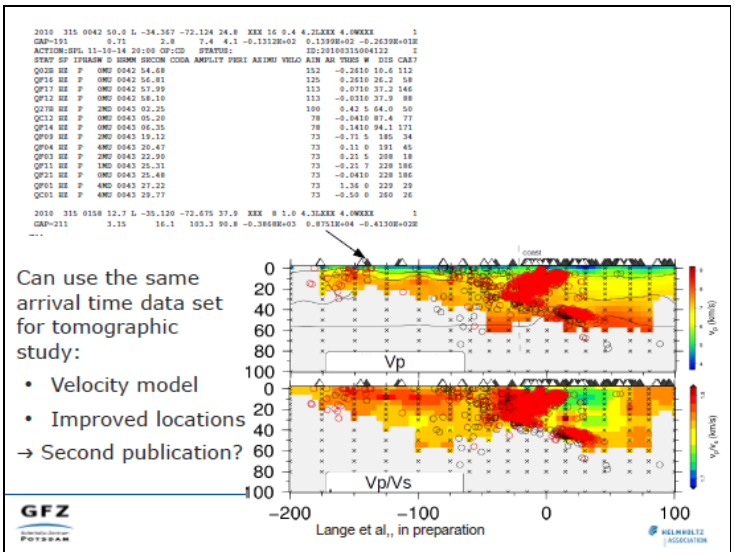
Folie 6



Folie 7



Folie 8



Folie 9

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures

Jon Claerbout
(Stanford geophysicst)

GFZ

Folie 10

An article about observational *science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures as well as reference to the underlying data, which must be open-access.*

paraphrasing Jon Claerbout
(Stanford geophysicst)

GFZ

Folie 11

An article about observational *science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures as well as reference to the underlying data, which must be open-access.*

paraphrasing Jon Claerbout
(Stanford geophysicst)

For any given observation-based publication, only time can tell whether the ideas and hypothesis inspired by the data or the observations themselves will have the more lasting impact and will be the reason for which the paper is cited

GFZ

Folie 12

Scientific value chain for Lange et al (2012)

Process	Product	Archival/distribution
Field work	Raw data, often proprietary format	USB drive
Conversion and metadata generation (geometry, instrument calibration)	Standardised waveform data	Seismological data archive (z.B. GEOFON) – possibly time-limited access restrictions
Data analysis 1: Picking of travel times, Earthquake location using 1D-model	Bulletin (Picks + Lokationen)	Some institutional repositories, e.g. ISC Temporary exp, project internal: local, email, dropbox
Data analysis 2: Tomographic inversion	3D-Geschwindigkeits-modell, verbesserte Lokationen	Local, email, dropbox
Seismotectonic interpretation; higher order relationships	New scientific knowledge (Cartoons, integrative plots)	ISI publication (Commercial publisher or scientific agency; (p)reprint servers)

GFZ

Folie 13

Private data organisation (computer-based analysis)

- Lab book: still widely used but varying in quality – more notebook than record of reference
- Unstructured file-based notes: `README` and `doit.csh` files
- In geoscience little danger of disputes of priority; proof of exact timing not usually required
- Challenges: finding old data; avoiding overwriting successful runs with new parameter sets; inability to reproduce prior results due to missing obsolete software

Personal 'low-tech' solutions, mostly based on UNIX tools:

- `unison/rsync/rsnapshot`
- `locate + grep`
- `subversion (svn)`

Also popular
- Dropbox

GFZ

Folie 14

Private data organisation (computer-based analysis)

Evaluation of existing practice

- Existing standard tools generally up to the job; challenge is in convincing people to use them (svn/Trac system in my section only used by 2-3 people)
- Necessity to move from `doit.csh` type note-taking to `make` based build system

GFZ

Folie 15

Informal data sharing

Why are email and similar informal approaches so popular?

- Immediate use; no need to register, no limitation on data type (except size → Dropbox)
- Full control over data release in each instance; can send 'latest version'
- Email allows informality in instructions and imposed restrictions.

Disadvantages:

- Maintenance does not scale well: each request generates work; hard to keep up-to-date with improvements to derived data
- Danger of data loss if data generator leaves science (common for PhD students, postdocs, but also senior faculty retire etc)

In spite of disadvantages and availability of superior tools, real challenge to get adoption in the wider community (`PanMetaDocs`)

GFZ

Folie 16

Vision I: static data and software post-publication

- Data available in form suitable for further processing at all stages of the value chain with suitable meta-information (geometry etc)
- Results reproducible using either standard open source software, or software supplied; parameter sets controlling software behaviour fully specified in digestible form
- Data discoverable by being fully tagged and geo-referenced

GFZ

Folie 17

Vision I: static data and software post-publication

- Data available in form suitable for further processing at all stages of the value chain with suitable meta-information (geometry etc)
- Results reproducible using either standard open source software, or software supplied; parameter sets controlling software behaviour fully specified in digestible form
- ~~Data discoverable by being fully tagged and geo-referenced~~

GFZ

Folie 18

Barriers to data sharing

Technological
- Lack of storage
- Proprietary data formats
- Interactive software

Personal
- Time and effort necessary for organising and documenting
- Reservation for future steps of scientific value chain
- Feeling of ownership (fear of data being stolen)
- Hiding irregularities / selective data presentation
- Software: retain edge / enforce collaboration by not releasing of software

- Social
  - Lack of attribution, impact hard to track
  - Data production given low value compared to publication record
  - Danger to reputation due to data misuse by others

GFZ

Folie 19

Overcoming barriers 'Carrots and sticks'

- Improving attribution and recognition
  - DOIs are the technical pre-requisite; great for consistent datasets early in the value chain, but they are not on the same footing as publications (e.g. don't count for *h* factor)
  - Formats for data publications (but in practice used rarely)
  - Journals editors and reviewers must pay more attention to proper data attribution (author guidelines; targeted questions to reviewers)

GFZ

Folie 20

Overcoming barriers 'Carrots and sticks'

- Improving attribution and recognition
  - DOIs are the technical pre-requisite; great for consistent datasets early in the value chain, but they are not on the same footing as publications (e.g. don't count for *h* factor)
  - Formats for data publications (but in practice used rarely)
  - Journals editors and reviewers must pay more attention to proper data attribution (author guidelines; targeted questions to reviewers)

Vision II: attribution
Use of data and data products fully referenced; data production rewarded on equal footing with paper production

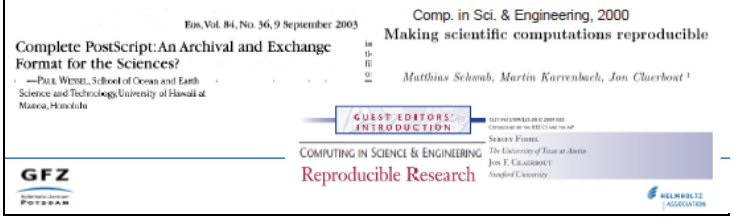
GFZ

Folie 21

Overcoming barriers
'Carrots and sticks'

- Supplementary data as pre-requisite of publication (journals/funding bodies)
  - Require authors to supply underlying data and software for all figures, moving toward *Reproducible Research*
  - Respect exclusive use for intermediate data: force archival but allow time-limited restrictions similar to primary data to replace distribution-on-demand

NB: Many journal guidelines are already requiring data sharing on demand, but this policy is hard to police, and often fails for reasons outlined earlier


Eos, Vol. 84, No. 36, 9 September 2003

Complete PostScript: An Archival and Exchange Format for the Sciences?

