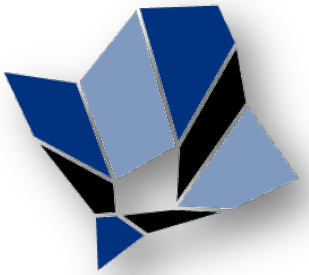


# DFG-Projekt RADIESCHEN

**Ra**hmenbedingungen einer **di**sziplinübergr**e**ifenden
For**sch**ungsdat**en**infrastruktur

## Materialien des
## Experten-Workshops

vom

17.April 2012

GFZ, Potsdam

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

# Einleitung

Der vorliegende Band enthält Materialien zum Experten-Workshop „Elemente einer übergreifenden Forschungsdaten-Infrastruktur: Eine für Alle?“. Der Workshop wurde veranstaltet durch das DFG-Projekt „Rahmenbedingungen einer disziplin-übergreifenden Forschungsdateninfrastruktur (Radieschen)“ und fand am 17.April 2012 am GeoForschungsZentrum (GFZ) in Potsdam statt.

Die Materialien beinhalten den Workshop-Report mit einer Zusammenfassung der Ergebnisse, sowie die Präsentationen der Veranstaltung. Das Workshop-Programm (S. 12) gibt einen Überblick über die Themen der einzelnen Sessions und deren Redner.
Der Workshop wurde eröffnet mit einer Begrüßung durch Dr. Jörn Lauterjung (GFZ), gefolgt durch die Keynotes von Peter Wittenburg (MPI für Psycholinguistik) zu „Forschungsdaten-Infrastrukturen: Bauen wir einen neuen Turm von Babylon“ (S. 13) und Dr. Stefan Winkler-Nees (DFG) zum Thema „Forschungsdaten-Infrastrukturen in Deutschland“ (S. 28). Dr. Jens Klump präsentierte das DFG-Projekt „Radieschen“ (S.42) und leitete über in vier parallele Break-Out-Sessions.
Session 1 „Policies und Anreize“ stellte die Frage nach sinnvollen und notwendigen Richtlinien im Umgang mit Forschungsdaten. Heinz Pampel (Helmholtz-Gemeinschaft – Open Access) und Dr. Peter Lampen (ISAS e.V.) gaben Impulsvorträge (S. 52).
Session 2 „Einbindung in den Forschungsprozeß“ befasste sich mit der Frage, ob die Daten zur Infrastruktur oder die Infrastruktur zu den Daten kommt. Matthias Razum und Anton Güntsch trugen hierzu vor (S. 82).
Session 3 „Generische vs. Disziplinspezifische Dienste“ stellte die Erfolgskriterien disziplin-übergreifender Dienste zur Diskussion. Brigitte Hausstein (GESIS) gab zu diesem Thema einen Impulsvortrag (S. 110).
Die Möglichkeiten und Grenzen der Auslagerung und Zentralisierung von Diensten bildete das Thema von Session 4 (S.145). Hier steht der Vortrag von Prof. Dr. Nagel, TU Dresden – ZIH, zur Verfügung.
Die Teilnehmerliste der Veranstaltung bildet den Abschluss der Materialien-Sammlung (S. 164).

Wir danken allen Experten, die mit ihren Vorträgen und Diskussionsbeiträgen zum Gelingen dieses Workshops beigetragen haben.

Weitere Informationen zum DFG-Projekt „Radieschen“ und dessen Aktivitäten finden Sie unter:
www.forschungsdaten.org

Potsdam, April 2012

*Das „Radieschen“-Konsortium*

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

# DFG-Projekt Radieschen – Experten-Workshop
# Elemente einer übergreifenden Forschungsdaten-Infrastruktur: Eine für Alle?

17. April 2012, GeoForschungsZentrum Potsdam (GFZ)

## Workshop-Bericht

Der Aufwand für die Gewinnung von Daten in so unterschiedlichen Disziplinen wie Geowissenschaften, Soziologie, Medizin oder Astrophysik liegt allein in Deutschland in der Größenordnung von mehreren Milliarden Euro. Viele dieser Datensätze sind einmalig und in der Erfassung nicht wiederholbar. Dabei handelt es sich sowohl um große Datenmengen, wie sie zum Beispiel in der Fernerkundung durch Satellitenaufnahmen entstehen, als auch um eher kleine Datensätze, die beispielsweise in Laborexperimenten erzeugt werden. Wissenschaftler und Institutionen stehen vor der großen Herausforderung, die Vielfalt dieser Daten systematisch zu erfassen, zu speichern und für eine weitere Nutzung vorzubereiten. Für diese Aufgaben müssen neue Infrastrukturen aufgebaut sowie organisatorische Lösungen für den Umgang mit Daten in der Forschung gefunden werden. Eine noch offene Frage ist, welche dieser Dienste sinnvoll zentral von disziplinübergreifenden Infrastrukturen angeboten werden können, und welche Dienste besser lokal aufgebaut werden.

**Abb. 1: Begrüßung durch Dr. Jörn Lauterjung, GFZ,**

Am 17. April diskutierten 80 Wissenschaftler verschiedener Disziplinen und Forschungseinrichtungen im Rahmen eines Workshops am GFZ, wie dieser Herausforderung begegnet werden kann. Der Workshop "Elemente einer übergreifenden Forschungsdaten-Infrastruktur: Eine für Alle?" behandelte Policies und Anreize, die Einbindung der Datenzyklen in den Forschungsprozess, generische vs. disziplin-spezifische Dienste, sowie die Möglichkeiten und Grenzen der Auslagerung und Zentralisierung von Diensten.

## Keynotes

Nach Begrüßungsworten durch **Dr. Jörn Lauterjung**, GeoForschungszentrum Potsdam (GFZ), eröffnete **Peter Wittenburg** (MPI für Psycholinguistik) mit seiner Keynote "Forschungsdaten-Infrastrukturen: Bauen wir einen neuen Turm von Babylon?" das Plenum. Der Vortrag begann mit einem Blick auf unterschiedliche Typen von Datensätzen. Er unterschied in 'Big Data' und 'Smart Data'. Big Data definiert sich vor allem durch Menge und stellt Herausforderungen an Transport, Rechenleistung und Speicherplatz. Smart Data dagegen definiert sich vor allem durch Heterogenität und stellt Herausforderungen an Datenmanagement, Datenqualität und Kuration sowie Interoperabilität. Integration, Kuration und Interoperabilität beschrieb Peter Wittenburg als die Kernpunkte im Umgang mit wissenschaftlichen Daten und Datensätzen. Jedoch sind bereits die Definitionen von Integration und Interoperabilität vielfältig.

**Abb. 2: Keynote-Sprecher Peter Wittenburg, MPI für Psycholinguistik**

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

Daten-Infrastrukturen sind komplex und in der Theorie, nicht aber in der Praxis einfach umzusetzen. Daher wird die Ausbildung von sog. 'Data Scientists' gefordert, welche sich im Detail und mit vertieften Kenntnissen der Problematik widmen. Im europäischen Kontext beschäftigen sich eine Reihe von Projekten mit Daten-Infrastrukturen, darunter Clarin, EPOS, EUDAT und auf nationaler Ebene das DFG-Projekt Radieschen. Die wichtigsten Kernaussagen sind, dass

- der aktuelle Status sehr unterschiedlich ist,
- sich alles in Bewegung befindet und noch vieles beeinflussbar ist,
- Heterogenität besteht, aber Harmonisierung möglich scheint,
- ein Trend zu kompatiblen Lösungen besteht, aber auch ein Wissensmangel in der Umsetzung.

Abschließend behandelte Peter Wittenburg die Frage, ob es gelingen kann die Komplexität zu beherrschen, diskutierte mögliche Lösungsansätze und stellte die Initiative ‚Data Access and Interoperability Task Force' (DAITF) vor. Die EU, USA, Australien und Kanada sind bereit die Task Force zu unterstützen, weitere Länder sind ebenfalls an einer Teilnahme interessiert.

**Dr. Stefan Winkler-Nees** (DFG) betrachtete in seiner Keynote „Forschungsdaten-Infrastrukturen in Deutschland - Sind wir auf dem Weg in Richtung ‚Open Science'?" den Umgang mit Forschungsdaten in Deutschland sowie die Rolle der AG Forschungsdaten der Allianz der deutschen Wissenschaftsorganisationen in diesem Kontext. Er beschrieb Forschungsdaten als Grundlage wissenschaftlicher Erkenntnis, die bislang nur unzureichend genutzt wird, nur eingeschränkt zugänglich ist und nicht längerfristig zur Verfügung steht. Die Vision der Wissenschaftsorganisationen und potentiell auch der Wissenschaftler ist, dass "Forschungsdaten möglichst frei und überregional zugänglich und langfristig verfügbar sein" sollten. Stefan Winkler-Nees bemerkte, dass "Open Science" kein neues Thema sei, sondern bereits zu Goethes Zeiten Diskussionsgegenstand war. Er beschrieb die Vielfalt an Akteuren, welche in Deutschland mit Forschungsdaten interagieren: Wissenschaftler, Institutionen, Informationsinfrastrukturen, Wissenschafts-organisationen und Fördereinrichtungen. In diesem Kontext wies er auf das neu erschienene Handbuch "Langzeitarchivierung von Forschungsdaten: Eine Bestandsaufnahme" hin.

**Abb. 3: Keynote-Sprecher Dr. Stefan Winkler-Nees, DFG**

Dr. Winkler-Nees gab einen Überblick über die Struktur der deutschen Forschungsförderung und die Allianz der Deutschen Wissenschaftsorganisationen. Die DFG arbeitet gemeinsam mit den Allianzpartnern zusammen in der Schwerpunktinitiative "Digitale Information". Schwerpunkte der gemeinsamen Arbeit sind: Nationale Lizenzierung, Open Access, Nationale Hosting-Strategien, Forschungsdaten, Virtuelle Forschungsumgebungen und Rechtliche Rahmenbedingungen. Herr Winkler-Nees stellte die Grundsätze der Allianz zum Umgang mit Forschungsdaten vor und beschrieb die Handlungsfelder der Allianz, die sich in 2012 auf Forschungsdatenmanagement, die Nachnutzung und Verfügbarkeit von Forschungsdaten, Kostenstrukturen und juristische Rahmenbedingungen konzentrieren. In diesem Kontext empfahl er das Information Literacy Program der Australian National University als Beispiel für einen institutionellen Leitfaden. Abschließend beschrieb Herr Winkler-Nees kurz die Position der EU Kommission zu 'Open Science for the 21th century'.

## Projekt-Präsentation Radieschen

**Jens Klump** 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 http://www.forschungsdaten.org/

**Abb. 4: Präsentation ‚Radieschen‘ , Dr. Jens Klump, GFZ**

Ziel ist es, diese Website auch über die Lebenszeit des Projekts hinaus zu einer Informationsplattform für Forschungsdaten-Infrastrukturen auszubauen.

## Breakout Sessions

National und international werden derzeit in einer Vielzahl von Aktivitäten Infrastrukturen für Forschungsdaten aufgebaut. Gemeinsam ist diesen Infrastrukturen, dass sie sowohl bessere als auch effizientere Forschung ermöglichen sollen, indem sie den Umgang mit Forschungsdaten unterstützen. Einer von vielen Aspekten, in dem sich die Vorhaben unterscheiden, ist, wie generisch oder disziplinspezifisch sie sind. Die wichtigsten Fragestellungen in diesem Kontext sind:

- Welches sind die organisatorischen, finanziellen und technischen Rahmenbedingungen für die Dienste, die in einer disziplinübergreifenden Forschungsdaten-Infrastruktur in Deutschland angeboten werden sollten?
- Nach welchem Verfahren und nach welchen Kriterien sollte entschieden werden, ob Dienste stärker an dem generischen oder disziplin-spezifischen Modell orientiert aufgebaut werden?

In vier Break-Out-Sessions wurden diese und weitere Fragen diskutiert, wichtige Faktoren für eine übergreifende Forschungsdaten-Infrastruktur herausgearbeitet und Ideen für mögliche Umsetzungsmaßnahmen gesammelt.

## Session 1 – Policies und Anreize: Was sind sinnvolle und notwendige Richtlinien im Umgang mit Forschungsdaten?

Die Session wurde mit zwei Impulsreferaten eröffnet. Heinz Pampel (Helmholtz-Gemeinschaft, Open Access) sprach zum Thema „Data Policies – Zuckerbrot oder Peitsche“ gefolgt von Dr. Peter Lampen (ISAS e.V.), der über „Daten-Repositorien für die Spektroskopie“ berichtete.

In der anschließenden Diskussion wurde noch einmal das Problem der mangelnden Bereitschaft von Wissenschaftlern, ihre Daten offen zugänglich zu machen, erörtert. Dabei wurden zahlreiche Gründe genannt, die dafür verantwortlich sein können. Die Aufbereitung der Daten für die Veröffentlichung ist aufwändig. Viele Wissenschaftler haben ohnehin mit einer hohen Arbeitsbelastung zu kämpfen und wollen oder können diese zusätzliche Arbeit nicht leisten, zumal diese immer noch nicht entsprechend gewürdigt

wird (bspw. durch einen Gewinn an Reputation). Als mögliche Lösungen für dieses Problem wurde der Einsatz von Data Scientists, die die Aufbereitung der Daten übernehmen könnten, oder die Schaffung von Stellen mit 75% Forschungs- und 25% Datenpflege-Anteil vorgeschlagen. Aufwändig ist auch die Überprüfung der Daten vor der Veröffentlichung durch die Reviewer, welche diese ebenfalls in vielen Fällen nicht leisten können. Hier wurde eine Überarbeitung des Peer-Review-Verfahrens angemahnt. In vielen Bereichen fehlt die für ein geregeltes Forschungsdatenmanagement notwendige Infrastruktur, ebenso wie die Mittel, eine solche zu schaffen. Weiterhin fehlen vielen Wissenschaftlern die notwendigen Kenntnisse und Fertigkeiten des Datenmanagements. Es wurde ein hoher Bedarf an entsprechenden Schulungen und an der Integration des Themas in die Ausbildung konstatiert. In einigen Disziplinen, z.B. den Wirtschafts- und Sozialwissenschaften und der Medizin, stellen auch datenschutzrechtliche Bedenken eine Hürde dar, da in diesen Wissenschaften häufig mit personenbezogenen Daten geforscht wird. In manchen Disziplinen, als Beispiel wurden ebenfalls die Wirtschaftswissenschaften genannt, liegt ein großer Teil der Forschungsdaten in proprietären Formaten vor, was eine Veröffentlichung sehr schwierig macht. Schließlich wurde darauf hingewiesen, dass es auch viele Wissenschaftler gibt, welche die von ihnen erhobenen Daten gewissermaßen als ihr Eigentum betrachten und die Daten aus diesem Grund nicht mit anderen teilen wollen.

Trotz der zahlreichen möglichen Hindernisse ist es in einigen Disziplinen gelungen, in der Community ein Bewusstsein für die Notwendigkeit zur Zusammenarbeit und zum Teilen von Forschungsdaten zu schaffen. Als Erfolgsbeispiele wurden das Language Archive des MPI für Psycholinguistik in Nijmegen und die Molekularbiologie genannt, in der die für die Disziplin relevanten Journals klare Anforderungen stellen: Wer in diesen Journals die Ergebnisse seiner Arbeit veröffentlichen möchte, muss auch die zugehörigen Daten öffentlich zugänglich machen. Als weitere Beispiele führten die Workshop-Teilnehmer die Wirtschafts- und Sozialwissenschaften an, für die mittlerweile eine Infrastruktur mit zahlreichen Datenzentren besteht, sowie die Astrophysik, in der es ebenfalls üblich ist, zusammen mit einer Publikation auch die Forschungsdaten zu veröffentlichen.

Ausgehend von diesen Beispielen wurden mögliche Anreize für Datenmanagement und Data Sharing diskutiert. Es wurde festgestellt, dass gerade die angeführten Erfolgsgeschichten darauf basieren, dass der äußere Druck sehr groß war. Die Wissenschaftler sahen sich zur Kooperation gezwungen: Im Fall der Molekularbiologie war das Human Genome Project der Anlass. Man glaubte damals nicht, solch ein Vorhaben ohne die gemeinsame Anstrengung einer Vielzahl von Forschern bewältigen zu können. Im Fall des Language Archive sah sich die Community zur Zusammenarbeit gezwungen, um eine Infrastruktur zur Bewahrung von sprach- und kulturwissenschaftlichen Materialien zu schaffen, ohne die der Verlust einmaliger, unersetzlicher Aufnahmen ausgestorbener Sprachen gedroht hätte.

Viele Workshop-Teilnehmer waren deshalb der Ansicht, dass durch Förderorganisationen und Journals, aber auch durch die Universitäten ein entsprechender Druck in Form verbindlicher Regeln bzw. Vorschriften geschaffen werden sollte. Andere vertraten jedoch die Auffassung, dass äußerer Druck allein nicht das Mittel der Wahl sein kann. Vielmehr müsse man den Wissenschaftlern die Vorteile des Veröffentlichens der Daten näherbringen: dass nämlich die Ergebnisse ihrer Arbeit besser sichtbar sind. Auch muss darauf hingearbeitet werden, dass Wissenschaftler, die ihre Daten veröffentlichen, dafür mehr Anerkennung erfahren, als dies in den meisten Disziplinen bisher üblich ist. Es sollte daher auf eine entsprechende Änderung des wissenschaftlichen Reputationssystems hingewirkt werden.

Weiterhin berichteten einige Teilnehmer von ihren Erfahrungen bei der Implementierung bzw. dem Versuch der Implementierung von Policies in der eigenen Institution. Dies ist ein langwieriger und nicht immer

einfacher Prozess. Hürden bei der Schaffung institutioneller Policies sind z.B. mangelnde Ressourcen (Finanzen und Personal) und die Schwierigkeit, die Ansprüche mehrerer Communities unter einen Hut zu bringen. Auch ist die Frage bisher ungelöst, wie ein nachhaltiges Konzept mit Drittmittelfinanzierung umgesetzt werden kann.

Die Diskussion hat gezeigt, dass sich im Themenfeld Policies und Anreize noch zahlreiche Herausforderungen stellen. Gleichwohl gibt es Hoffnung, wie die oben angeführten Erfolgsgeschichten zeigen. Auch wurde berichtet, dass in der jüngeren Generation – den heutigen Studierenden – das Bewusstsein für die Wichtigkeit des Datenmanagements und den Nutzen des Veröffentlichens und Teilens von Daten gewachsen ist. Die Anforderung der DFG, bei der Antragstellung über das eigene Datenmanagement nachzudenken und dies im Antrag zu explizieren, trägt laut der Diskutanten ebenfalls zu einem Umdenken bei. Nicht zuletzt induziert auch in vielen Disziplinen die komplexe Natur der Forschungsgegenstände (z.B. globale Veränderungen, Biodiversität) den Wunsch zu umfangreicher Zusammenarbeit und mithin die Einsicht in die Notwendigkeit, sich mit dem Thema Datenmanagement auseinanderzusetzen.

## Session 2 – Einbindung in den Forschungsprozess: Kommen die Daten zur Infrastruktur oder die Infrastruktur zu den Daten?

Session 2 beschäftigte sich mit der Frage „Wie kommen die Daten in die Infrastruktur?“ In seinen einleitenden Worten wies Jens Klump darauf hin, dass eine Forschungsdateninfrastruktur ohne Daten sinnlos ist, die Aufbereitung von Daten und deren Überführung in die Infrastruktur jedoch der aufwendigste und teuerste Schritt im Lebenszyklus von Forschungsdaten ist.

Die einleitenden Impulsreferate stellten Beispiele aus der Biodiversitätsforschung (Anton Güntsch, BMBG Berlin) und aus der Materialforschung (Matthias Razum, FIZ Karlsruhe) vor. Die vorgestellten Beispiele diskutierten, ob eine Forschungsdaten-Infrastruktur besser **zentral oder dezentral organisiert** wird, und wie in beiden Szenarien die Erfassung und Vernetzung der Daten unterstützt werden kann.

Die anschließende Diskussion begann mit der Frage, wem die Aufgabe der Dokumentation von Forschungsdaten übertragen werden sollte, wie also die **Rollen und** die **Verantwortung** für die Daten verteilt sind.

Ein offensichtliches Problem besteht darin, dass es sich für die Wissenschaftler heute kaum lohnt, Zeit in die Pflege von Daten zu investieren. In manchen Fällen wird dies durch **Delegieren dieser Aufgabe an technisches Personal** möglich sein, dessen primäre Aufgabe die Pflege der Daten ist.

In den meisten Fällen wird es nicht möglich sein, die Aufgabe der Datenpflege zu delegieren, sie verbleibt beim Wissenschaftler. In diesem Fall müssen geeignete Anreize geschaffen werden. Genannt wurden:

- Datenveröffentlichung als Publikationsform
- Erhöhte Zitationsraten durch Verfügbarkeit der Daten
- Reproduzierbarkeit der Ergebnisse
- Einbindung der Daten in Linked Open Data

In der bisherigen Praxis haben diese Anreize allerdings bisher noch wenig Wirkung gezeigt. Das Potenzial der Datenveröffentlichung als Anreiz, die eigenen Daten zur Nachnutzung aufzubereiten, wurde von allen als hoch eingeschätzt. Als unbedingt notwendig wird hier die Entwicklung eines Systems der Qualitätssicherung ähnlich dem Peer-Review gesehen. Im Zusammenhang mit der Frage der Reproduzierbarkeit von wissenschaftlichen Ergebnissen durch die Veröffentlichung von Daten als begleitendes Material wurde auch

die Veröffentlichung von Software-Code diskutiert. Insbesondere bei der Veröffentlichung von Modellergebnissen ist die Nachvollziehbarkeit nur gegeben, wenn auch der Code, mit dem das Modell gerechnet wurde, mit zur Verfügung steht.

In Feldern, in denen heute schon erfolgreich Daten zur Nachnutzung bereitgestellt werden, besteht ein enger Zusammenhang zur kollaborativen Arbeitsweise der Wissenschaftler in diesen Feldern. So kann z.B. die Möglichkeit, Fahrtzeit auf einem Forschungsschiff zu beantragen, daran gebunden werden, dass die Daten einer abgeschlossenen Fahrt bis zu einem bestimmten Stichtag einem akkreditierten Datenarchiv übergeben werden. Dieses Verfahren wird bereits von der US National Science Foundation (NSF) eingesetzt.

Einigkeit herrschte darüber, dass die Erfassung der Metadaten von Forschungsdaten noch zu arbeitsaufwendig ist und die Werkzeuge dazu noch zu wenig benutzerfreundlich sind. Ziel muss eine weitgehende Automatisierung der Annotation von Forschungsdaten mit Metadaten sein, die verbleibenden Arbeiten müssen sich nahtlos in den Arbeitsalltag der Forscher einfügen.

## Session 3 – Generische vs. Disziplinspezifische Dienste: Was sind die Erfolgskriterien disziplinübergreifender Dienste?

Der Moderator Jens Ludwig (SUB Göttingen) führte in die Session und deren geplanten Ablauf ein. Die Diskussion hatte als Ziel, Kriterien und erfolgreiche, aber auch "mahnende" Beispiele zu sammeln. Fragestellungen der Session waren

- Die Erarbeitung einer Definition der Anforderungen an eine Forschungsdateninfrastruktur.
- Welche Einflussmöglichkeiten bestehen?
- Was sind Erfolgskriterien für disziplinübergreifende Dienste?
- Welches Potential haben disziplinübergreifende Dienste?

Die Session begann mit einem Impulsreferat. Brigitte Hausstein (GESIS) sprach zum Thema „da|ra – DOI-Registrierung für Sozial- und Geisteswissenschaften“.

In der anschließenden Diskussion kam die Frage auf, warum im Projekt da|ra eine neues Schema verwendet wird und nicht DDI (Data Documentation Initiative) / Cesta. Dies liegt laut Frau Hausstein daran, dass da|ra ein Subset von DDI / Cesta ist. Bei da|ra handelt es sich um eine reale Umsetzung und nicht nur ein theoretisches Modell. Probleme im Umfeld von da|ra sind die offene Suche über alle Metadaten sowie Lizenzeinschränkungen bei einigen Quellen. Die Diskussionen dazu sind im Fluss (freie zugängige Dokumente vs. kostenpflichtiger Zugriff). Ein wichtiger Aspekt bei da|ra bildet die Zielgruppe und die Akzeptanz durch die Zielgruppe. DataCite ist ein Cross-Community-Service. Die Metadaten sind allgemein und nicht community-spezifisch. da|ra als ein fachspezifisches Datenarchiv eine Ergänzung zu DataCite. Eine weitere Frage war, ob da|ra ein disziplinübergreifendes System darstellt. Frau Hausstein bejahte dies aus Sicht der Sozial- und Geisteswissenschaften. Aus Sicht der Wissenschaft allgemein verneinte sie dies jedoch. Es bestand Konsens darüber, dass da|ra ein community-spezifischer Dienst ist, während DataCite ein allgemeineres Dienstangebot darstellt.

In der Diskussion zu **Semantic Web / URI** wurde erörtert, dass es bereits Identifier für Daten gibt und sich demzufolge die Frage stellt, wie man mit diesen umgeht. Persistent Identifier (PID / DOI) sind bereits verbreitet, da sie offenbar gut zur Verwaltung von Massendaten ("sehr viele Anfragen") genutzt werden können. Andere Dienste, wie das klassische DNS, sind für Massendaten weniger geeigntet. Des Weiteren bieten sich, neben PID, noch andere Dienste an, z.B. Langzeitarchivierung (LZA) als Dienst für den wissenschaftlichen Alltag. In der Diskussion zu **Identifikation und Speicherung von Forschungsdaten**

(Bitstream Preservation) ergab sich, dass für Persistent Identifier ein Service zur Speicherung und Verwaltung der Metadaten, der Suche in den Metadaten sowie zur Auflösung der gefundenen Referenz benötigt wird. Außerdem müssen Identität, Integrität und Authentizität der Daten geklärt werden können. Der Einsatz der PID erfordert eine Registratur durch externe Anbieter. In Bezug auf **Daten** wurde diskutiert, dass Daten repliziert werden und die Sicherung der Identität, Integrität und Authentizität eine lokale Angelegenheit darstellt. Zum Thema **Langzeitarchivierung** wurde Bitstream Preservation als disziplinübergreifend definiert. Ebenso wie das "inhaltlich lesbar machen" der Daten. Kuration, Indexierung und Auffindbarkeit jedoch sind disziplinspezifisch.

In der Diskussion um „mahnende Beispiele“ ergab sich, dass

- bei Grids Software-Stacks zwar vorhanden sind, deren Funktionalität aber nicht optimal ist. Es gibt viele community-spezifische Softwarelösungen, Cross Community-Projekte (z.B. AI-Services) sind jedoch weltweit gescheitert.
- Mail mit X.400 hohe Komplexität bietet, jedoch in der Praxis als überholt gilt.

Die Session-Teilnehmer kamen überein, dass folgende Erfolgskriterien für die Schaffung disziplin-übergreifender Dienste gelten:

- Der Dienst darf nicht zu komplex sein
- Der Dienst muss sich auf jeden Fall am Bedarf orientieren
- Ein erster Schritt kann immer ein disziplinspezifischer Dienst (Prototyp) sein

Kriterien für einen disziplinübergreifenden Community-Dienst definieren sich wie folgt: Bedarfs-orientiertheit, Einfachheit, Vertrauen und Langlebigkeit, Umsetzbarkeit und Anwendbarkeit, sowie Anpassbarkeit und Flexibilität. Die Problematik generische vs. einfache Dienste ist also ein lösbares Problem.

## Session 4 – Möglichkeiten und Grenzen der Auslagerung und Zentralisierung von Diensten

Zu Beginn der Session gab es jeweils ein Impulsreferat von Prof. Dr. Wolfgang Nagel (TU Dresden, ZIH) und Stefan Heinzel (Rechenzentrum der MPG, Garching). Herr Nagel betrachtete in seiner Präsentation die Folgen der Veränderung der IT-Landschaft und erörterte verschiedene Möglichkeiten zum Umgang mit Forschungsdaten (Selbstverwaltung, Auslagerung an wissenschaftlichen Rechenzentren oder kommerzielle Anbieter). Herr Heinzel erläuterte in seiner Präsentation die Situation der Max-Planck-Gesellschaft in dieser Hinsicht.

Diskutiert wurden die Vor- und Nachteile der Selbstverwaltung von Forschungsdaten. Als Vorteil wurden die Möglichkeiten zur Kontrolle, der Selbstbestimmung und die flexible Anpassung der Konzepte an die eigenen Bedürfnisse gesehen. Nachteilig bewertet wurden der benötigte zeitliche Aufwand, die Menge der Daten, sowie Kosten für Hardware, Software, Personal und Unterhaltung der Services. Eine offene Frage bleibt, wie Nachhaltigkeit erreicht werden kann.

Thematisiert wurde die Auslagerung der Forschungsdaten an ein wissenschaftliches Rechenzentrum. Vorteil dieses Vorgehens ist, dass hier Expertise und Erfahrung im Umgang mit Forschungsdaten und eine Nähe zu den wissenschaftlichen Anwendern besteht. Nachteile sind, dass vielfach keine Service Level Agreements (SLA) bestehen. Des Weiteren besteht die Gefahr eines „Vendor-Lock-in“, d. h. es kann schwierig bis unmöglich sein, die Daten in ihrer ursprünglichen Form wieder zurück zu erhalten.

Weitere Überlegungen zu wissenschaftlichen Rechenzentren waren, ob es nicht sinnvoll wäre, wenn sich mehrere Zentren bzw. Gesellschaften zum Betrieb eines gemeinsamen Rechenzentrums zusammenschließen würden. Dagegen wiederum spricht, dass die nationale Ebene für ein Vorhaben dieser Art zu klein ist. Ein solches Zentrum sollte auf europäischer oder internationaler Ebene angesiedelt und Community-spezifisch sein. Schlussfolgerung war, dass eine zentrale Struktur von Nöten ist, um welche herum die Services gelagert werden können. Ebenso benötigt werden aber auch kleinere Strukturen in einem kleinen, lokalen Rechenzentrum. Beide Konzepte sind relevant und wichtig.

Bei einer Auslagerung der Daten an einen kommerziellen Anbieter sind die Vorteile, dass diese oftmals am günstigsten sind und Service Level Agreements (SLA) bieten. Nachteile sind die mit den Anforderungen steigenden Kosten, die vom Anwender aufgebracht werden müssen. D.h. der Anwender muss sich bereits im Voraus darüber im Klaren sein, was er braucht und möchte. Auch hier besteht die Gefahr des „Vendor-Lock-in“. Im Falle eines kommerziellen Anbieters hat der Nutzer keine absolute Sicherheit, ob die Datenschutzbestimmungen eingehalten werden. Es bleibt unklar, ob geistiges Eigentum sicher und ob es ausreichend ist, dieses mit Verschlüsselung zu sichern. Falls ja, stellt sich die Frage, ob es möglich sein wird, verschlüsselte Daten auch dann noch zu lesen, wenn der zugehörige Schlüssel seine Gültigkeit verloren hat. (Viele Schlüssel und Zertifikate verlieren nach 3 Jahren ihre Gültigkeit.) Auch SLAs schützen nicht immer vor Zwischenfällen. Zudem besteht bei kommerziellen Anbietern immer die Gefahr, dass existierende Dienste durch andere Firmen gekauft werden. Vielfach funktioniert dann nichts mehr so wie vorher. Bei der Verwendung kommerzieller Dienste sollte man demzufolge Vorsicht walten lassen.

Als eine Herausforderung wurde der Datentransport identifiziert. Die Datenmenge wächst schneller als die Bandbreiten des Internet. Es sollte somit eine Abwägung zwischen lokaler und zentraler Speicherung getroffen werden.

Diskutiert wurde die Nutzung von Diensten wie Skype und Dropbox, deren Verwendung aus Gründen des Datenschutzes im wissenschaftlichen Bereich meist nicht gestattet ist, die jedoch privat häufig genutzt werden, da sie einfach zu bedienen und praktisch sind.

In diesem Kontext wurde erörtert, warum die Wissenschaft nicht eigene Dienste äquivalent zu Skype und Dropbox entwickelt, um nicht in den Konflikt der Datenschutzproblematik der kommerziellen Dienste zu kommen. Wenn ein Dienst Mehrwert liefert und leicht nutzbar ist, wird er auch benutzt. Zudem sind Stabilität und Standardisierung wichtig, dies ist aber noch nicht genug kommuniziert.

Diskutiert wurde, welche Bedeutung die räumliche Nähe des Wissenschaftlers zu den Datenzentren habe. Fazit war, dass vor allem der Zugang zu Expertise wichtig sei. Erfahrungswerte sagen, dass der lokale Ansatz von Nutzern bevorzugt wird. Deswegen sind lokale Rechenzentren mit Expertise wichtig, dort müssen die Daten aber nicht liegen.

Auch stellte sich die Frage, welche Software und Anwendungen ein Wissenschaftler braucht. Was ist der adäquate Nachfolger des Laborbuches, das bisher alle Daten fasste? Die Software oder die Anwendung muss auf jeden Fall verständlich sein. Wünschenswert wäre, wenn sie so leicht zu bedienen wäre wie eine sog. App. Allerdings ist der Wissenschaftsprozess kompliziert und kann nicht als einfache App gestaltet werden. Defizite bestehen im Datenbereich bei Smart Tools und an Knowhow. Für manche Probleme fehlen zudem die Dienste, obwohl die Rechenleistung vorhanden ist.

Zum Thema Infrastruktur wurde angemerkt, dass mehr als nur technische Infrastruktur benötigt wird. Es wurde erörtert, welches die Hemmnisse übergreifender Infrastrukturen sind und ob die Bund-Land-Problematik eine Rolle hierbei spielt. Dabei blieb die Frage offen, wer übergreifende Infrastrukturen fördert.

Wäre eine zentrale User-ID für Wissenschaftler in Deutschland möglich? Welche Software-Infrastruktur wird benötigt, um den Wissenschaftler bestmöglich zu unterstützen? Es wurde geschlussfolgert, dass Infrastrukturen aus der Community heraus entstehen müssen, damit sie genutzt werden.