—Paul Wessel, School of Ocean and Earth Science and Technology, University of Hawaii at Manoa, Honolulu

Comp. in Sci. & Engineering, 2000

Making scientific computations reproducible

Matthias Schwab, Martin Karrenbach, Jon Claerbout

GUEST EDITORS' INTRODUCTION

COMPUTING IN SCIENCE & ENGINEERING

Reproducible Research

GFZ

Folie 22

Modest steps

- Aim for reproducible research in the private sphere
- Better enforcement of existing journal policies
- End practice of converting vector graphics to jpg's for publication (in post-production, but also by users in composing WORD documents)
- Encourage sharing and reproducible research by awarding openness grades

GFZ

Folie 23

Gold:

- Underlying primary data open and data usage described precisely (time frame, scope)
- All necessary software open and supplied or accessible from repository with full versioning history
- Secondary data products supplied
- Complete build sequence specified ('Reproducible research'); (by necessity) interactive steps clearly described, and results of all interactive steps supplied

Silver:

- As Gold, but also allow proprietary software (if versioned and usage documented) and reasonable time-restrictions on distribution of data to be in place

Bronze:

- Only require (1) and (3) from Gold standard, limited time-restrictions allowed

To encourage adoption, can make this voluntary, and at a later stage make „Bronze" or „Silver" requirement of publication.

Possible exceptions:

- Collaborations with industry: allow longer time restrictions, or selective obfuscation of data (e.g. geographic location)
- Can define standard proprietary software which is allowed even for Gold standard, e.g. Excel, matlab (BUT: versioning problem)

GFZ

Folie 24

What about dynamic data sets/products?

Time-dependency through

1) Growth: natural growth of most continuous datasets (e.g. continuous seismological data and earthquake catalogues)
2) Corrections: Fixing of obvious mistakes (e.g. polarity, timing for seismological data; discovery of bugs in algorithms)
3) Improvements: improved results due to: (data products)
   - Growth of and corrections to primary data
   - Improved methodology (e.g. location in 3D instead of 1D model)

GFZ

Folie 25

Vision III: time-dependency

- Provide access to the latest and best in primary data, data products and software but make available full time history of changes and easy reconstruction of any point in time
- Strive for automated analysis and provide self-updating data products which automatically assimilate additional data having become available since original paper
- For steps which cannot be automated: commit effort to keep data-product updated, or isolate manual steps and provide clear procedures for updating

GFZ

Folie 26

Summary

Data should be available and discoverable.

Data products should be re-usable, reproducible, and capable of assimilating new data.

The full time history must be available for both as well as for the software used.

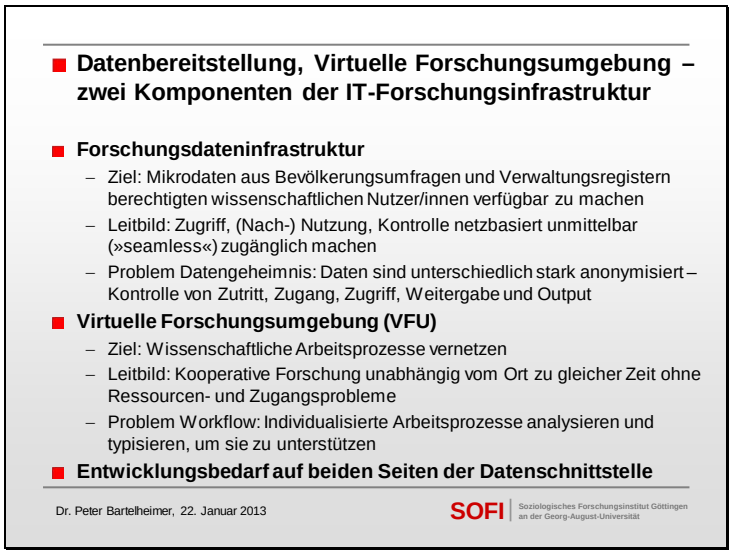
Sharing of data and software must be rewarded.

GFZ

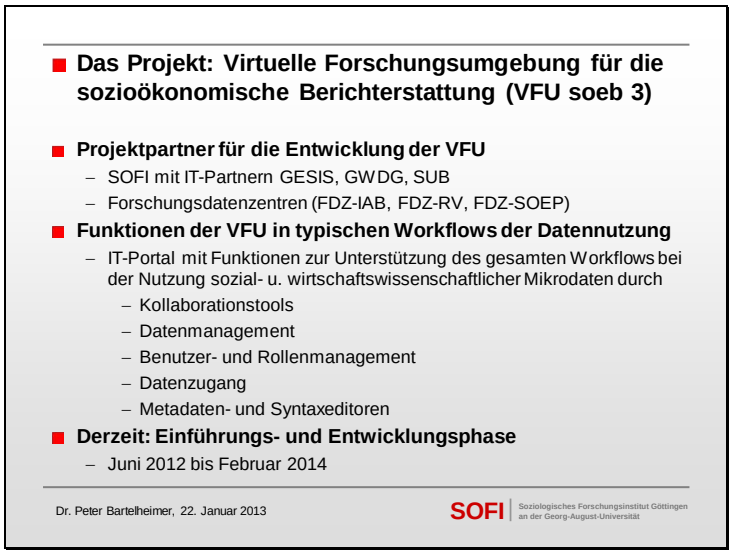
Folie 27

„Gruppendomäne ‚VFUs'" – Dr. Peter Bartelheimer (Uni Göttingen)

Folie 1



Folie 2



Folie 3

Anwendungsfall: Dritter Bericht zur sozio-ökonomischen Entwicklung in Deutschland (soeb 3)

- **Verbundvorhaben 2013 - 2015 in Vorbereitung (Antragslage)**
  - Bis zu 24 beteiligte wissenschaftliche Einrichtungen mit bis zu 29 Arbeitspaketen (AP)
  - Nutzung von 66 verschiedenen Forschungsdatensätzen, Datenbanken
  - Neun der beim RatSWD akkreditierten FDZ als Datengeber
  - An VFU-Entwicklung und Verbund beteiligt: FDZ-IAB, FDZ-RV, FDZ-SOEP
- **Schwerpunkte der Datennutzung**
  - Neue sozial- und wirtschaftswissenschaftliche Panel-Datensätze
  - 13 Datensätze kollaborativ genutzt, davon sieben intensiv (> zwei AP)
- **Projektbegleitende Weiterentwicklung der VFU**
  - Operative Nutzung im Verbund
  - Teilnahme von Verbundpartnern in Nutzungsstudien
- **Ziel: Bereitstellung einer »nachhaltigen« VFU-Plattform**

Folie 4

Welche Daten (»Objektklassen«) und Metadaten?

- **Unveränderliche Daten**
  - Forschungsdaten (Ausgangsdaten)
    - Datensätze und Dokumentation (»Studien«)
    - Metadaten auf Studien-, Datensatz und Variablenebene
  - Statistikprogramme (z.B. R), statistische Packages (z.B. ado-Files)
- **Im Workflow erzeugte Daten)**
  - Syntaxdateien, Syntax-Memos, Metadaten zu Syntax
  - Output-Dateien, Metadaten zu Outputs, z.B. Log-Dateien
    - Tabellen, Grafiken, Textdateien (diverse Datenformate)
- **Speicherorte, Orte der Nutzung außerhalb der VFU**
  - Forschungsdatenzentren, Daten(service)einrichtungen
  - Internet, Intranet
  - Lokale Workstation, Intranet

Folie 5

An der Datenschnittstelle: Wie kommen Forschungsdaten in die VFU?

- **Bereitstellung für lokale Nutzung**
  - Public / Scientific Use Files –
- **Bereitstellung für Rechnen in Dateneinrichtungen**
  - Gastarbeitsplätze: interaktiver lesender Zugriff auf Originaldaten, Outputkontrolle
  - Kontrolliertes Fernrechnen, zwei Varianten
    - Versenden von Syntax (»job submission«, »Remote Execution«), Übermittlung kontrollierter Outputs
    - »Remote Data Access«: interaktiver lesender Zugriff über gesicherte Verbindung (»safe centers«), Outputkontrolle
- **Jedes FDZ gestaltet Fernrechnen anders**
- **Anforderungen an Authentifizierung, Rechte- u. Rollenverwaltung**
  - Zugangskontrolle, Zugriffskontrolle, Weitergabekontrolle

Folie 6

VFU muss alle Datenzugangswege unterstützen

- **Rechnen in der VFU als Variante externer Bereitstellung**
  - Public Use Files (PUF) / Scientific Use Files – nach Vereinbarung
  - Nutzung der Metadaten und Syntax in der VFU beim Rechnen in der VFU – mit R oder lizenzierten Statistikprogrammen
- **»Job Submission«**
  - Einheitliche Nutzungserfahrung durch Integration der verschiedenen Schnittstellen in VFU-Portlet?
  - Nutzung der Metadaten und Syntax in der VFU für Job-Erstellung
  - Historisierung der Jobs anhand von Metadaten
- **»Remote Data Access«**
  - Anmeldung über Aufruf der VFU an gesicherten Standorten oder zertifizierten Forschungseinrichtungen
  - Nutzung der Metadaten und Syntax in der VFU beim interaktiven Fernrechnen

Folie 7

Hinter der Datenschnittstelle: Workflow und VFU

- **Aufarbeiten des Forschungsstands**
  - Arbeit mit Literatur: Recherche, Exzerpte, Bibliografie
- **Arbeit mit Forschungsdatensätzen**
  - Planung, Organisation, Dokumentation und Ausführung
    - Datensatzgenerierung
    - Recodierung und Analysen
    - Ergebnispräsentation
    - Datensicherung
- **Upload und Download**
  - von Dokumenten, Syntax, Outputs
- **Editieren**
  - von Syntax und Metadaten
- **Ergebnisdokumentation, Publikation (Open Access)**

Folie 8

Hinter der Datenschnittstelle: Wie kommen Syntaxdateien in die VFU?

- **Syntaxeditor als technische Entwicklungsaufgabe**
  - Funktionen des Erstellen und Nutzung von Syntax in der VFU
  - Upload / Download zwischen lokaler Workstation und VFU
  - Versionierung und Dateivergleich
  - Ggf. Volltextsuche
- **Soziale Aspekte – Arbeitsweise in der VFU**
  - Wer lädt welche Syntax hoch und wann?
    - Lauffähige (vollständige) Syntax oder Syntaxbausteine
  - Kontrolle über eigenes Arbeitsprodukt (geistiges Eigentum)
    - Rechte- und Rollenverwaltung: dateibezogene Rechte
  - Verhaltensanforderungen
    - z.B. Namenskonventionen, »kontrollierte Vokabulare«
  - Integrität zwischen lokaler Workstation und VFU?

Folie 9

Hinter der Datenschnittstelle: Wie kommen Metadaten in die VFU?

- **Metadatenschema als konzeptionelle Entwicklungsaufgabe**
  - Welcher Metadatenstandard (z.B. DDI 3.1, DDI 2.5)
  - Für Metadaten zu Syntax kein etablierter Standard
  - Dokumentationstiefe: Einschluss von Variablenebene?
- **Metadateneditor als technische Entwicklungsaufgabe**
  - Integration in VFU
    - Metadatenverwaltung in SQL-Datenbank
    - Verknüpfung mit »Objekten« in Dateiverwaltung
  - Suchfunktion: strukturierte Suche und Volltextsuche
  - Metadaten: Eingabe und Extraktion: z.B. aus Syntax, aus Datensätzen
- **Soziale Aspekte – Arbeitsweise in der VFU**
  - Verhaltensanforderungen und individuelle Arbeitsweise
    - Z.B. Erfassungsaufwand, Dokumentationstiefe, kontrollierte Vokabulare

Folie 10

Zum Schluss

- **Zwei Perspektiven auf VFU**
  - Kritisch für Akzeptanz datenhaltender Einrichtungen: Kontrollanforderungen
  - Kritisch für Akzeptanz durch Nutzer/innen: intuitive individuelle Anwendung
- **Nicht alle Daten stehen in der VFU**
  - Leistungsversprechen der VFU: einheitliche Arbeitsumgebung
  - Schrittweise mehr Daten integrieren
- **Entwicklungsbedarf auf beiden Seiten der Datenschnittstelle**
  - »Forschende FDZ« in VFU-Entwicklung und Nutzungsstudien integrieren
- **»Koevolution« technischer und sozialer Entwicklung**
  - Individualität der Forschungspraxis unterstützen
  - Entwicklung als transdisziplinäre Herausforderung
  - Selbstreflexion sozialwissenschaftlicher Datennutzung:
    - Analyse und Typisierung des Workflow ≠ Standardisierung

Folie 11

Mehr …

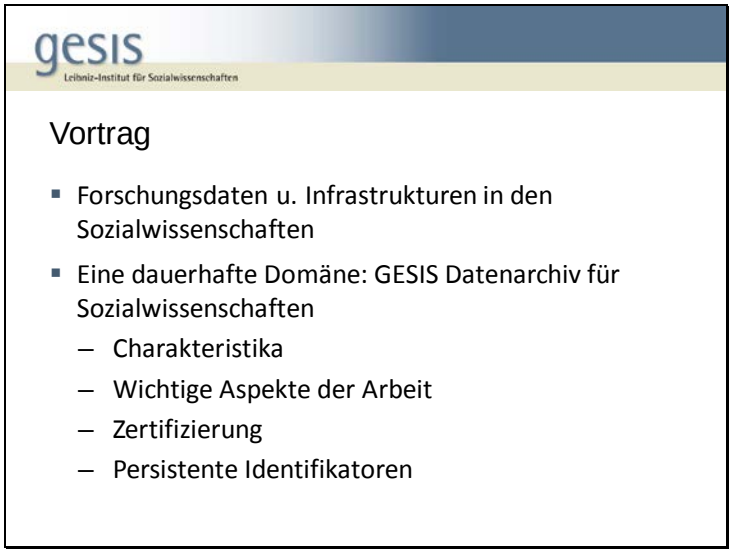
- **http://www.soeb.de**

Folie 12

Session 3 – Data Curation Continuum – Teil 2

„Persistente Domäne ‚PID, DLZA, Zertifikate'" – Reiner Mauer (GESIS)

Folie 1



Folie 2

gesis
Leibniz-Institut für Sozialwissenschaften

Forschungsdaten in den Sozialwissenschaften

- Sozialwissenschaften: Sammelbegriff für viele Fächer
  → vielfältige Datentypen (Statistik, Audio, Video, Text, …)
- Sekundär-/ Nachnutzung v. Forschungsdaten relativ verbreitet (Soziologie, Politologie)
- Lange Tradition von Datenarchiven
  erste Gründung 1940er-Jahre in USA; 1960 Zentralarchiv für Empirische Sozialforschung (jetzt GESIS Datenarchiv für Sozialwissenschaften)
- gegenwärtig allein in Europa 21 Datenarchive im CESSDA-Verbund organisiert (www.cessda.org)