In Bezug auf Forschungsdatenmanagement wurde generell erörtert, wie das Thema überhaupt angegangen werden sollte. Sollte das Problem des Forschungsdatenmanagements aus der Sicht des Wissenschaftlers betrachtet werden? Oder bekommt man damit nur die lokale Sicht? Auch Wissenschaftler, die wenig Daten produzieren und nutzen, haben Anspruch auf ein angemessenes Datenmanagement. Die Differenzierung in „Big Data“ und „Small Data“ wurde kritisiert. Fazit der Diskussion war, dass am Ende derjenige gewinnt, welcher den besten Zugang zu Daten hat. Konsens bestand darüber, dass Wissenschaftler nicht alleine für das Datenmanagement verantwortlich sein sollten, sondern vor allem der Arbeitgeber und die Community.

Zu Fragen der Finanzierung wurde angemerkt, dass es auf Universitätsebene schwierig ist zu kommunizieren, dass mehr Geld für IT benötigt wird und oft das Bewusstsein für die Relevanz und den heutigen Stellenwert von IT fehlt.

**Abb. 5: Die Teilnehmer des Experten-Workshops am Tagungsort**

## Fazit

Die Wissenschaftler kamen überein, die Frage des zentralen oder dezentralen Angebots von Diensten stark von Volumen der Daten bestimmt wird. Im Vordergrund sollte in erster Linie die Diskussion um „Heterogene vs. Homogene Daten“ stehen und weniger die Frage, ob es sich um ‚Big Data‘ oder ‚Small Data‘ handelt. Die vordringlichste Frage sei, ob Infrastrukturen existieren, um heterogene, hochkomplexe Daten einfach darzustellen und ob Infrastrukturen existieren, um aus sehr großen Datenmengen Informationen zu extrahieren. Bislang existiert eine Vielfalt von Formaten für sowohl große als auch kleine, heterogene als auch homogene Datensätze, welche schwer zu verwalten ist.

Eine weitere Schlussfolgerung der Diskussion war, dass Werkzeuge für das Forschungsdaten-Management ein Niveau erlangen müssen, welches mit kommerziellen Werkzeugen vergleichbar ist. Desweiteren sollten Policies, Infrastrukturen und Anreizsysteme im Gleichklang behandelt und weiterentwickelt werden.

In der Diskussion wurde deutlich, dass auch ein kultureller Wandel in der Bewertung eines systematischeren Umgangs mit Forschungsdaten nötig ist und nachhaltige, nutzerfreundliche Anwendungen entwickelt werden müssen, welche sich nahtlos in die wissenschaftlichen Arbeitsabläufe einfügen.

# DFG-Projekt Radieschen – Experten-Workshop

## Elemente einer übergreifenden Forschungsdaten-Infrastruktur: Eine für Alle?

### 17.April 2012, GeoForschungsZentrum Potsdam (GFZ)

| | | |
|---|---|---|
| **10:30 – 11:00** | **Registrierung** | |
| **11:00 – 12:20** | **Eröffnungs-Session** | Moderation: Dr. Jens Klump<br>Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| | **Begrüßung** | Dr. Jörn Lauterjung<br>Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| | **Keynote 1** | Peter Wittenburg, Max-Planck-Institut für Psycholinguistik, Njimegen |
| | **Keynote 2** | Dr. Stefan Winkler-Nees, Deutsche Forschungsgemeinschaft (DFG) |
| | **Projektpräsentation Radieschen & Vorstellung der Breakout-Sessions** | Moderation: Dr. Jens Klump<br>Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| **12:20 – 13:00** | **Mittagspause** | |
| **13:00 – 15:00** | **Break-Out-Sessions** | |
| | **Session 1** – Policies und Anreize: Was sind sinnvolle und notwendige Richtlinien im Umgang mit Forschungsdaten | Heinz Pampel, Helmholtz Open Access Koordinationsbüro<br>Dr. Peter Lampen, Leibniz-Institut für Analytische Wissenschaften (IAIS)<br>Moderation: Stefan Strathmann, Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) |
| | **Session 2** – Einbindung in den Forschungsprozess: Kommen die Daten zur Infrastruktur oder die Infrastruktur zu den Daten? | Matthias Razum, FIZ Karlsruhe<br>Anton Guentsch, Botanischer Garten und Botanisches Museum Berlin (BGBM)<br>Moderation: Dr. Jens Klump<br>Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| | **Session 3** – Generische vs. Disziplinspezifische Dienste: Was sind die Erfolgskriterien disziplinübergreifender Dienste? | Brigitte Hausstein, GESIS, da\|ra<br>Moderation: Jens Ludwig, Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) |
| | **Session 4** – Möglichkeiten und Grenzen der Auslagerung und Zentralisierung von Diensten | Prof. Dr. Wolfgang Nagel, TU Dresden - ZIH<br>Stefan Heinzel, Rechenzentrum der Max-Planck-Gesellschaft am Institut für Plasmaphysik<br>Moderation: Dr. Harry Enke, Leibniz-Institut für Astrophysik Potsdam (AIP) |
| **15:00 – 15:30** | **Kaffeepause** | |
| **15:30 – 16:30** | **Vorstellung und Diskussion der Session-Ergebnisse** | Moderation: Dr. Leonie Schäfer<br>Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| **16:30 – 17:00** | **Zusammenfassende Diskussion und Perspektiven** | Moderation: Dr. Jens Klump<br>Deutsches GeoForschungsZentrum Potsdam (GFZ) |

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

MAX-PLANCK-GESELLSCHAFT

# Forschungsdaten-Infrastrukturen:
## bauen wir eine neuen Turm von Babylon?

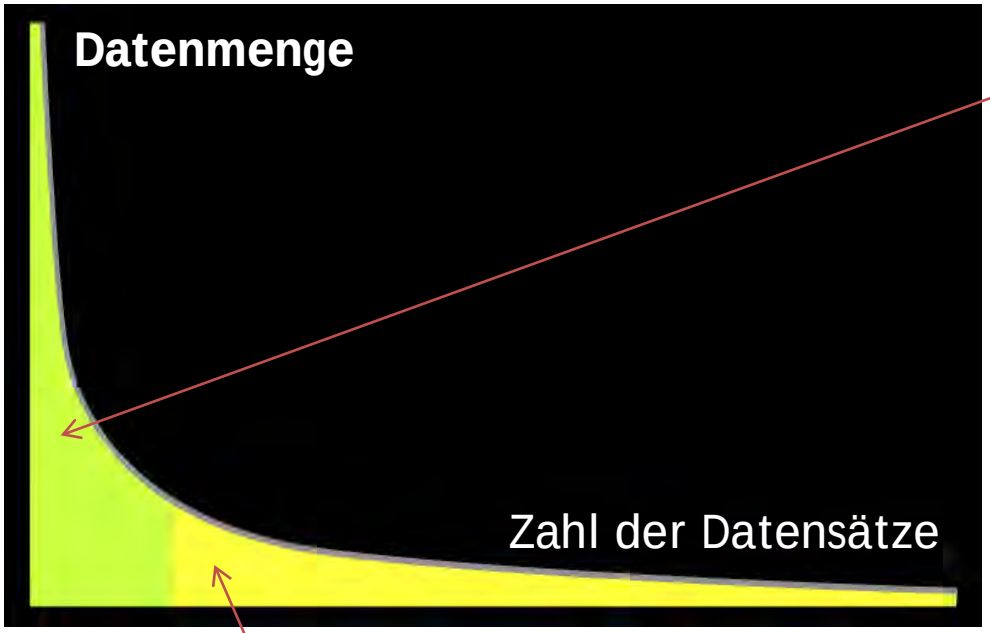
Peter Wittenburg
The Language Archive – Max Planck Institut für Psycholinguistik

# Integration und Interoperabilität

- e-Science & e-Humanities
  - “Daten sind die Währung der modernen Wissenschaft”
  - also brauchen Wiss. einen einfachen Zugriff auf die Datensätze
  - Datensätze:
    - verteilt über viele Repositorien => (virtuelle) Integration
    - sehr oft in schlechtem/r Zustand und Qualität => Kuration
    - durch verschiedene Teams erzeugt unter Verwendung verschiedener Konventionen (Struktur, Semantik) => Interoperabilität
  - folgerichtig sind I-K-I die mit am meisten verwendeten Wörter (Terminologie-Diff.) auf Konferenzen wie z.B. ICRI
- Große Frage: Wie können wir die I+I Barrieren überwinden, um große heterogene, distribuierte Datensätze effizient analysieren zu können? Das K Problem ignoriere ich mal eben.

# Datensätze

"Most of the bytes are at the high end, but most of the datasets are at the low end" – Jim Gray




"Big Data" durch Menge
- zumeist "raw" Daten
- im allgemeinen regulär

Herausforderungen
- Transport!!!, CPU!!, Storage!
- robuste automatische "Workflows" + Algorithmen
- Finden von Mustern
- Fitten von Parametern etc.

"Big Data" durch Datensätze
- zumeist abg
- sehr hetero
- spezielle Str
  semantische

Herausforderungen
- Datenmanagement
- Datenqualität und Kuration
- Interoperabilität

Science"
*mbole nur ein Beispiel*
elatierte Datensätze
nittels Wissen

# Was ist Interoperabilität?

1. Wikipedia: Interoperability is a property of a system, whose interfaces are completely understood, to work with other systems, present or future, without any restricted access or implementation.

2. IEEE: Interoperability is the ability of two or more systems or components to exchange information and to use the information that has been exchanged.

3. O'Brian/Marakas: Being able to accomplish end-user applications using different types of computer system, operating systems, and application software, interconnected by different types of local and wide-area networks.

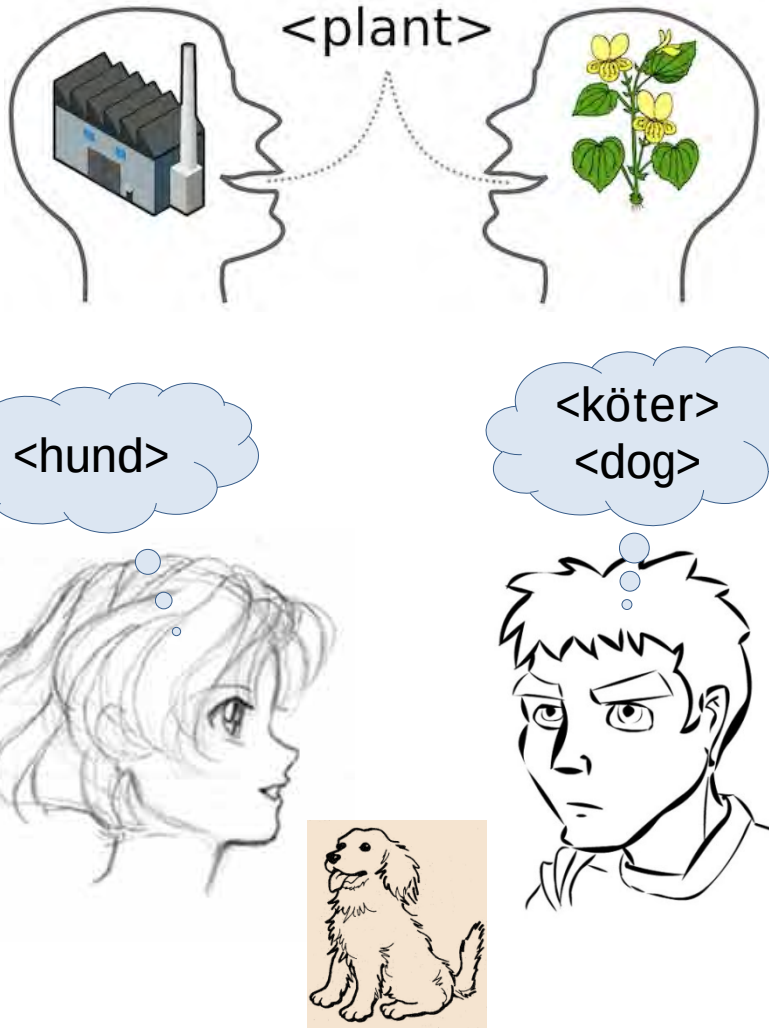
4. OSLC: To be interoperable one should actively be engaged in the ongoing process of ensuring that the systems, procedures and culture of an organization are managed in such a way as to maximise opportunities for exchange and re-use of information.

# Was ist Interoperabilität?

- Es geht um Brücken, die zwei bis N “Systemen” die Interaktion ermöglicht.

Menschen - Menschen
Menschen - Maschine
Maschine - Maschine
(relevant für Big Data)

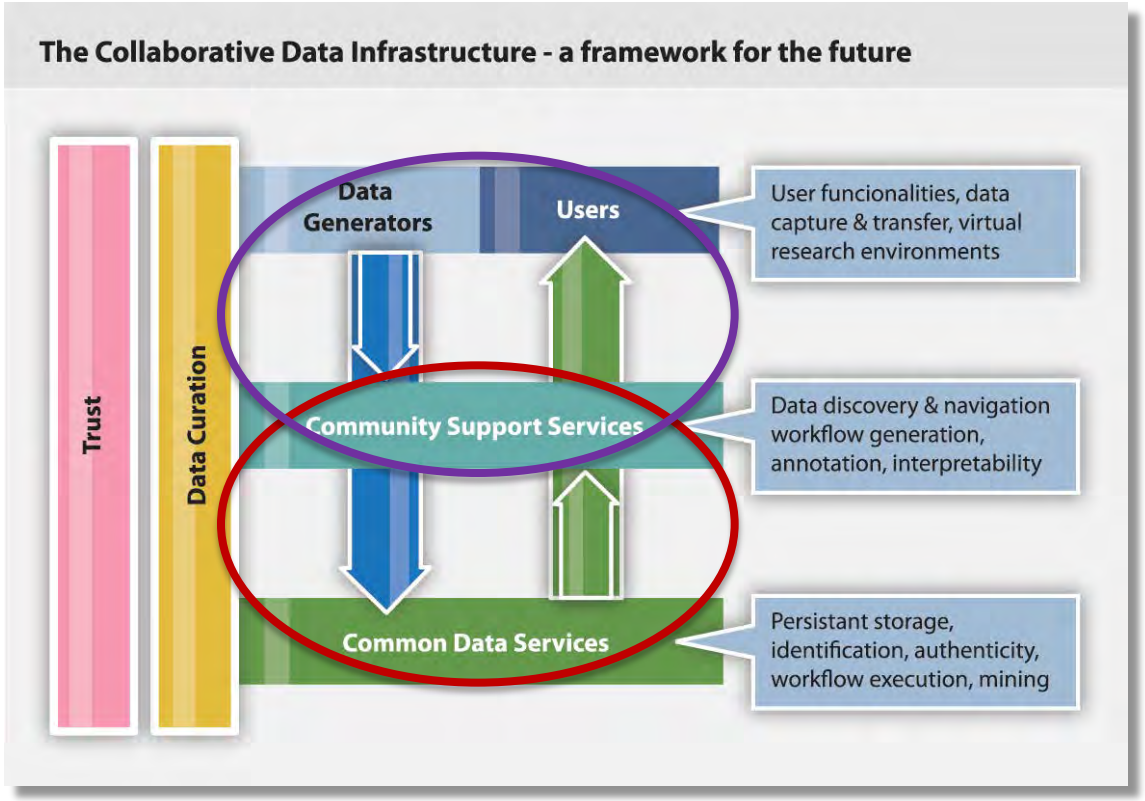
- bilität
- (techn. encoding, format, structure, API, protocol)
- Semantische Interoperabilität



# Was ist Integration?

- Die Interoperabilität sagt etwas darüber aus, ob zwei/N Systeme miteinander interagieren können.
- Der Begriff Integration beschreibt, wie sich derartige Systeme überhaupt erst einmal finden.

# Daten-Infrastrukturen sind komplex




Theoretisch einfach:

- eine Architektur, Datenorganisation + begrenzte Standards durchsetzen
- funktioniert eh nicht
- UND es ist bereits viel investiert worden

Wie sollen wir die inhärente I+C+I Komplexität verstehen?
Wie sollen wir die inhärente Komplexität verbergen?