Folie 3

gesis
Leibniz-Institut für Sozialwissenschaften

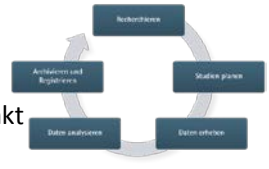
Forschungsdaten in den Sozialwissenschaften (2)

- Gründung des **Rats für Sozial- u. Wirtschaftsdaten (RatSWD)** in 2004 u. paralleler Aufbau von **Forschungsdatenzentren** verbreitern Datenzugänglichkeit deutlich
- SoWi-Dateninfrastrukturlandschaft entwickelt sich insgesamt sehr dynamisch, aber
  - → positive Entwicklung führt auch zu gewisser Unübersichtlichkeit im Datenzugang
  - → viele Daten existieren (?) weiterhin im ‚Verborgenen'
  - → Wachsende Anzahl an Akteuren in der ‚Access'-Domäne, aber persistente Domäne (noch) unterentwickelt bzgl. Breite u. Tiefe der Abdeckung

Folie 4

gesis
Leibniz-Institut für Sozialwissenschaften

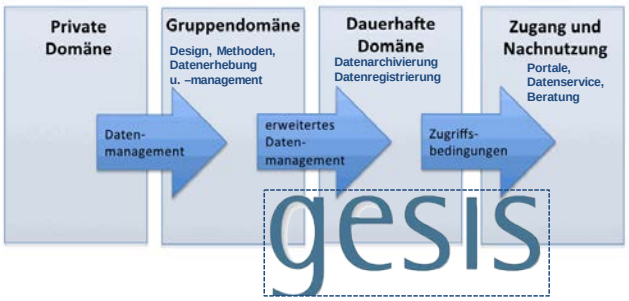
Was ist GESIS?



- Infrastruktureinrichtung für die Sozialwissenschaften, Schwerpunkt empirische Sozialforschung
  - **Forschungsdaten** (Erhebung, Archivierung, Registrierung, Analyse)
  - **Integrierte Fachinformationen** (Daten, Literatur, Projekte)
  - **Methodenberatung u. –entwicklung**
  - **Wissensvermittlung** (Summer School, Workshops, Seminare etc.)
- Basis der Serviceangebote ist eigene Forschung
- Langzeitarchivierung von Forschungsdaten explizit als Zielsetzung in der Satzung verankert (§ 2, 2c)

Folie 5

GESIS – Aktivitäten im Curation Continuum



[Grafik: WissGrid, Leitfaden zum Forschungsdaten-Management, eigene Textergänzungen (blau)]

Folie 6

GESIS Datenarchiv: persistente Domäne

- 50+ Jahre Archivierung von Forschungsdaten
  - Kontinuität trotz (oder wegen?) institutionellem Wandel gewahrt
  - technologischer Wandel extrem ; veränderte Anforderungen u. Nutzungsszenarien, organisatorischer u. rechtlicher Rahmen
- Fokus auf Kuratieren u. Bereitstellen für Nachnutzung
  - Kuratieren dient nicht nur dem Erhalt, auch d. Verbesserung der Nachnutzbarkeit u. dem Schaffen von Mehrwert
  - Langzeitarchivierung lange Zeit eher implizite Aufgabenstellung
- Fachwissenschaftlicher/ disziplinärer Ansatz
  - eigene Forschung u. Einbettung in Forschungszusammenhänge sind Teil des Selbstverständnisses (Ausdifferenzierung der Funktionen hat allerdings stark zugenommen)

Folie 7

Datenarchiv: Persistenz, LZA (2)

- Archivierungseinheiten (Studien) bzw. AIPs sind ….
  - … **heterogen:** bestehen aus mehreren bis vielen (im Extremfall mehreren hundert) Objekten , die in ihrer Zusammensetzung variieren (ein od. mehrere Datensätze, Messinstrumente, Metadaten, begleitende Materialien …)
  - … **nicht statisch:** Daten u. Metadaten werden im Archiv (kontinuierlich) verändert bzw. erzeugt (korrigiert, erweitert um neue Datenpunkte, integriert, aufgewertet)

→ **LZA an wichtigen Stellen von manuellen/ intellektuellen Prozessen bestimmt** (bspw. Ingest inkl. Eingangskontrolle und Datenbereinigung, Datenaufbereitung u. –dokumentation)

Folie 8

gesis
Leibniz-Institut für Sozialwissenschaften

Datenarchiv: Persistenz, LZA (3)

- eher **kleine Daten**
  → vergleichsweise geringe Speicheranforderungen
- kleine Daten**, aber komplexe Inhalte**
  → Aufbereitung, Dokumentation u. Metadaten spielen wichtige Rolle beim Kuratieren; sonst keine (sachgerechte) Nachnutzung möglich
  → SoWi verfügt mit **DDI** über einen komplexen, internationalen Metadatenstandard <ddi>
  (www.ddialliance.org/)

Folie 9

gesis
Leibniz-Institut für Sozialwissenschaften

Datenarchivierung@GESIS: Funktional

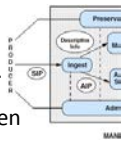
- **Akquisition** (aktiv, passiv)
- **Aufnahme ins Archiv** (Ingest)
- **Datenaufbereitung u. –dokumentation**
  (Standard für alle, Added-value für ausgewählte Studien)
- **Langzeitarchivierung**
- **Datenservice (Access)**: Beratung, Datenzugang
  (Download, Online-Analyse, manuelle Bereitstellung)
- **Datenregistrierung (da|ra)**: Vergabe von DOIs im DataCite Verbund

10

Folie 10

gesis
Leibniz-Institut für Sozialwissenschaften

Ingest / Aufnahme ins Archiv (1)



- Vorbereitung d. Archivierung mit Datengeber
  - Ziel ist es möglichst früh im Lebenszyklus anzusetzen
  - Klärung bzgl. Umfang, Formaten, Aufbereitungs- u. Dokumentationszielen, Zugangsklasse
  - Abschluss Archivierungsvertrag
- Eingangskontrolle SIP und Basis-Aufbereitung
  - Technische Kontrollen (Formate, Lesbarkeit, Virenfreiheit …)
  - Vollständigkeit u. Nutzbarkeit
  - Konsistenz der Daten
  - **ggf. korrigierende Aufbereitungen**

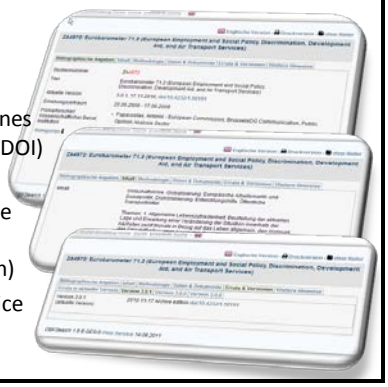
11

Folie 11

gesis
Leibniz-Institut für Sozialwissenschaften

Ingest / Aufnahme ins Archiv (2)

- Studienbeschreibung (inhaltliche, methodische, technische Charakteristika)
- Versionierung u. Vergabe eines persistenten Identifikators (DOI)
- Erzeugung AIP (Originale, aufbereite und normalisierte Versionen, technische u. administrative Metadaten)
- Erzeugung der für den Service bestimmten Objekte (DIP)



Folie 12

gesis
Leibniz-Institut für Sozialwissenschaften

Datenaufbereitung / -dokumentation

FDZ ALLBUS
FDZ Amtliche Mikrodaten
FDZ Internationale Umfrageprogramme
FDZ Wahlen

Adding value

- hauptsächlich für *ausgewählte* komparative Studien, die kontinuierlich durchgeführt werden bzw. mehrere Länder umfassen
- maßgeschneiderte Unterstützung für große Umfrageprogramme
- häufig ist GESIS bereits an der Datenerhebung beteiligt
- Service u. Forschung dazu i.d.R. durch GESIS FDZs

ALLBUS · GLES · CSES · European Values Study · ISSP

13

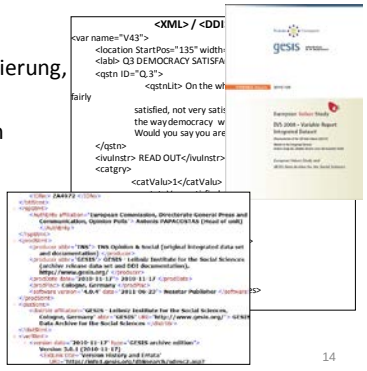
Folie 13

gesis
Leibniz-Institut für Sozialwissenschaften

Datenaufbereitung / -dokumentation (2)

Adding value

- Standardisierung, Harmonisierung, Integration
- Umfassende Produktion von strukturierten Metadaten
- Kontextinformationen
- Variablenreports, Datenhandbücher, Methodenberichte, ….



14

Folie 14

gesis
Leibniz-Institut für Sozialwissenschaften

Access / Datenbereitstellung

- 40.000+ Datenweitergaben in 2012
- Weit überwiegend akademische Nutzung: Lehre und Forschung
- Je nach Angebot 30%-70% internationale Nutzer
- Datenzugang über Online-Portale und (individuelle) Bereitstellung
- (Meta)Datenportale: Retrieval, Download, Online-Analyse
- Zugang zu Datenbeständen internationaler Partnerarchive (ICPSR, CESSDA, IFDO)

15

Folie 15

gesis
Leibniz-Institut für Sozialwissenschaften

Langzeitarchivierung: Substanzerhaltung

- physischer Schutz vor Ort
- räumlich getrennte und redundante Datenhaltung
- Diversivität eingesetzter Speichertechnik
- regelmäßige Medienmigration bzw. Refreshment

16

Folie 16

gesis
Leibniz-Institut für Sozialwissenschaften

Langzeitarchivierung: Erhalt von Nutzbarkeit und Interpretierbarkeit

- Bedrohung der digitalen Bestände durch technischen Fortschritt ist eine Konstante in der 50jährigen Archivarbeit
- Maßnahmen, die auf Erhalt des bitstreams abzielen nicht ausreichend

→ Erhalt wird hauptsächlich durch Migrationsstrategien erreicht (zur Überbrückung auch Emulation bzw. Virtualisierung)

17

Folie 17

gesis
Leibniz-Institut für Sozialwissenschaften

Langzeitarchivierung: Erhalt von Nutzbarkeit und Interpretierbarkeit (2)

Maßnahmen:

- Dokumentation und Metadaten, Metadaten, Metadaten, ….
- Definierte und standardisierte Formate (Ingest):
  erleichtert Monitoring, senkt Migrationsbedarf, verringert Ressourcenbedarf bei Migration
- Verfolgen d. technischen Entwicklung (insbesondere Software und damit verbundene Dateiformate)
- Formatmigrationen
  nur wenn Gefahr der Beeinträchtigung der Nutzbarkeit oder wenn mit der Migration so große Vorteile für die Nutzung oder die Archivarbeit einhergehen, dass der Aufwand zu rechtfertigen ist

18

Folie 18

gesis
Leibniz-Institut für Sozialwissenschaften

Zertifizierung: „Vertrauenswürdige Archive“

Vertrauen in der dauerhaften Domäne von großer Bedeutung:

- Datenproduzenten u. -nutzer, Kooperationspartner, Forschungsförderer müssen darauf vertrauen und sich davon überzeugen können, dass ein Langzeitarchiv in der Lage ist, dauerhaft
  – Authentizität,
  – Integrität,
  – Verfügbarkeit und Zugänglichkeit,
  – sowie Interpretierbarkeit der Archivobjekte zu erhalten

→ **Kurz u. mittelfristig evtl. kein Problem, aber dauerhaft …?**

Folie 19

gesis
Leibniz-Institut für Sozialwissenschaften

Zertifizierung (2)

- Vertrauen muss erworben werden, aber **Vertrauenswürdigkeit** kann durch eine Zertifizierung nachgewiesen werden (wird mittelfristig ein ‚Muss‘ in der persistenten Domäne)
- Zertifizierung **auch und gerade für etablierte Institutionen relevant,** wie bspw. das GESIS Datenarchiv
  → Grundlegende Abläufe und Verfahren wurden lange vor OAIS, DIN 31644 od. ISO 16363 etabliert
  → Zertifizierung hilft bei der Identifikation v. Schwachstellen u. kann zur Verbesserung der Abläufe u. Services beitragen

Folie 20

Zertifizierung (3)

- **Optionen**:
  - Data Seal of Approval (DSA): Selbst-Audit mit Review
  - DIN 31644: Kriterien für vertrauenswürdige digitale Langzeitarchive
  - ISO 16363: Audit and certification of trustworthy digital repositories
- 3-stufiges Verfahren "**European framework for audit and certification of digital repositories**":
  - **Basis**: DSA
  - **Erweitert**: DSA plus extern begutachtetes self-audit (ISO od. DIN)
  - **Formal**: DSA plus externes Audit und Zertifizierung (ISO od. DIN)

Folie 21

Zertifizierung (4): GESIS Datenarchiv

- Zunächst Basiszertifizierung (gerade angelaufen); anschließend ‚internes' Selbst-Audit DIN
- Weitere Schritte derzeit noch nicht entschieden, sehr wahrscheinlich ‚erweiterte' Zertifizierung:
  → entsprechende Angebote für die ‚erweiterte' u. ‚formale' Zertifizierung sind in Entwicklung (z.B. **nestor** für DIN)
  → kritische Entscheidungsgröße ist der dafür notwendige Ressourceneinsatz

Folie 22

Persistente Identifikatoren (PID): Was ist das?

- Ein Identifier (Name), der **eindeutig** auf **Dauer** mit einem **digitalen Objekt** verknüpft ist
- Zielsetzung: digitale Objekte langfristig identifizierbar, referenzierbar und verfügbar zu halten
- Verschiedene technische Systeme, verschiedene Policies, verschiedene Vor- u. Nachteile

→ Zu möglichen Risiken u. Nebenwirkungen fragen Sie Ihren Arzt oder Apotheker

Folie 23

Persistente Identifikatoren (PID): Wozu?

(1) Zitation: Daten werden selten od. ungenügend zitiert

- Probleme:
  - → Forschungsergebnisse sind nicht replizierbar (weil Daten nicht auffindbar; Unsicherheit bzgl. Versionen u. Quellen)
  - → Zitationen weisen **Reputation** zu; es **fehlen Anreize, Daten zu publizieren und damit zu teilen**
- Zitation mit PIDs eindeutig; erlauben schnellen Zugriff; können für bibliometrische/ szientometrische Analysen sowie Impact-Messung verwendet werden

(2) Verlinkung (Daten mit Daten, Daten mit Publikationen, Daten mit …)

Folie 24

PID: Fragen zur Implementierung

- Welches System?
  Technologie, Stabilität, Kosten, Zusatzdienste, Metadaten, Registrierungsprozess
- Was soll mit welcher Granularität registriert werden?
- Was geschieht, wenn das Objekt verändert wird? Wann wird ein neuer PID vergeben?
- Wie soll mit Objekten umgegangen werden, die auch an anderen Orten/ bei anderen Institutionen vorgehalten werden?
- Wohin löst der PID auf? Direkt auf das Objekt, landing page …?