- wir brauchen “Data Scientists” mit tiefen Kenntnissen
- diese müssen Interviews führen und an vielen konkreten I+C+I Projekten mitarbeiten

# warum Radieschen und EUDAT?

und die so wichtigen Radieschen Interviews

Kernaussagen:

- verschiedener Stand, alles bewegt sich und vieles ist beeinflussbar
- Heterogenität, aber Harmonisierungen sind möglich
- je näher an Wissenschaft, desto idiosynkratischer die Lösungen
- Trend zu kompatiblen Lösungen, aber Wissensmangel

# Können wir die Komplexität beherrschen?

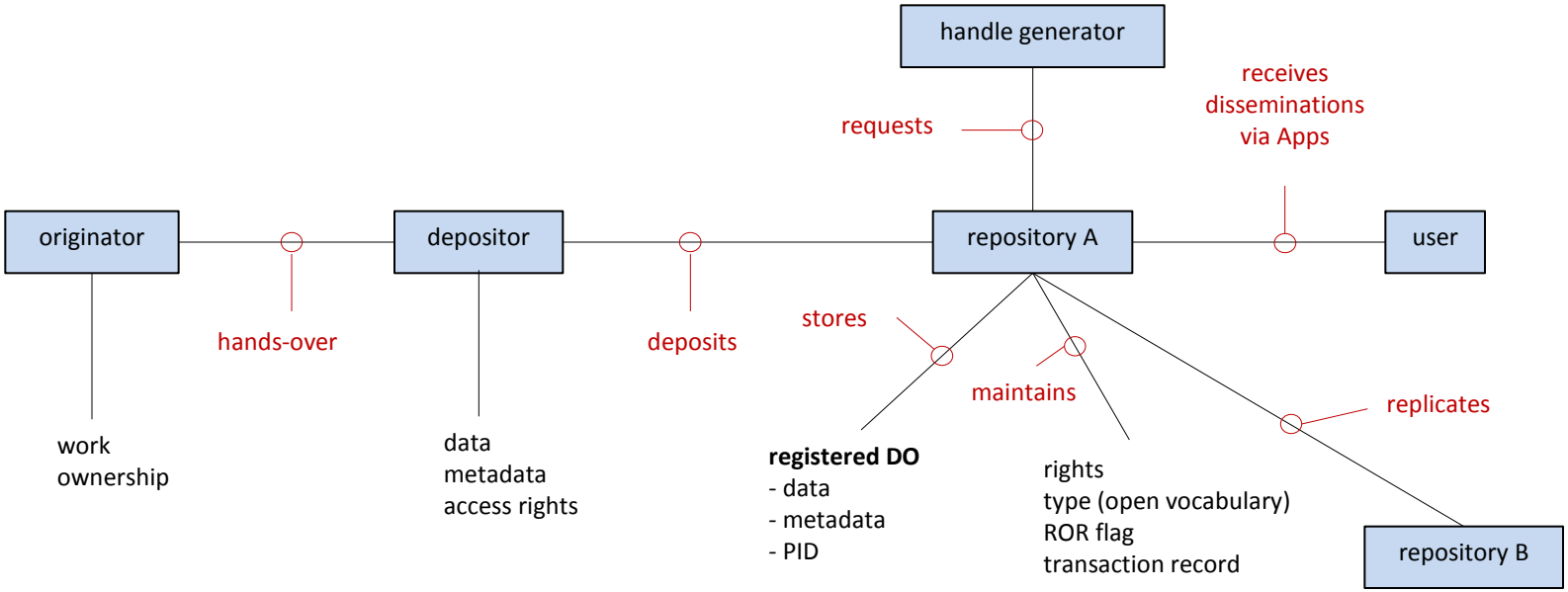
- der Turm als Wahrzeichen menschlicher Hybris
- das Resultat: Fluch der vielen Sprachen (Verlust des GOLD-Standards) und damit die Unmöglichkeit, Komplexität zu beherrschen
- müssen den Spieß wieder umdrehen ☺

## Wie kann das gelingen?

- konkrete Architekturen/Organisationen analysieren und abstrahieren
- individuelle Komponenten (Building Blocks) identifizieren, deren Eigenschaften definieren, diese realisieren und vertrauenswürdigen, starken Organisationen übergeben
- wo möglich auf soliden Komponenten und Protokollen aufbauen
- nichts Neues also

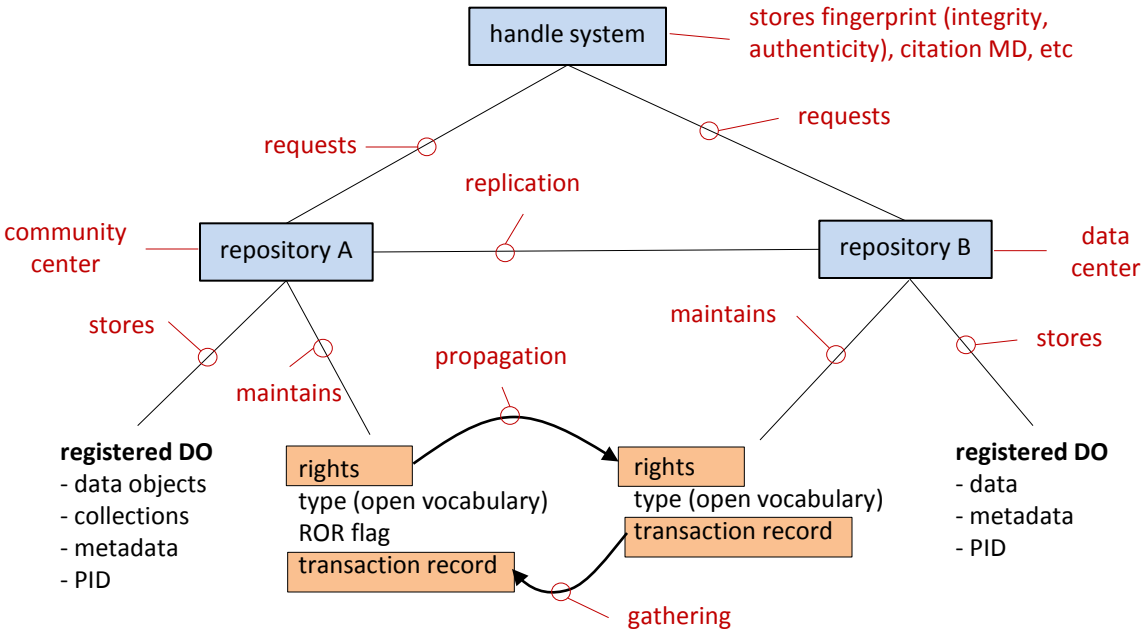
# Trennung externe/interne Eigenschaften

## Basis ist die Digitale Objekt Architektur (Kahn & Wilensky)




- jedes Datenobjekt hat externe Eigenschaften: einen Inhalt, Checksum, PID, MD, Relationen, Rechte, Lokationen, ROR Flagge, etc.
- diese sind für nahezu alle Disziplinen gleich
- brauchen separate Komponenten und Policy-Rules zur Manipulation
- Email Geschichte: brauchen Physiker und GW verschiedene Systeme?

# REPLIX - ein Beispiel




sichere Replikation auf Kollektions-Niveau mittels deklarativer Statements
- Beibehaltung aller externen Eigenschaften
- Verwendung externer Registraturen (MD, PIDs, Rechte DB+HT Access, etc.)
- Erfolgs-Checks

- saubere Trennung in Funktions- und Verantwortungs-Domänen (iRODS nur für die Policy-Rules und den Workflow, externer Handle Server von EPIC, eigene CLARIN MD Domäne, eigene Rechte-Verwaltung, etc.)
- weitere Extensionen: citation MD a la DataCite, transaction record aggregation
- Setup des LAT SW Stacks bei RENCI - Demo der Replikation nach 1 Monat fertig

# Welche Komponenten gibt es?

## Registraturen - Integration

- Persistente Identifikatoren (DOI, EPIC = Handles)
- “Trusted” Zentren
- Metadaten (Daten, Tools)
- Personen-IDs (ORCID)
- etc

## Registraturen - Interop

- Schemas
- Konzepte
- Relationen
- Ontolo

hinzukommt die maschinelle Nutzung
(Portale sind für menschliche Nutzung)
wir müssen die Attribute/APIs dringend harmonisieren

siehe Katasterämter

- Abmessungen, Eigner, Belastungen, Leitungen, etc.
- klare Autoritäten
- jeder kennt sie

# wie weiter - voher ein Blick zurück

- es gab TCP/IP und das ISO/OSI Model und endlose Theorie-Debatten
- es war NSF, die dem ein Ende bereitete, um vorwärts zu kommen
    - ein Projekt mit einem Netzwerk von Zentren wurde als Anker genommen
    - es war danach sehr schnell kein Thema mehr

- wer kann heute den Knoten durchhacken?
    - es ist genug geredet und wir kennen zumeist die verschiedenen Ansätze
    - es bedarf Entscheidungen
        - wir brauchen z.B. nur ein funktionierendes PID System
        - vielleicht verschiedene Business-Modelle
    - NSF ist kurz davor einen Call zu “Building Blocks” zu vergeben
    - und Europa?

# haben globalen Diskurs begonnen

- Data Access and Interoperability Task Force

  www.daitf.org

  – klingt wie IETF was kein Zufall ist
  – ein bottom-up Prozess ist essentiell
    getrieben durch die Wissenschaft
  – haben bereits mit zwei Workshops begonnen
    (Januar UIPIU, März Kopenhagen)
  – es gibt viele andere Initiativen - Interaktion erforderlich
  – von der NSF gibt es einen Governance Vorschlag: DWF

- EC ist bereit, den Prozess zu unterstützen
  EUDAT und OpenAIRE haben bereits Mittel
  iCORDI wird sich ganz auf den Prozess fokussieren
- US, Aussi, Kanada sind auch bereit
- andere Länder sind interessiert und waren in Kopenhagen ebenfalls vertreten
- Interessenten sollten sich melden - nächster Workshop ist in Washington

*Vielen Dank für Ihre Aufmerksamkeit.*

# Forschungsdaten-Infrastrukturen in Deutschland

## *Sind wir auf dem Weg in Richtung „Open Science“?*

Radieschen Experten Workshop
Dr. Stefan Winkler-Nees; Potsdam, 17. April 2012

# „Open Science“ ist kein neues Thema ...

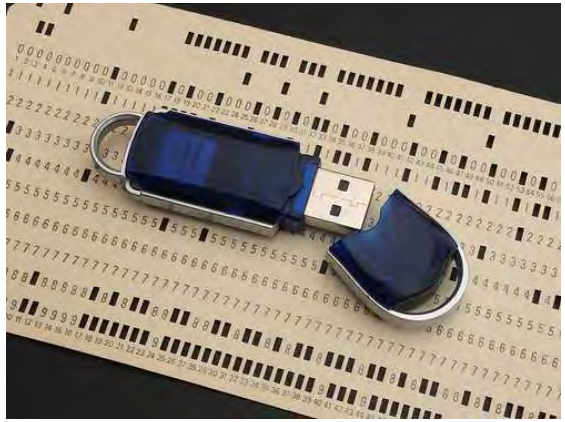
*"Die Deutschen, und sie nicht allein, besitzen die Gabe, die Wissenschaften unzugänglich zu machen."*

Johann Wolfgang von Goethe (1830)

Goethe‘s Werke: Vollständige Ausgabe letzter Hand. 22. Band: Wilhelm Meisters Wanderjahre oder die Entsagenden.

# Um was geht es? Forschungsdaten ...

- ... sind die Grundlage wissenschaftlicher Erkenntnis.
- ... werden heute nur unzureichend genutzt.
- ... sind nur eingeschränkt zugänglich.
- ... stehen nicht langfristig zur Verfügung.



Ian-S; http://www.flickr.com/photos/ian-s/2152798588; cc: by-nc-nd

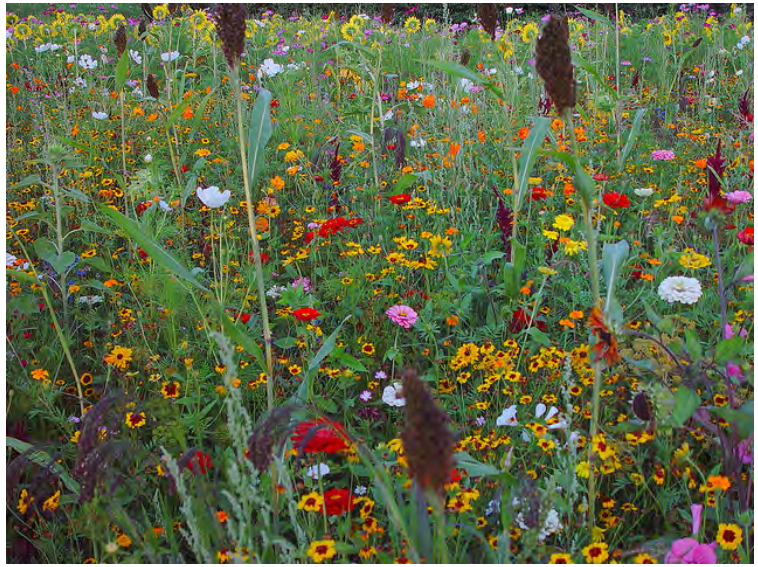
**Die Vision der Wissenschaftsorganisationen (und auch der Wissenschaftler/innen?):**

- Forschungsdaten sollen **möglichst frei** und **überregional zugänglich** und **langfristig verfügbar** sein.

# Bestandsaufnahme

## Forschungsdaten in Deutschland: eine bunte Wiese

- Wissenschaftler/innen
- Institutionen
- Informationsinfrastrukturen
- Wissenschaftsorganisationen
- Fördereinrichtungen



(Ela2007; http://www.flickr.com/photos/64479867@N00/3801475165/; cc BY-NC-ND 2.0)

# Bestandsaufnahme, siehe:

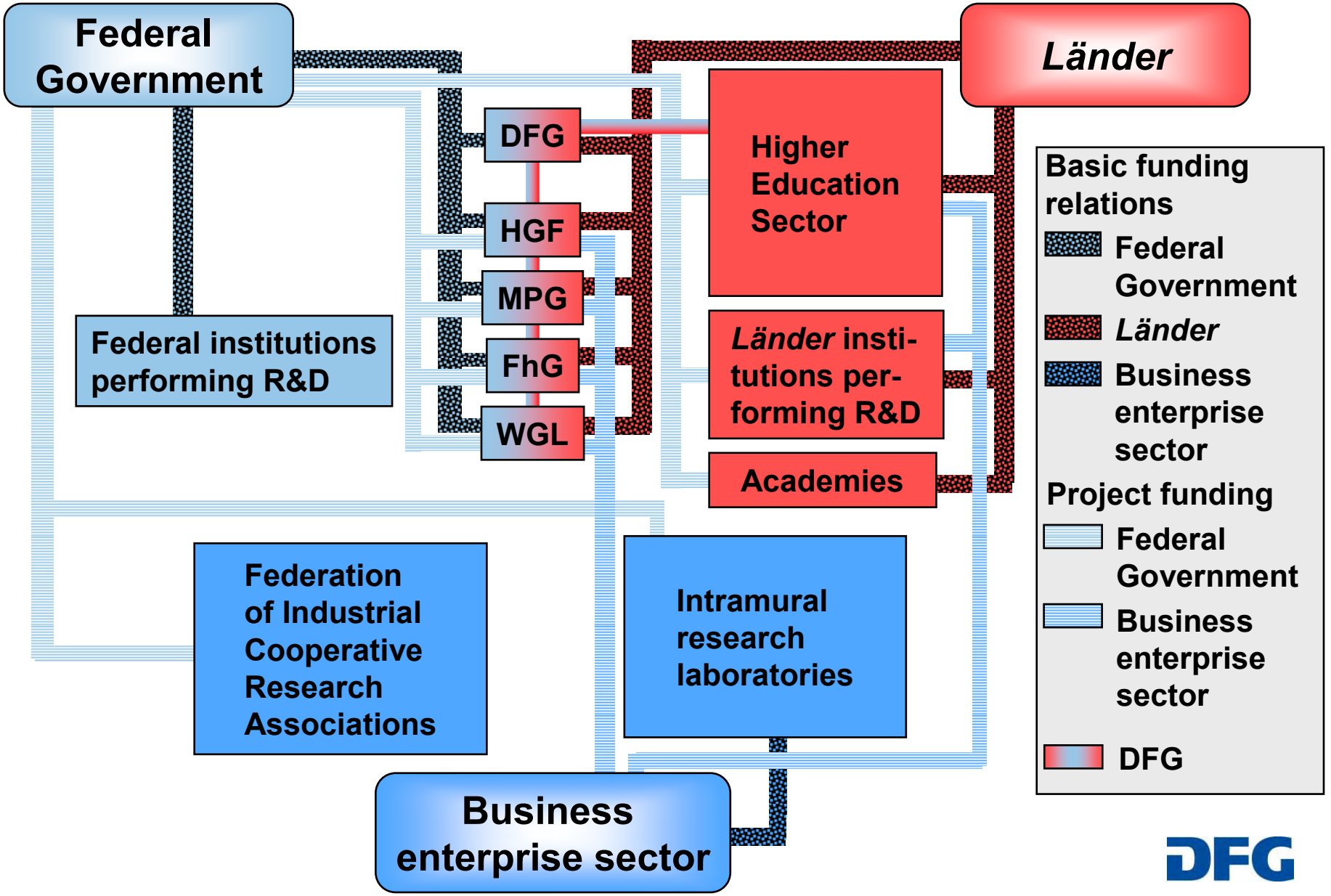
**Langzeitarchivierung von Forschungsdaten:**
**Eine Bestandsaufnahme**