Folie 25

Herausforderungen

- Dateninfrastruktur zunehmend verteilter
  → Stärkere Bedeutung strukturierende u. kooperative Dienste
- Bedeutung komplexer Forschungsdesigns und neuer Datenformen wächst
  → LZA muss mit immer komplexeren Forschungsdaten umgehen;
- Verlinkung/Integration von Daten aus unterschiedlichen Datenquellen u. Disziplinen nimmt zu
  → höhere Anforderungen an Daten u. Metadaten; rechtliche Fragen, Lizenzfragen; Qualität: Aktualisierung, Versionierung, autoritative Versionen

Folie 26

gesis
Leibniz-Institut für Sozialwissenschaften

„Eternity is really long, especially near the end“
[Woody Allen]

Vielen Dank!

Kontakt: Reiner Mauer, reiner.mauer@gesis.org

Folie 27

gesis
Leibniz-Institut für Sozialwissenschaften

GESIS Datenarchiv: Datenquellen

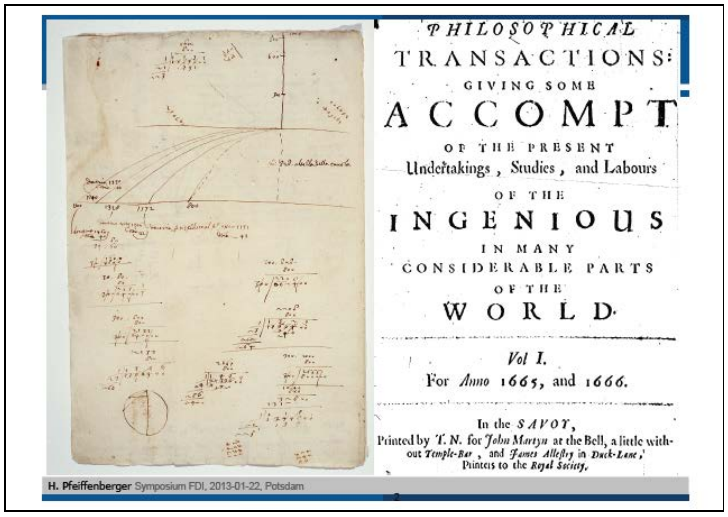
Daten wurden / werden

- … **selbst bzw. unter Beteiligung von GESIS erhoben**:
  z. B. ALLBUS, EVS, ISSP, GLES
- … **akquiriert** (externe Daten):
  Großteil des Archiv-Bestands, wie z.B. Politbarometer, Eurobarometer, Mediaanalyse, Reiseanalyse, viele Einzelstudien)
- … **entwickelt / produziert / transformiert**
  Produktion zeit- u./od. ländervergleichender Datensätze: Eurobarometer, EVS, ISSP, Politbarometer, Daten der historischen Statistik (HISTAT)

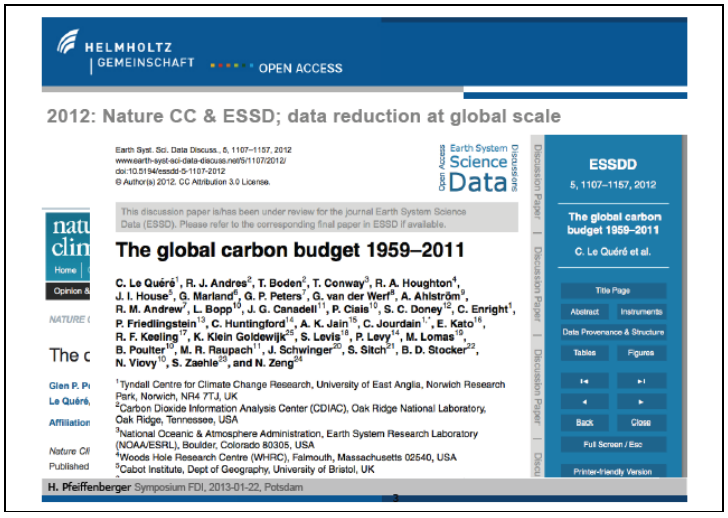
Folie 28

„Zugangsdomäne ‚Portale, Best Practices'" - Dr. Hans Pfeiffenberger (AWI)