Version 1.0 – 2012

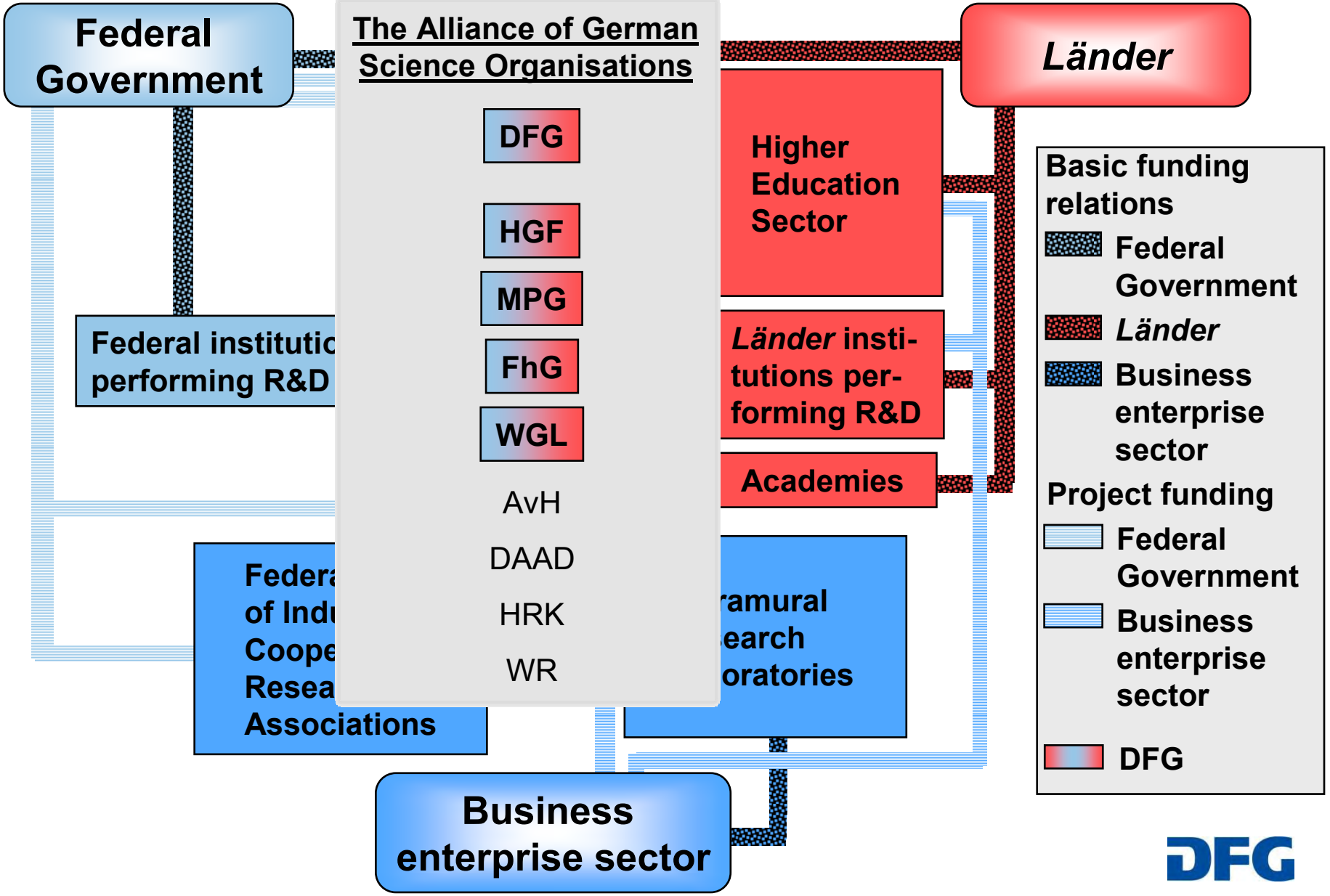
Hrsg. von: Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig

(http://nestor.sub.uni-goettingen.de/bestandsaufnahme/index.php)

# Structure of German Research Funding



# Structure of German Research Funding

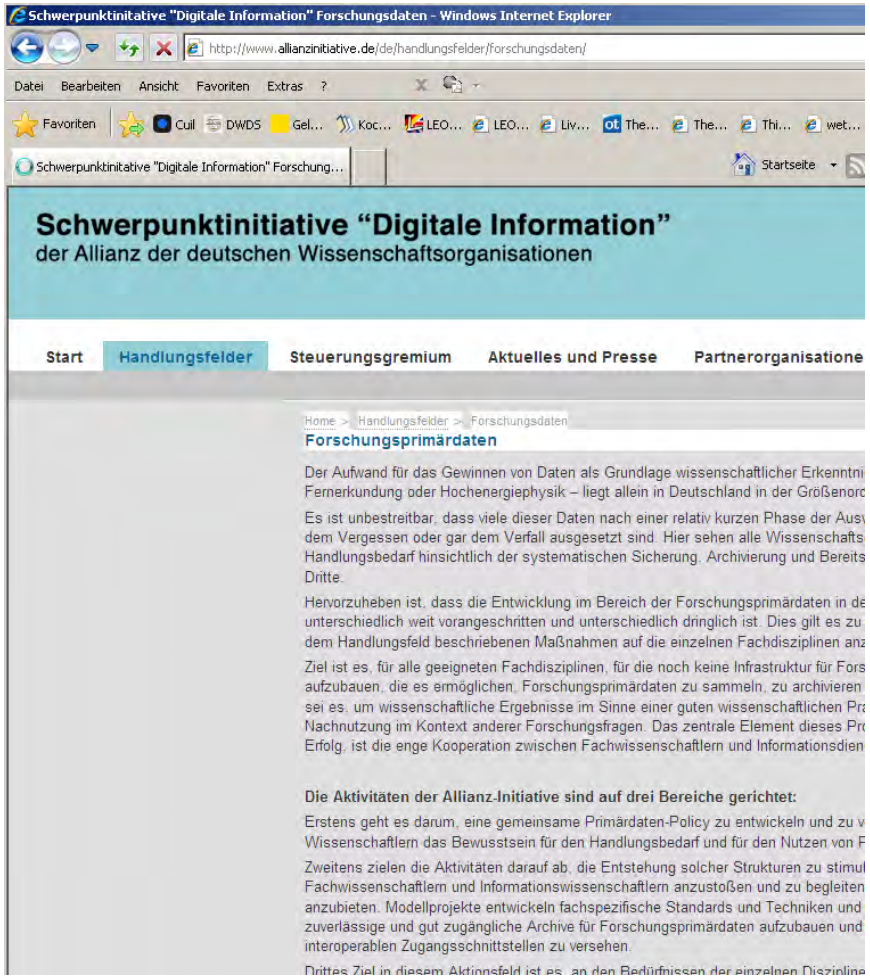


# Die Allianz der Deutschen Wissenschaftsorganisationen

- **Die DFG gemeinsamt mit den Allianz Partnern:**
  Max-Planck-Gesellschaft
  Helmholtz Gemeinschaft
  Leibniz Gemeinschaft
  Frauenhofer Gesellschaft
  und weitere

- **22. Juni 2008: Schwerpunktinitiative "Digitale Information"**

- **Die Schwerpunkte sind:**
  Nationale Lizenzierung
  Open Access
  Nationale Hosting Strategie
  **Forschungsdaten**
  Virtuelle Forschungsumgebungen
  Rechtliche Rahmenbedingungen



# Die Allianz Grundsätze

Die Grundsätze der Allianz zum Umgang mit Forschungsdaten 2010

- Sicherung und Zugänglichkeit
- Unterschiede der wissen-schaftlichen Disziplinen
- Wissenschaftliche Anerken-nung
- Lehre und Qualifizierung
- Verwendung von Standards
- Entwicklung von Infrastruk-turen

**Grundsätze zum Umgang mit Forschungsdaten**

Die Grundsätze wurden von

der Allianz der deutschen Wissenschaftsorganisationen

am 24. Juni 2010 verabschiedet

Alexander von Humboldt-Stiftung
Deutsche Akademie der Naturforscher Leopoldina
Deutsche Forschungsgemeinschaft (DFG)
Deutscher Akademischer Austausch Dienst (DAAD)
Fraunhofer-Gesellschaft
Helmholtz-Gemeinschaft
Hochschulrektorenkonferenz (HRK)
Leibniz-Gemeinschaft
Max-Planck-Gesellschaft
Wissenschaftsrat

Die Grundsätze sind auch abrufbar unter:
http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten

# Handlungsfelder der Allianz in 2012

*Forschungsdatenmanagement*

Es werden zwei Stränge verfolgt:

- Fokussieren des WissGrid Leitfadens auf eine institutionelle Policy
- Entwickeln einer fachspezifische Policy.

*Nachnutzung und Verfügbarkeit von Forschungsdaten*

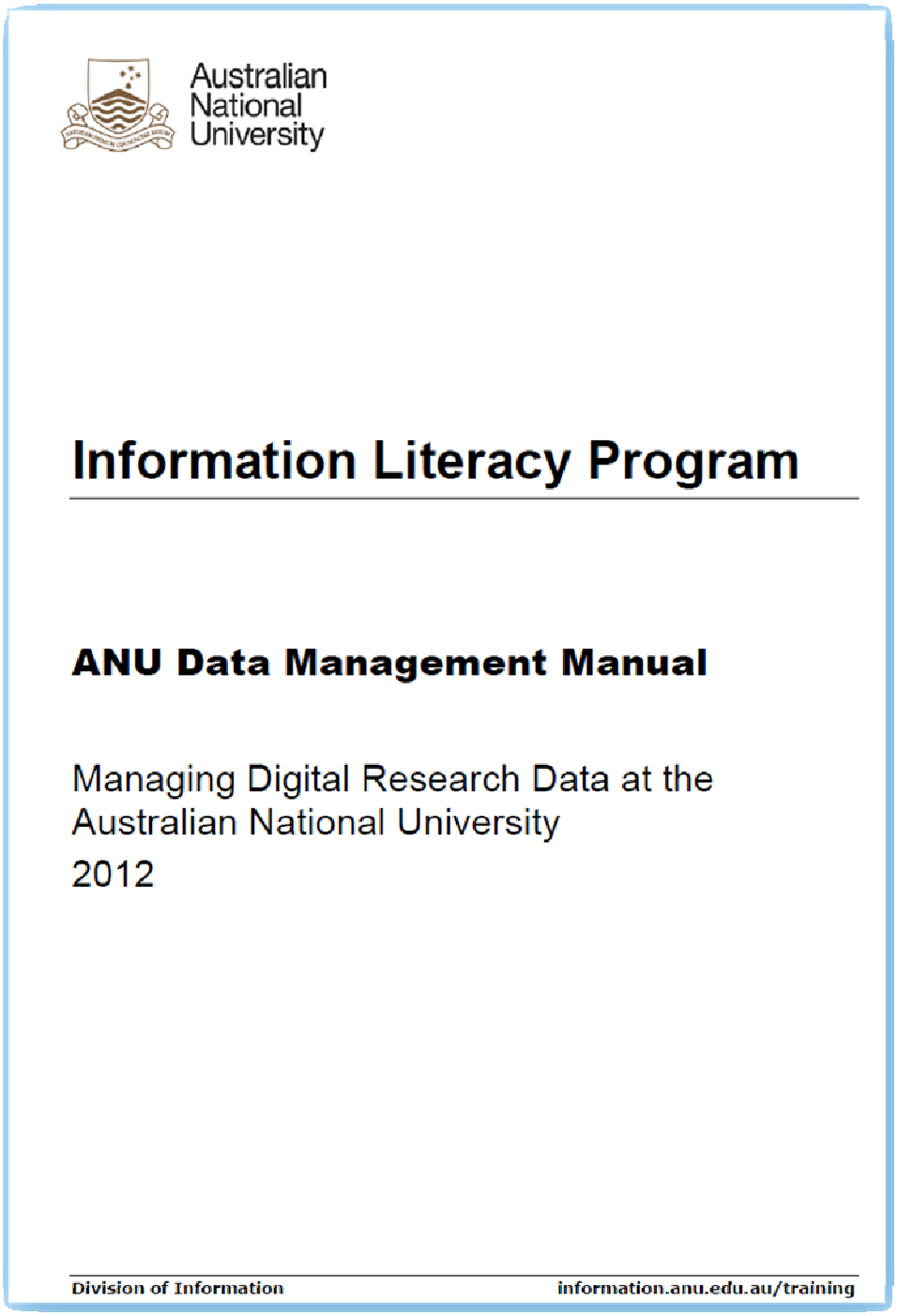
- Sammlung von Success-Stories.

# Beispiel eines institutionellen Leitfadens

Bietet:

- Allg. Informationen zum Daten Management
- Anforderungen und Vorteile
- Methoden des Daten Management
- Angaben zu ANU's Daten Management Dienstleistungen
- Anleitung zum Schreiben eines Daten Management Plans

http://information.anu.edu.au/training_and_skills_development/information_literacy/



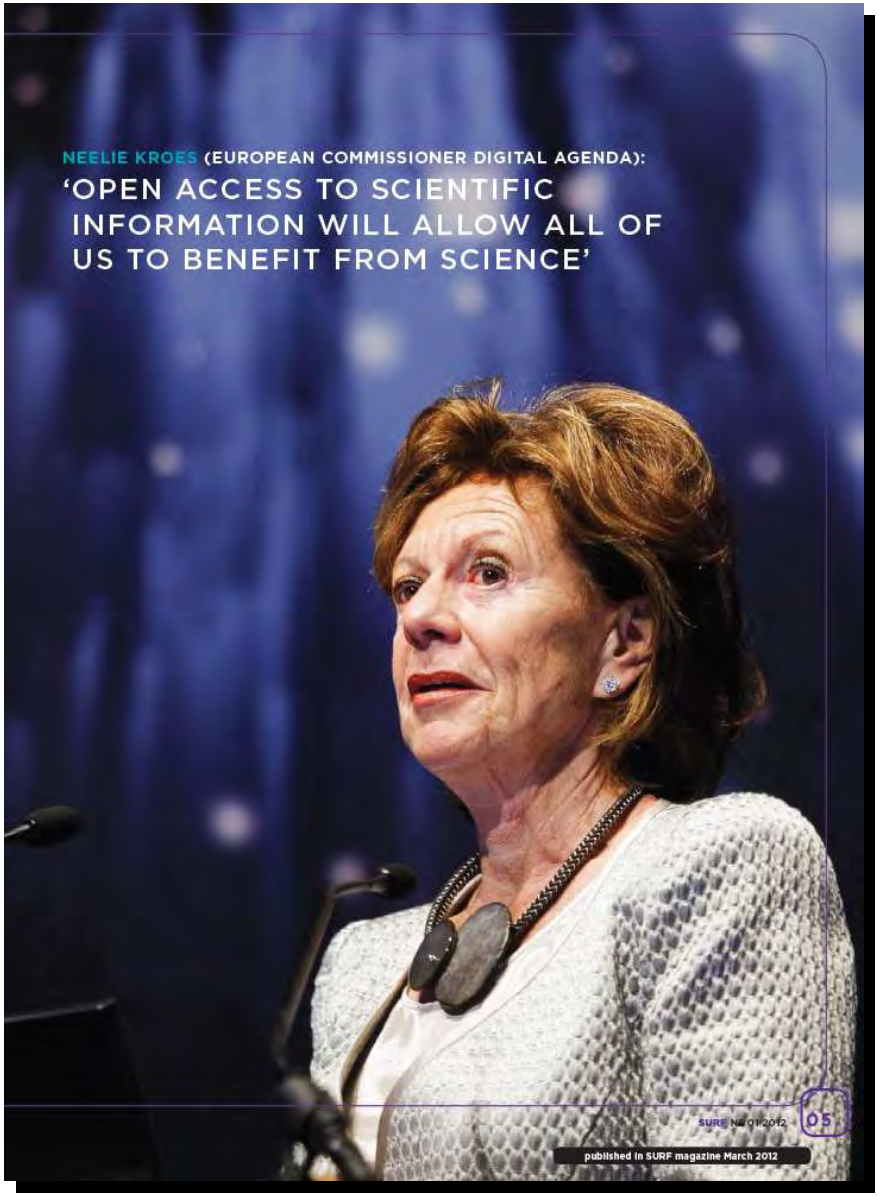
# Handlungsfelder der Allianz in 2012

*Kostenstrukturen*

- Untersuchung der Kostenstrukturen einzelne Datenzentren, evtl. mit Hilfe einer Studie.

*Juristische Rahmenbedingungen*

- Erstellung eines Übersichtspapiers mit einer Statusbeschreibung, Handlungsoptionen und Vorschlägen.


NEELIE KROES (EUROPEAN COMMISSIONER DIGITAL AGENDA):
'OPEN ACCESS TO SCIENTIFIC INFORMATION WILL ALLOW ALL OF US TO BENEFIT FROM SCIENCE'
SURF NR 01 2012
05
published in SURF magazine March 2012

Open Science for the 21st century
A declaration of ALL European Academies
presented at a special session
with Mme Neelie Kroes, Vice-President of the European Commission,
and Commissioner in charge of the Digital Agenda
on occasion of the ALLEA General Assembly held at
Accademia Nazionale dei Lincei, Rome, on 11-12 April 2012
ALLEA | ALL European Academies
The European Federation of National Academies of Sciences and Humanities

# Vielen Dank für Ihre Aufmerksamkeit!

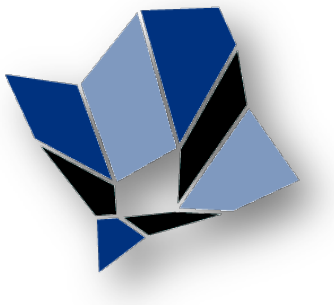
Stefan.Winkler-Nees@dfg.de



DFG

# DFG-Projekt Radieschen

## Rahmenbedingungen einer disziplin-übergreifenen Forschungsdateninfrastruktur



# Projektdaten

- Laufzeit: Mai 2011 – April 2013
- 6 Partner aus unterschiedlichen Disziplinen und Wissenschaftsorganisationen
- Förderung durch
  DFG „Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS)“

# Projektpartner

ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen

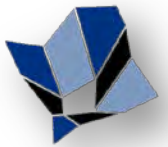
GFZ
Helmholtz-Zentrum
POTSDAM

MAX-PLANCK-INSTITUT
FÜR PSYCHOLINGUISTIK

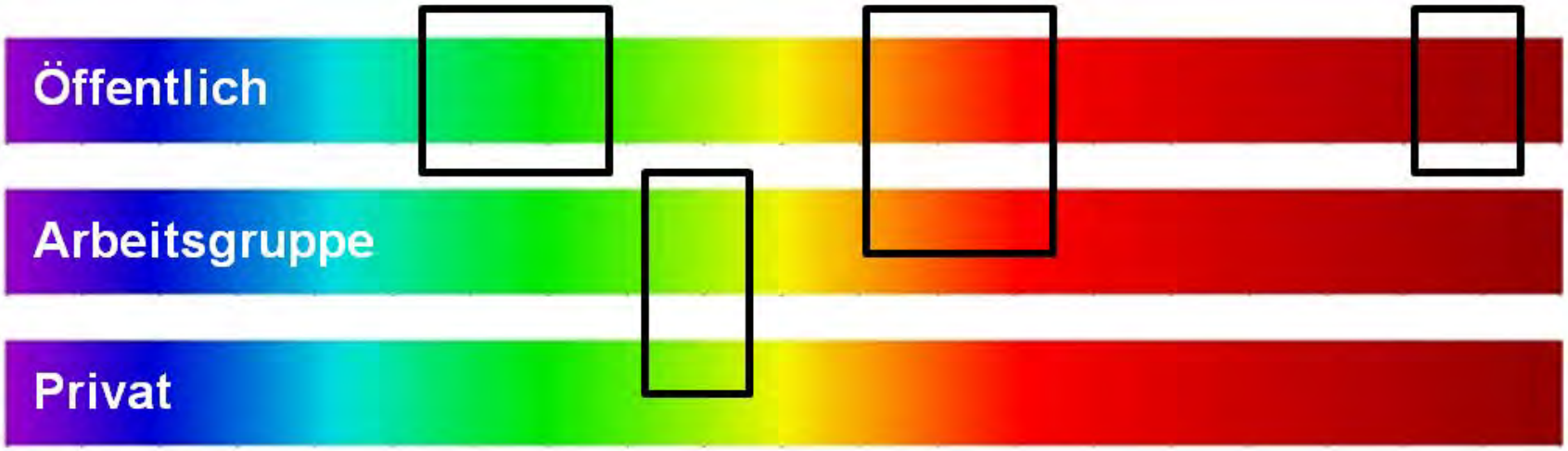
DKRZ
DEUTSCHES
KLIMARECHENZENTRUM

SUB | Niedersächsische Staats- und Universitätsbibliothek Göttingen

AIP

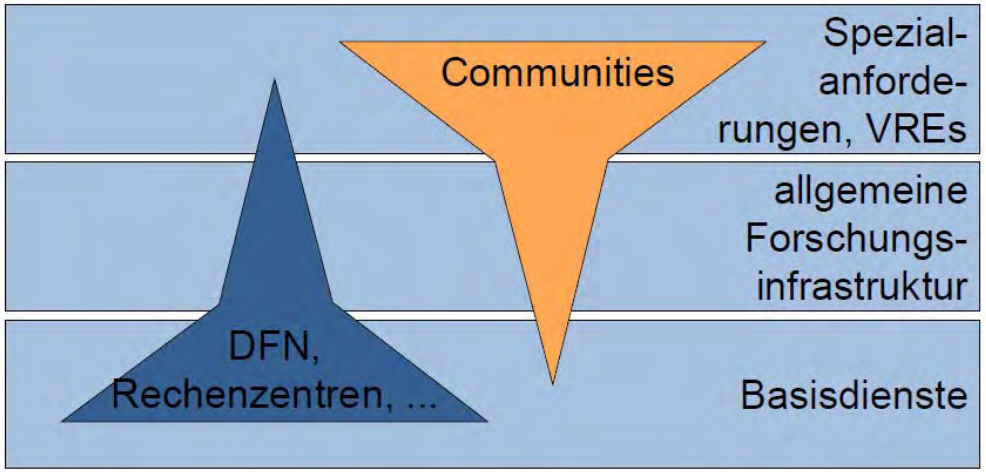


# Status Informationsinfrastrukturen




- Breites Spektrum an Anforderungen
- Radieschen untersucht, welche Bereiche des Spektrums bereits abgedeckt sind

# Verortung der Akteure

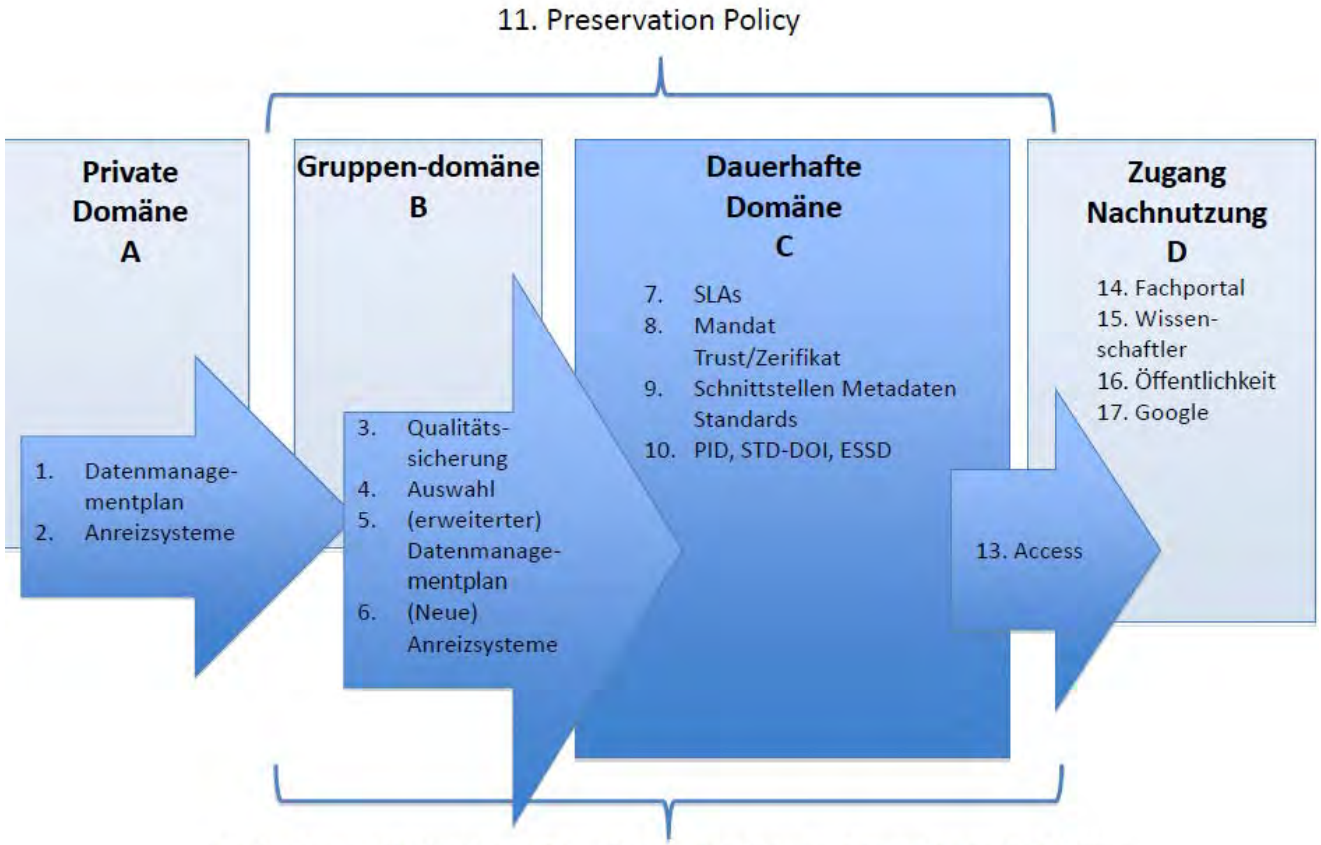



Akteure unterscheiden sich nach

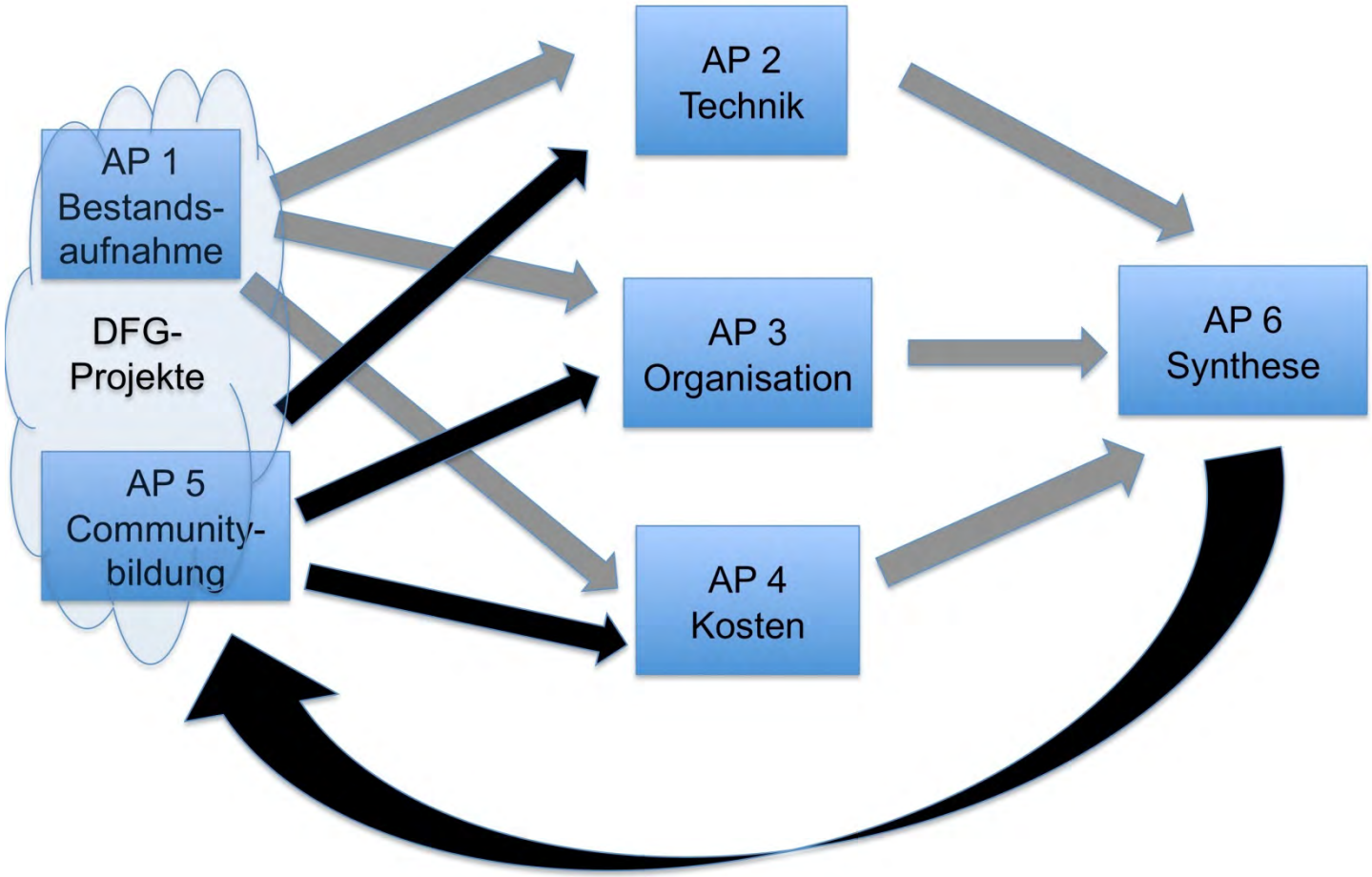
- Arbeitsschwerpunkten
- Vorstellungen
- Herangehensweisen

=> Diskussionsbedarf „Forschungsinfrastruktur“

# Arbeitsablauf im Umgang mit Forschungsdaten



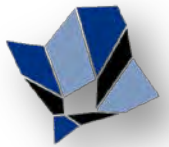
# Arbeitspakete & Projektablauf



# Projektziele

- Bericht Technik:
  - Anforderungen und Entwicklungsbedarf
  - Dokumentation vorbildlicher, disziplinspezifischer Lösungen
- Bericht Organisation:
  - Empfehlungen zur Einbettung des Forschungsdaten-Managements in die Forschungsinfrastruktur
- Bericht Kosten:
  - Analyse der Kosten des Forschungsdaten-Managements
  - Fokus auf generische Faktoren und Risikoabschätzungen

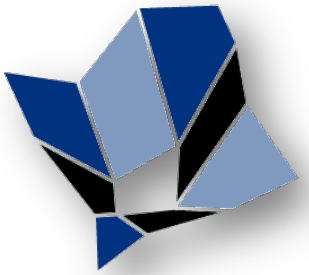
=> **Roadmap** für eine Forschungsdaten-Infrastruktur in Deutschland



# www.forschungsdaten.org

# Danke für Ihre Aufmerksamkeit !

Kontakt:
info@forschungsdaten.org
http://www.forschungsdaten.org/



# DFG-Projekt RADIESCHEN

Workshop Materialien

des Experten-Workshops

vom 17.April 2012

## Session 1 – Policies und Anreize:

Was sind sinnvolle und notwendige Richtlinien

im Umgang mit Forschungsdaten

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

# „Data Policies“ - Zuckerbrot oder Peitsche?

## Session 1: Policies und Anreize: Was sind sinnvolle und notwendige Richtlinien im Umgang mit Forschungsdaten?

Experten-Workshop „Forschungsdaten-Infrastruktur“ | Potsdam, 17.04.2012

Heinz Pampel | Helmholtz-Gemeinschaft

HELMHOLTZ | GEMEINSCHAFT

Open Access

# UNSERE MISSION

- Forschung zur **Lösung wichtiger Zukunftsfragen** von Gesellschaft, Wissenschaft und Wirtschaft – strategisch und langfristig orientiert
- Bau und Betrieb großer **Forschungsinfrastrukturen** („think big, act big“)
- Erkenntnisse zum **Nutzen** von Gesellschaft und Wirtschaft umsetzen

HELMHOLTZ GEMEINSCHAFT

Open Access

# EUROPEAN COMMISSION (2012)

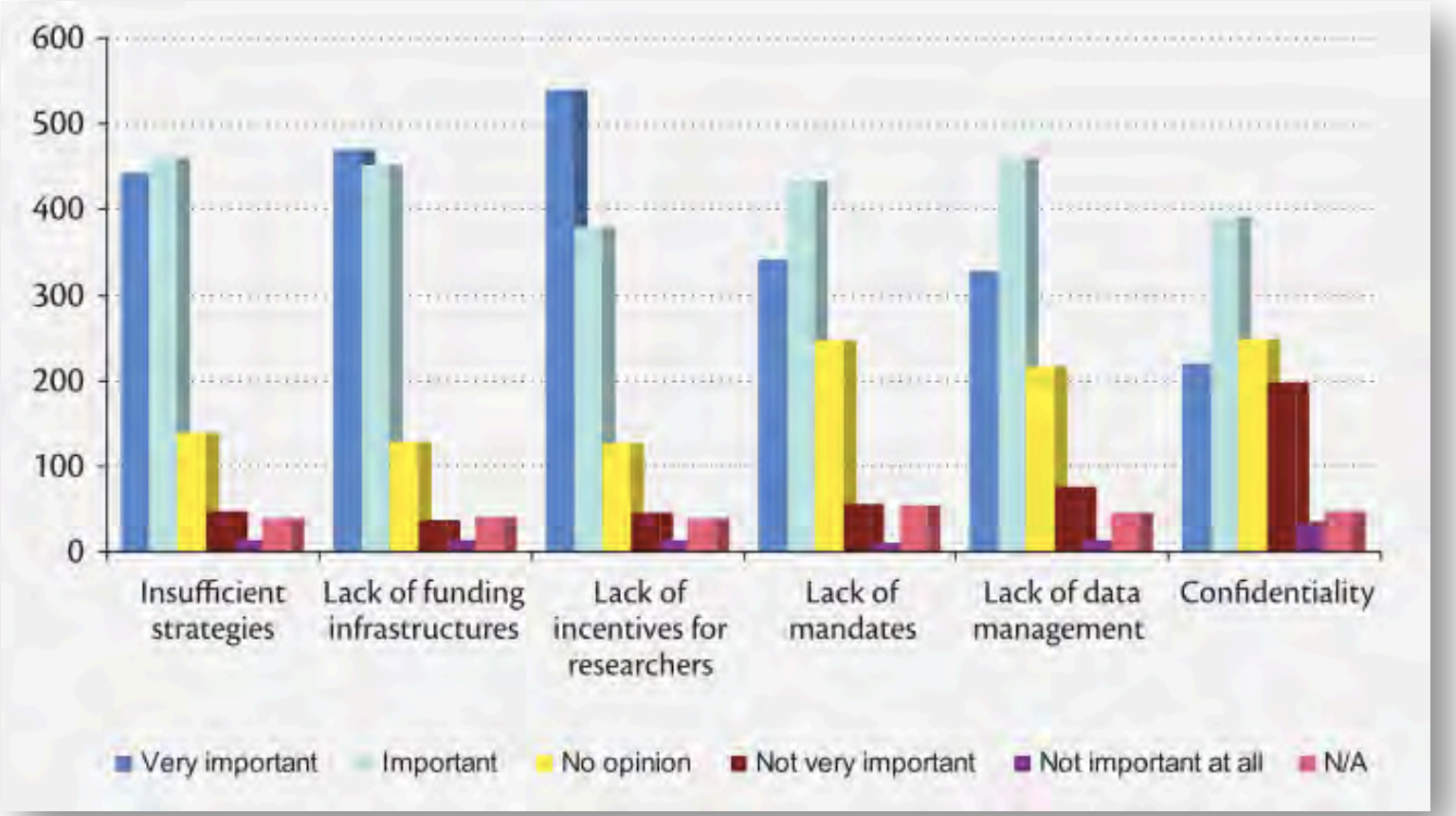
- „As for the question of access to research data, the vast majority of respondents (87 %) disagreed or disagreed strongly with the statement that there is **no** access problem for research data in Europe."
- „The barriers to access research data considered very important or important by respondents were:
  - lack of funding to develop and maintain the necessary infrastructures (80 %);
  - the insufficient credit given to researchers for making research data available (80 %);
  - and insufficient national/regional strategies/policies (79 %)."
- „There was strong support (90 % of responses) for research data that is publicly available and results from public funding to be, as a matter of principle, available for reuse and free of charge on the Internet."
- „Lower support (72 % of responses) was given for data resulting from partly publicly and partly privately funded research."