Folie 1

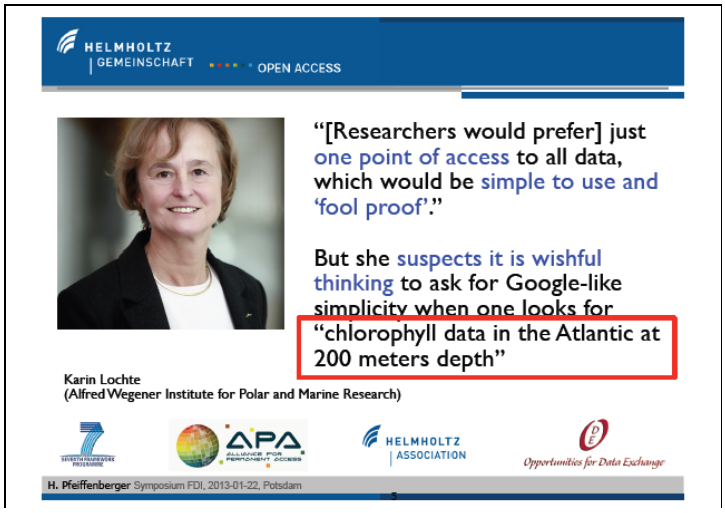


Folie 2

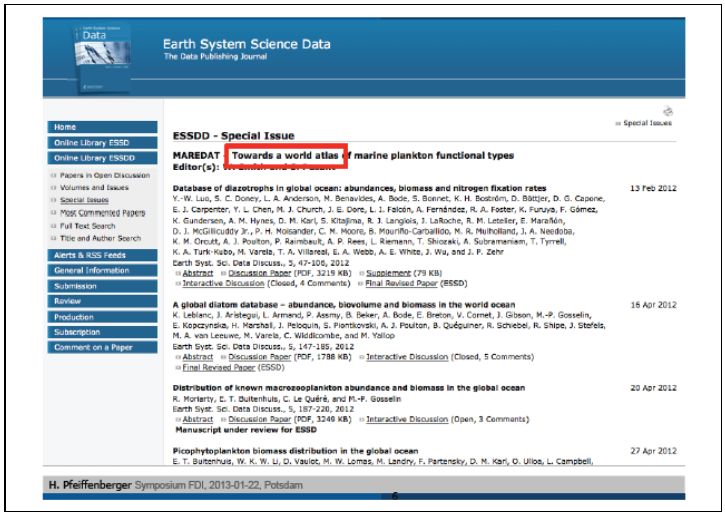


Folie 3

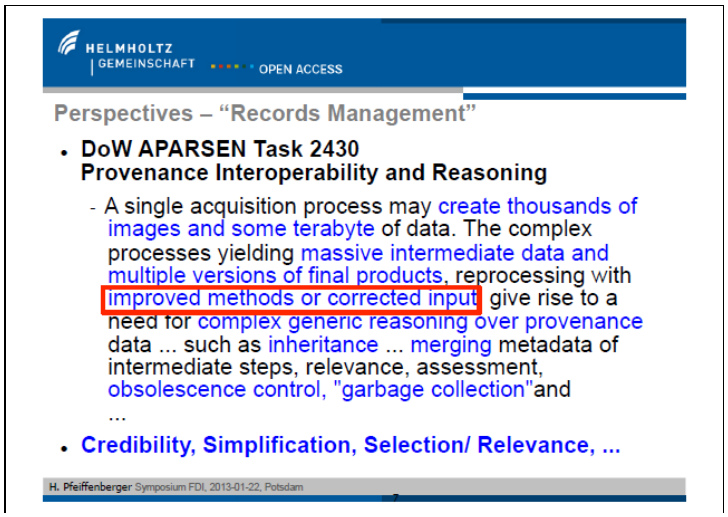
Folie 4



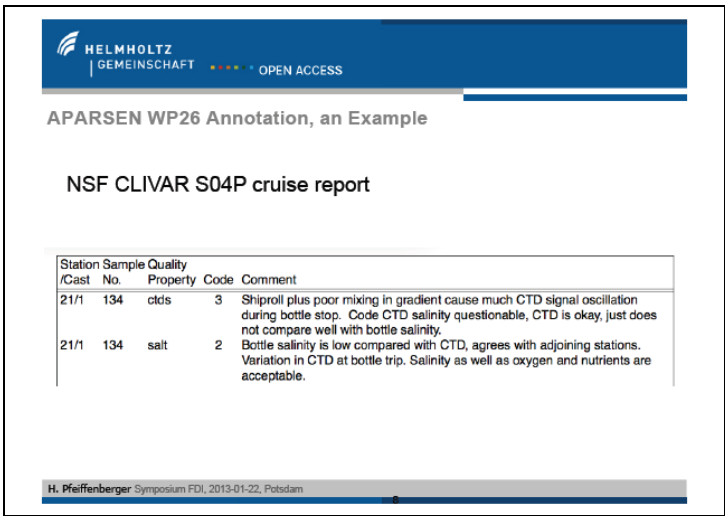
Folie 5



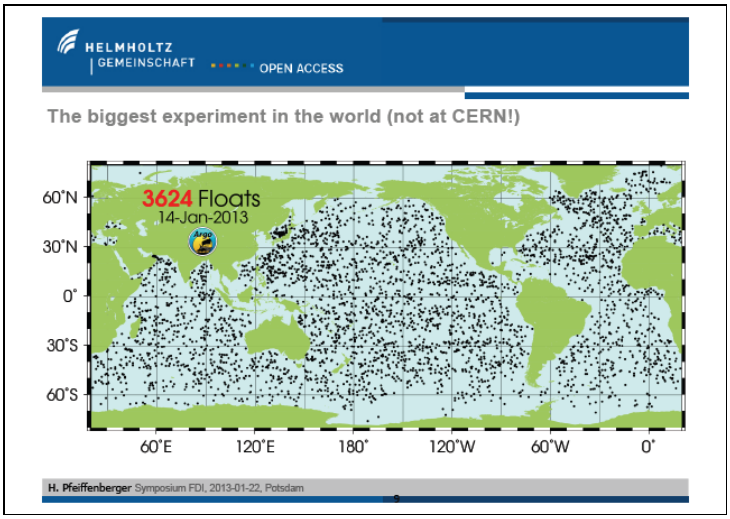
Folie 6



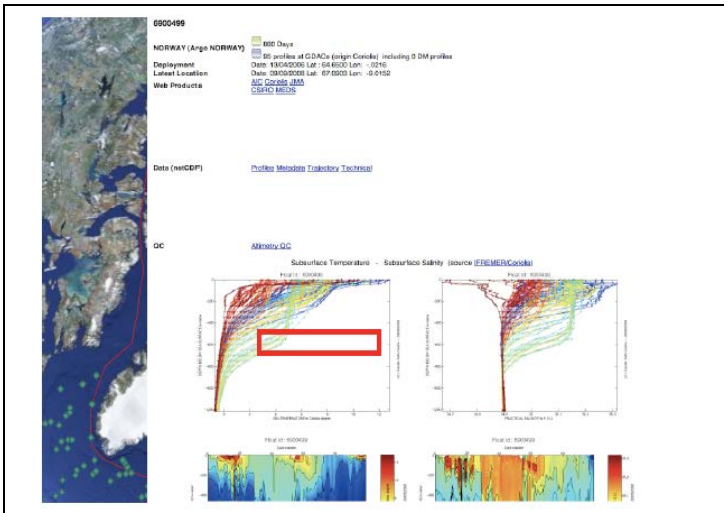
Folie 7



Folie 8



Folie 9

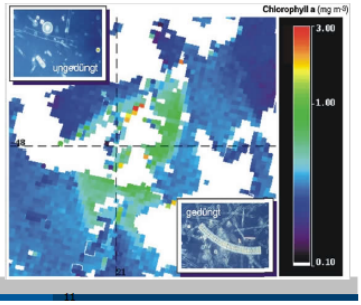


Folie 10

HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

An important, typical Experiment

- EISENEX / EIFEX : Two expeditions of "**Polarstern**" :
  With a few tons of iron fertilizer, south of Capetown ….
- EIFEX (2004):
  - 54 scientists and students from
  - 14 institutes and 3 companies from
  - 7 EU countries and South Africa
  - **Oceanographers**
  - **Biologists**
  - **Chemists…..**
- **"Biogeochemistry"**
- **+ Satellite observations !**

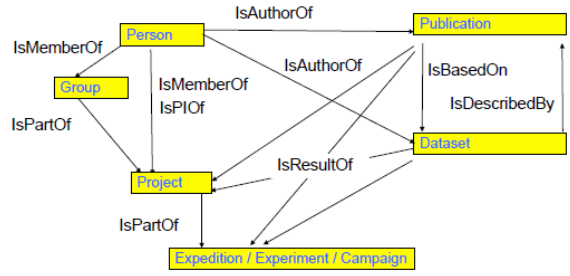


H. Pfeiffenberger Symposium FDI, 2013-01-22, Potsdam

Folie 11

HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

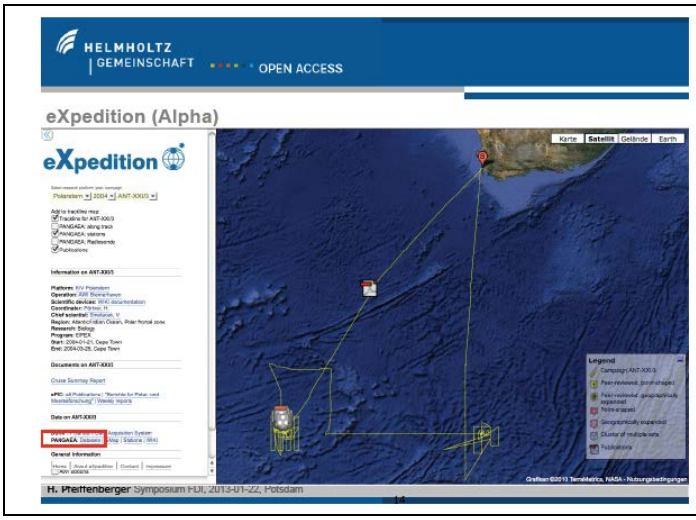
Pfeiffenberger, Macario, Text, Data and People, OAI4, CERN 2005