HELMHOLTZ GEMEINSCHAFT

Open Access

# EUROPEAN COMMISSION (2012)

- „How would you rate the importance of the following potential barriers to enhancing access to research data?




HELMHOLTZ GEMEINSCHAFT

Open Access

# KUIPERS & VAN DER HOEVEN (2009)

- „Researchers consider the **possibility of re-analysis of existing data** as the most important driver for the preservation of research data; 91% of the respondents thought this to be either important or very important.“
- „Researchers regard the **lack of sustainable hardware, software or support of computer environment may make the information inaccessible** as the most important threat to digital preservation. 80% believe this to be either important or very important.“
- „58% of the research respondents believe that an **international infrastructure for data preservation and access** should be to help guard against some of the above mentioned threats.“
- „25% of the researcher make their data openly available for everyone.“
- „Major barriers for sharing research data are **the fear of researchers regarding legal issues and the misuse of their data**.“

HELMHOLTZ GEMEINSCHAFT

Open Access

# ANFORDERUNGEN

Förder-
organisationen

Politik

Repositorien

Forschungs-
infrastrukturen

Wissenschaftler

Institutionen

Zeitschriften

Communities

Projekte

Gesellschaft

RRZE Icon Set (CC: BY-SA)

HELMHOLTZ GEMEINSCHAFT

Open Access

# BEISPIELE FÜR POLICIES

- Interdisziplinäre Policies
  - DFG: Vorschläge zur Sicherung guter wissenschaftlicher Praxis (1997)
  - Allianz: Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen (2003)
  - OECD: Principles and Guidelines for Access to Research Data from Public Funding (2007)
  - Allianz: Grundsätze zum Umgang mit Forschungsdaten (2010)
- Disziplinäre Policies
  - Bermuda Principles (1996)
  - IPY 2007-2008 (2007)
- Institutionelle Policies
  - U Siegen: Verfahrensrichtlinien zur Sicherung einer guten wissenschaftlichen Praxis (2001)
  - STFC: Scientific Data Policy (2011)
- Förderer Policies
  - NIH: Data Sharing Policy (2003)
  - DFG: Leitfaden für Antragsteller (2010)
- Journal Policies
  - Nature: Editorial Policy (2009)
  - PLoS ONE: Editorial Policy (o. J.)

# VAN DER GRAF & WAAIJERS (2011)

| Funding organisations requirements | Data plan | Access/ sharing | Long-term curation | Monitoring | Guidance | Costs |
|---|---|---|---|---|---|---|
| UK | | | | | | |
| AHRC | + | + | +/- | - | + | - |
| BBSRC | + | + | + | + | + | + |
| CRUK | + | + | + | + | - | - |
| EPSRC | + | + | + | + | - | + |
| ESRC | + | + | + | + | + | + |
| MRC | + | + | + | - | +/- | - |
| NERC | + | + | + | + | + | - |
| STFC | - | + | - | - | +/- | - |
| Wellcome Trust | + | + | + | + | + | + |
| Germany | | | | | | |
| Deutsche Forschungs Gemeinschaft | (+) | (+) | (+) | | | (+) |
| Denmark | | | | | | |
| Council for Independent Research | - | - | - | - | - | - |
| Council for Strategic Research | - | - | - | - | - | - |
| The Netherlands | | | | | | |
| NWO: arts and humanities social sciences | + | + | + | + | + | + |
| NWO: other scientific disciplines | - | - | - | - | - | - |
| STW | - | - | - | - | - | - |
| Senter Novem | - | - | - | - | - | - |

HELMHOLTZ GEMEINSCHAFT

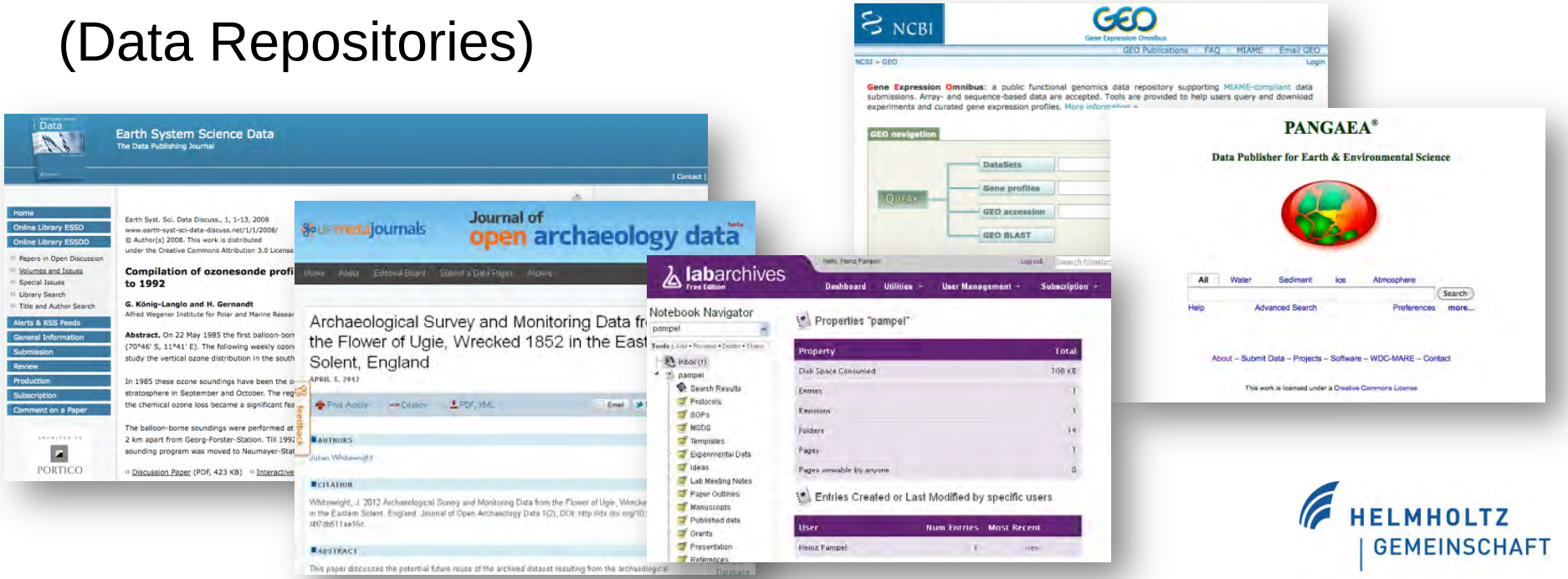
Open Access

# DATA POLICIES: ANMERKUNGEN

- Verschiedene Akteure mit gemeinsamen Interessen
  - Förderung von Nachprüfbarkeit und Nachnutzung
- Spannungsfeld zwischen Empfehlung und Verpflichtung
- Unterstützende Infrastrukturen, Dienstleister und „best practice“ werden benötigt
- Anpassungen im Laufe der Zeit nötig
- Ohne Anreize wenig Chancen auf Umsetzung
- Zentrale Akteure: Förderorganisationen u. Herausgebergremien
- Kulturelle Perspektive muss berücksichtigt werden
  - „sharing & openess“ als Kulturtechnik

# ANREIZE (IN AUSWAHL)

- Verankerung im Reputationssystem (Credit)
- Schaffung einer Publikationskultur (Data Journals)
- Angebot an wissenschaftsnahen Tools (VREs)
- Etablierung von dienstleistungsorientierten Infrastrukturen (Data Repositories)






HELMHOLTZ GEMEINSCHAFT

Open Access

# OPPORTUNITIES FOR DATA EXCHANGE

- http://ode-project.eu

Opportunities for Data Exchange

CERN · Science & Technology Facilities Council · HELMHOLTZ ASSOCIATION · CSC

APA ALLIANCE FOR PERMANENT ACCESS

BRITISH LIBRARY · DEUTSCHE NATIONAL BIBLIOTHEK · LIBER · stm

SEVENTH FRAMEWORK PROGRAMME

HELMHOLTZ GEMEINSCHAFT

Open Access

# OPPORTUNITIES FOR DATA EXCHANGE

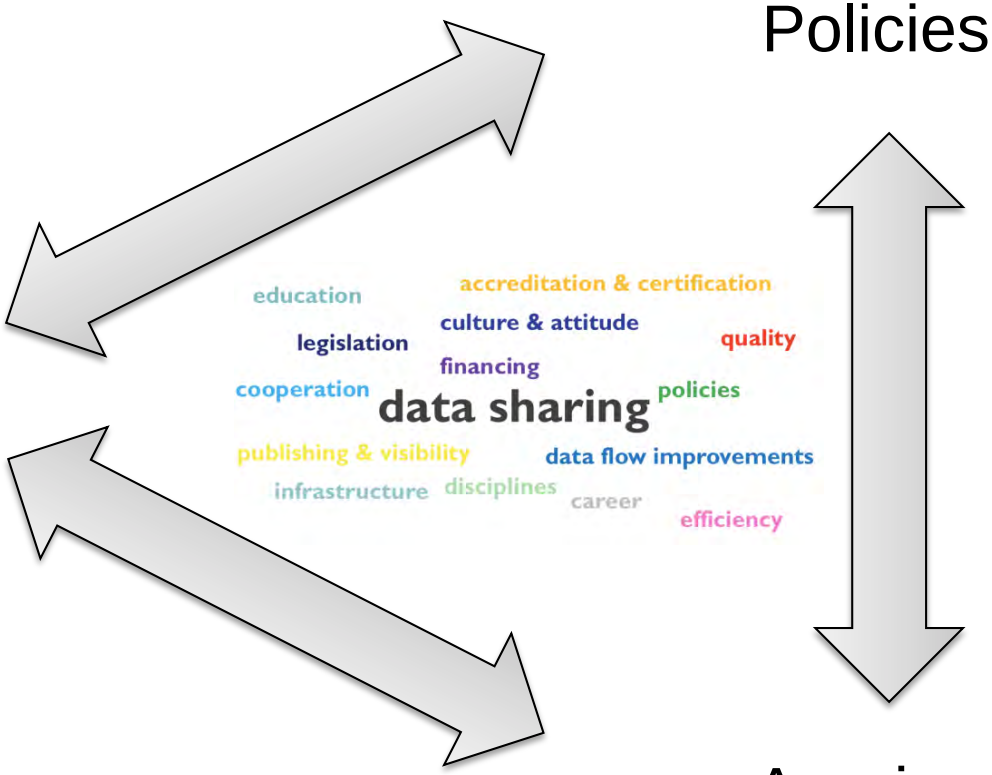
- Themenfelder, die das „data sharing“ beeinflussen:

Schäfer et al. 2011

HELMHOLTZ GEMEINSCHAFT
Open Access

# SPANNUNGSFELD

Bereitschaft zum „data sharing“




HELMHOLTZ GEMEINSCHAFT

Open Access

# POLICIES, ANREIZE UND…

- Verankerung in Workflows („Data Curation Continuum“)
- Entwicklung von Standards
- Berücksichtigung in der Ausbildung von Fachwissenschaftlern
- Schaffung von Ansprechpartnern und Dienstleistern
- Aufbau und Vernetzung von Infrastrukturen
- Kooperation und Kollaboration
- Entwicklung neuer bibliometrischer Metriken (altmetrics.org)
- Förderung von „Leuchttürmen“
- Übergreifende internationale Initiativen
- Forschung und Entwicklung

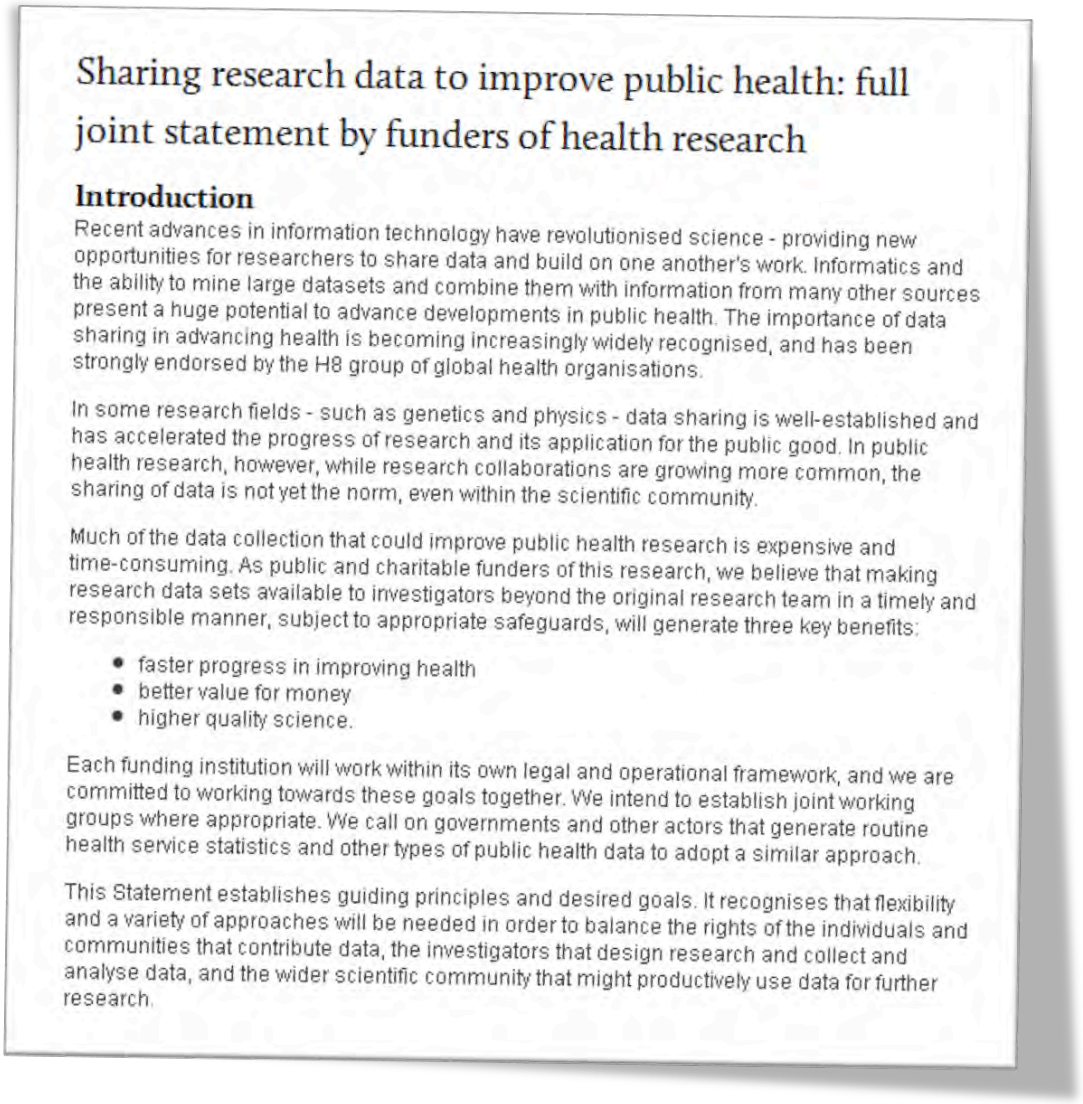
# FAZIT

- Policies und Anreize sind zwei von vielen Faktoren
- Data Policies benötigen unterstützende Maßnahmen
- Spannungsfeld zwischen Policies und Anreizen beachten
- Ohne Anreize wenig Chancen auf Umsetzung
- Rahmenbedingungen müssen definiert werden
  - W-Fragen
- Rahmenbedingungen müssen geschaffen werden
- Kooperation und internationale Initiativen fördern



HELMHOLTZ GEMEINSCHAFT
Open Access

# “SHARING RESEARCH DATA TO IMPROVE PUBLIC HEALTH”



Sharing research data to improve public health: full joint statement by funders of health research

**Introduction**

Recent advances in information technology have revolutionised science - providing new opportunities for researchers to share data and build on one another's work. Informatics and the ability to mine large datasets and combine them with information from many other sources present a huge potential to advance developments in public health. The importance of data sharing in advancing health is becoming increasingly widely recognised, and has been strongly endorsed by the H8 group of global health organisations.

In some research fields - such as genetics and physics - data sharing is well-established and has accelerated the progress of research and its application for the public good. In public health research, however, while research collaborations are growing more common, the sharing of data is not yet the norm, even within the scientific community.