H. Pfeiffenberger Symposium FDI, 2013-01-22, Potsdam

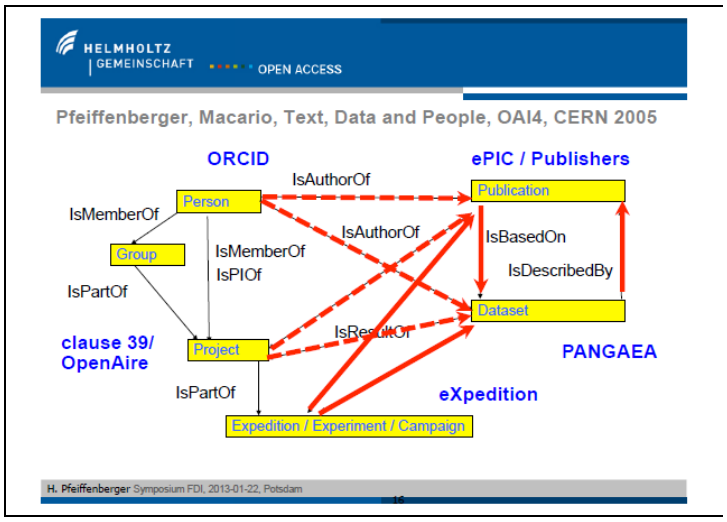
Folie 12

Folie 13



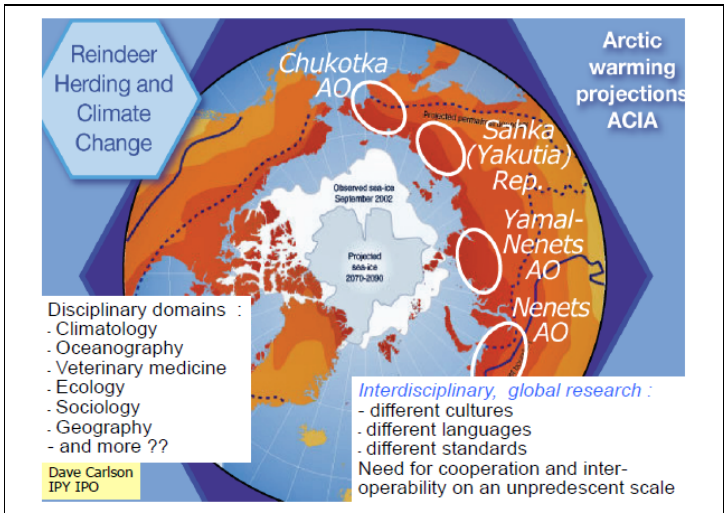
Folie 14

Folie 15



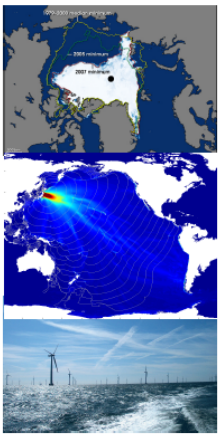
Folie 16

Folie 17



Folie 18

MaNIDA – Enabling Data-Intensive Marine Science



Global Change
- Assessing, understanding, and predicting environmental changes
- Human environmental impact

Hazards
- Risk analysis and support for disaster management
- Understanding environmental factors affecting human health

Resources
- Sustainable ecosystem management
- Energy from the ocean

Folie 19

Status of Marine Science Data in Germany

Extremely wide range of data sources
- Ship-based instruments
- Instruments in water column and at seafloor
- Air- and space-borne instruments
- Sensor networks (increasingly in the deep ocean)

- **"snapshots"** (individual projects) and **long-term monitoring**
- stored in **different data centers** (e.g. AWI/PANGAEA, BSH/DOD, HZG/COSYNA)

→ Requirement : Data-intensive research through **coherent data portal with common access strategy**

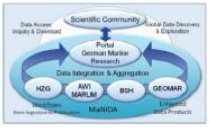
Folie 20

HELMHOLTZ GEMEINSCHAFT · OPEN ACCESS

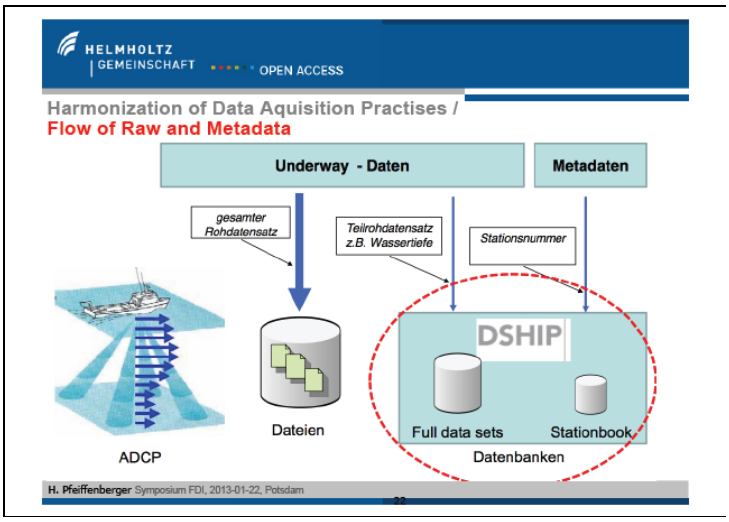
Data Flow from Research Vessels

- FS Polarstern
- FS Meteor
- FS Maria S. Merian
- FS Sonne
- FS Heincke
- FS Alkor
- FS Poseidon
- FS Elisabeth Mann-Borgese

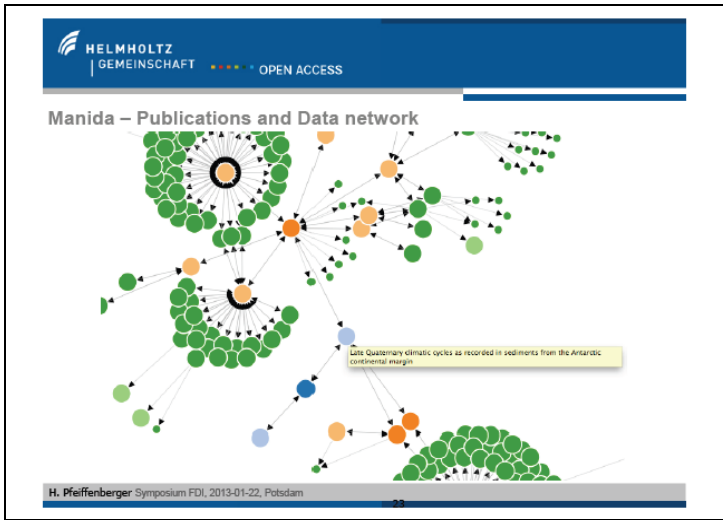
→ • BSH • GEOMAR • AWI →



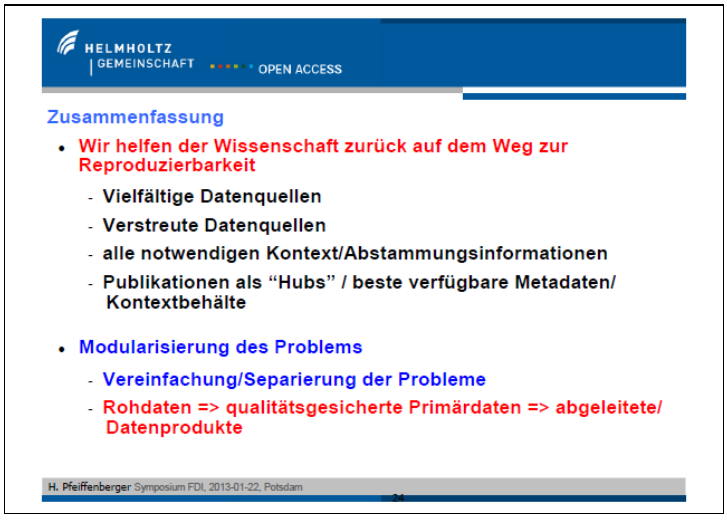
Folie 21



Folie 22



Folie 23



Folie 24


HELMHOLTZ
GEMEINSCHAFT
OPEN ACCESS
Thank you!
www.awi.de
www.manida.org
oa.helmholtz.de
H. Pfeiffenberger Symposium FDI, 2013-01-22, Potsdam

Folie 25