Much of the data collection that could improve public health research is expensive and time-consuming. As public and charitable funders of this research, we believe that making research data sets available to investigators beyond the original research team in a timely and responsible manner, subject to appropriate safeguards, will generate three key benefits:

- faster progress in improving health
- better value for money
- higher quality science.

Each funding institution will work within its own legal and operational framework, and we are committed to working towards these goals together. We intend to establish joint working groups where appropriate. We call on governments and other actors that generate routine health service statistics and other types of public health data to adopt a similar approach.

This Statement establishes guiding principles and desired goals. It recognises that flexibility and a variety of approaches will be needed in order to balance the rights of the individuals and communities that contribute data, the investigators that design research and collect and analyse data, and the wider scientific community that might productively use data for further research.

- Initiative siebzehn lebenswissenschaftlicher Förderorganisationen
- Walport & Brest (2011)

HELMHOLTZ GEMEINSCHAFT

Open Access

# LITERATUR


Allianz der Wissenschaftsorganisationen. (2010). Grundsätze zum Umgang mit Forschungsdaten. Retrieved from http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze/

Berliner Erklärung. (2003). Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Retrieved from http://oa.mpg.de/lang/de/berlin-prozess/berliner-erklarung/

Deutsche Forschungsgemeinschaft. (1998). Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft." Weinheim: Wiley-VCH. Retrieved from http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

Deutsche Forschungsgemeinschaft. (2010). Merkblatt für Anträge auf Sachbeihilfen mit Leitfaden für die Antragstellung. DFG-Vordruck 1.02 – 8/10. Retrieved from http://www.dfg.de/download/programme/emmy_noether_programm/antragstellung/1_02/1_02.pdf

European Commission. (2012). Online survey on scientific information in the digital age. Luxembourg: Publications Office of the European Union. Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf

International Polar Year 2007-2008 Data Policy. (2007). Retrieved from http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf

Kuipers, T., & Van der Hoeven, J. (2009). Insight into digital preservation of research output in Europe. Survey Report. Framework. Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

National Institutes of Health. (2003, April 12). Final NIH Statement on Sharing Research Data. Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

Organisation for Economic Co-operation and Development. (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. Paris: OECD Publications. Retrieved from http://www.oecd.org/dataoecd/9/61/38500813.pdf

Pampel, H., & Bertelmann, R. (2011). „Data Policies" im Spannungsfeld zwischen Empfehlung und Verpflichtung. In S. Büttner, H.-C. Hobohm, & L. Müller (Eds.), Handbuch Forschungsdatenmanagement (pp. 49-61). Bad Honnef: Bock + Herchen. Retrieved from http://opus.kobv.de/fhpotsdam/volltexte/2011/228/

PLoS ONE. (n.d.). PLoS ONE Editorial and Publishing Policies. Sharing of Materials, Methods, and Data. Retrieved February 20, 2011, from http://www.plosone.org/static/policies.action#sharing

Schäfer, A., Pampel, H., Pfeiffenberger, H., Dallmeier-Tiessen, S., Tissari, S., Darby, R., Giaretta, K., et al. (2011). Baseline Report on Drivers and Barriers in Data Sharing. Retrieved from http://ode-project.eu/outputs

Science and Technology Facilities Council. (2011). STFC scientific data policy. Retrieved from http://www.stfc.ac.uk/Resources/pdf/STFC_Scientific_Data_Policy.pdf

Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing. (1996). Retrieved from http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml

Treloar, A., & Harboe-Ree, C. (2008). Data management and the curation continuum. How the Monash experience is informingrepository relationships. Melbourne. Retrieved from http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

Universität Siegen. (2001). Grundsätze und Verfahrensrichtlinien zur Sicherung einer guten wissenschaftlichen Praxis an der Universität Gesamthochschule Siegen. Retrieved from http://www.uni-siegen.de/start/forschung/kombibox_forschung/grundsaetze_und_verfahrensrichtl_wiss_praxis.pdf

Van der Graaf, M., & Waaijers, L. (2011). A Surfboard for Riding the Wave. Towards a four country action programme on research data. Retrieved from http://www.knowledge-exchange.info/Admin/Public/DWSDownload.aspx?File=/Files/Filer/downloads/Primary+Research+Data/Surfboard+for+Riding+the+Wave/KE_Surfboard_Riding_the_Wave_Screen.pdf

Walport, M., & Brest, P. (2011). Sharing research data to improve public health. Lancet, 377(9765), 537-9. Retrieved from http://dx.doi.org/10.1016/S0140-6736(10)62234-9


HELMHOLTZ GEMEINSCHAFT

Open Access

# Danke für Ihre Aufmerksamkeit!

Heinz Pampel

open-access@helmholtz.de

http://oa.helmholtz.de

Alle Texte dieser Präsentation sind unter einem Creative Commons Namensnennung 3.0 Deutschland Lizenzvertrag lizenziert. http://creativecommons.org/licenses/by/3.0/de

HELMHOLTZ GEMEINSCHAFT

Open Access

# Daten-Repositorien für die Spektroskopie

Peter Lampen

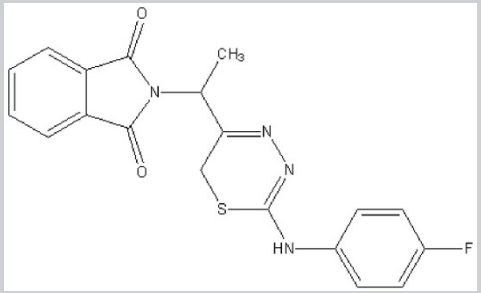
Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V.

Dortmund

# Vergleich zwischen der Anzahl der bekannten chemischen Verbindungen und den größten Spektren-Sammlungen

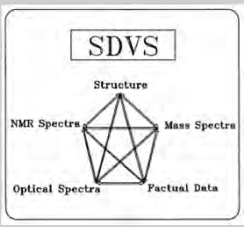
- Chemical Abstracts Service (CAS): 60.000.000 registrierte organische und anorganische Verbindungen (Stand 24. Mai 2011)

- CAS Registry Nummer: 1298016-92-8




- Wiley Registry of Mass Spectral Data, 9th Edition: 662.000 EI-MS-Spektren
- NIST/EPA/NIH 2011 Mass Spectral Library: 243.893 EI-MS-Spektren
- SpecInfo (Wiley-VCH): 333.500 $^{13}$C-NMR-, 91.450 $^{1}$H-NMR-, 86.000 MS- und 30.000 IR-Spektren
- Sadtler IR Libraries (Bio-Rad): 224.000 IR-Spektren

# BMBF-Projekte




- 1987 – 1989: Spektrendatenbanken-Verbundsystem (SDVS)
- Verantwortlich: FIZ Karlsruhe und FIZ Chemie, Berlin
- Ziel: Verknüpfung von chemischen Strukturen mit NMR-, MS-, IR- und UV/VIS-Spektren sowie mit physikalischen Daten
- Beteiligt (u.a.): MPI für Kohlenforschung, Mülheim und ISAS Dortmund als Spektren-Bewertungsstellen
- 1990 – 1995: Informationssystem-Spektroskopie
- Gründung der Fa. Chemical Concepts GmbH, Weinheim als Tochterfirma des VCH

Chemical Concepts

- Ziel: Wirtschaftlich profitables Unternehmen nach Projektende
- Seit 1996 gehört Chemical Concepts zusammen mit Wiley-VCH zu John Wiley & Sons Inc.

# Probleme der Spektren-Bewertungsstellen

- Geeignete Kriterien zur Bewertung der Spektren fehlten
- Die Empfehlungen der Ausschüsse des SDVS-Projekts waren sehr allgemein gehalten
- Die Daten lagen in den Binärformaten der Spektrometer-Hersteller vor
- Der Aufbau dieser Binärfiles war in der Regel nicht bekannt
- Ein Format zur Speicherung der Metadaten (Spektrometer, Messparameter, Probe, Probenpräparation, Datenverarbeitung) fehlte
- Das MDL Molfile-Format zur Speicherung chemischer Strukturen war zunächst geschützt

# Das JCAMP-DX-Format

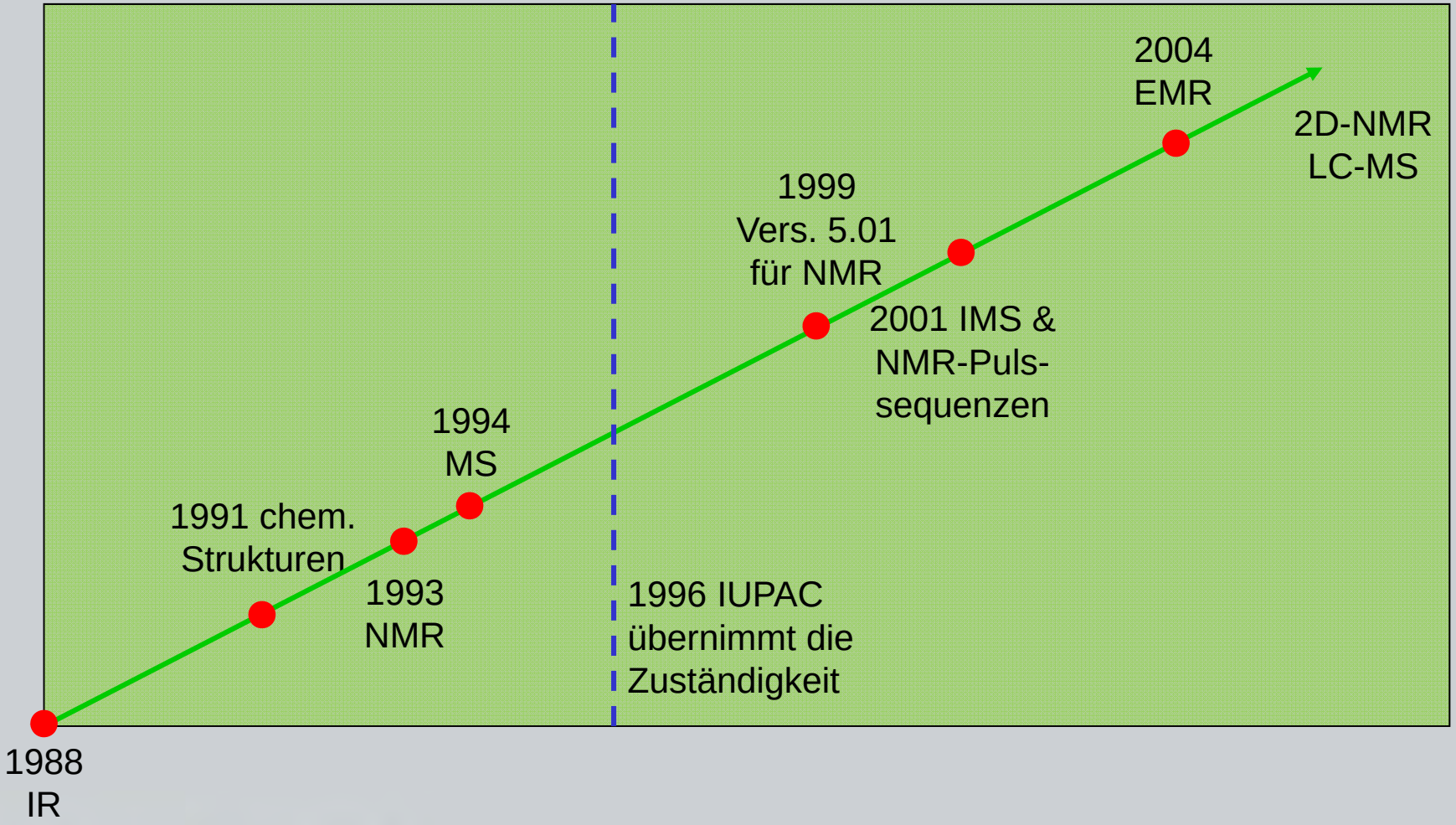
- R.S. McDonald and P.A. Wilks, Jr.: *JCAMP-DX: A standard Form for Exchange of Infrared Spectra in Computer Readable Form*, Appl. Spectrosc., **42**, 151-162 (1988)
- Entwickelt im Auftrag des Joint Committee on Atomic and Molecular Physical Data (JCAMP)
- ASCII-Format, in dem sowohl IR-Spekten als auch Metadaten gespeichert werden können
- Das Format beinhaltet eine Datenkompressions-Methode (DIFDUP) zur Speicherung der Spektren

# JCAMP-DX für IR – Beispiel

```
##TITLE= Knappe Beschreibung des Spektrums
##JCAMP-DX= 4.24
##DATA TYPE= INFRARED SPECTRUM
##ORIGIN= Name des verantwortlichen Wissenschaftlers
##OWNER= Name des Eigentümers der Daten
…
(optionale Records mit Metadaten)
##XUNITS= 1/CM
##YUNITS= ABSORBANCE
##XFACTOR= 4.88281E-4
##YFACTOR= 2.44140E-4
##FIRSTX= 4000.66
##LASTX= 400.16
##NPOINTS= 3735
##XYDATA= (X++(Y..Y))
8193341A66jkKmM%JjlJkLjnJMjTKn%K%UlLKjT%Km%KlM%kLkJjkjK%lKTlJLj%mjM%LpjNnK%Uk
…
831380H89JJ9K4K6L6j6
##END=
```

1988:
Datenübertragung typischerweise mit Kermit

# JCAMP-DX Entwicklung



# Das Analytical Data Interchange Protocol (ANDI)

- Entwickelt von D. Stranz et al. im Auftrag der Analytical Instrument Association (AIA), Alexandria VA, USA
- Die Protokolle werden heute von ASTM Subcommittee E13.15 (Analytical Data) betreut (ASTM-Standards E2077 und E1947)
- Es existieren Protokolle für Massen-Spektren und für Chromatogramme
- Sie verwenden das maschinen-unabhängige NetCDF-Format, entwickelt vom Unidata Program Center, Boulder CO, USA
- Für NetCDF existiert eine Programmierschnittstelle (API) in C mit Schnittstellen zu anderen Programmiersprachen (z.B. zu Matlab)
- Im NetCDF-Format können sowohl mehrdimensionale Datensätze als auch Metadaten gespeichert werden
- ANDI-MS wird von zahlreichen Geräteherstellern unterstützt

# Eurospec

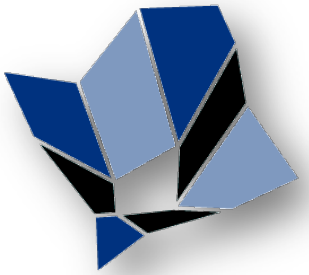
- Projekt im 5. EU-Rahmenprogramm (FP5), Teilprogramm Growth, 2002 – 2004
- Titel: Access to Research Spectroscopic Data and Associated Chemical Knowledge
- Ziel: Spektren aus der wissenschaftlichen Literatur online verfügbar zu machen.
- 7 Projektpartner aus 5 Ländern (CZ, D, F, GB, P), Koordinator: ISAS Dortmund, Projektleiter: A.N. Davies
- Vereinbarungen mit den Herausgebern und Verlagen wissenschaftlicher Zeitschriften waren angestrebt
- Die von den Autoren in der Datenbank gespeicherten Spektren sollen bei der Begutachtung der Papers genutzt werden können

# International Spectroscopic Data Base

- Ein Dateneingabe-System für Spektren, chemische Strukturen und bibliographische Informationen wurde entwickelt
- Als Front-End für die Datenbank-Suche wurde ein kommerzielles Webportal eingesetzt (Q-DIS/Panda)
- Das Systeme ist im Internet unter www.is-db.org erreichbar
- Nach dem Ende des EU-Projektes ist die Verantwortung vom IS-DB e.V., Dortmund übernommen worden
- Die Liste der interessierten Nutzer ist lang, jedoch die Anzahl der bereit gestellten Datensätze ist gering

# Das AnIML-Format

- ASTM E13.15 plant seit 2004 die Einführung der Analytical Information Markup Language (AnIML)
- Das Format verwendet XML Schema Definitionen (XSD)
- Es soll für jede Art analytisch-chemischer Daten geeignet sein
- Das Speichern mehrdimensionaler Datensätze ist möglich
- Metadaten können integriert werden
- Binärdaten werden im Base64-Format gespeichert
- Projekt-Webseite: animl.sourceforge.net
- Eine Kooperation mit dem IUPAC Subcommittee on Spectroscopic Data Standards ist vereinbart



# DFG-Projekt RADIESCHEN

Workshop Materialien
des Experten-Workshops
vom 17.April 2012

## Session 2 – Einbindung in den Forschungsprozess:
Kommen die Daten zur Infrastruktur
oder die Infrastruktur zu den Daten?

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

# Forschungs-datenmanagement – Datenakquise im Labor

**Matthias Razum**

**Experten-Workshop des DFG-Projekts Radieschen**

**Potsdam, 17.04.2012**

CC BY NC

FIZ Karlsruhe

# BW-eLabs

BW-eLabs gives access to virtual and remote experiments in the field of nano technology and digital holography.

Key concepts include:

- reproducability of experiments
- discoverability of and access to primary data
- storage and curation of all artifacts that emerge throughout the research process

The project was funded for 2 ½ years by the Ministry of Science, Research, and Art, Baden-Württemberg

Ministerium für Wissenschaft, Forschung und Kunst
Baden-Württemberg

# BW-eLabs

Project partners were

- University of Stuttgart
  - Institute of Applied Optics
  - Computing Center
  - Library
- Stuttgart Media University
- University of Freiburg
  - Freiburg Materials Research Center
  - IT Services
- FIZ Karlsruhe

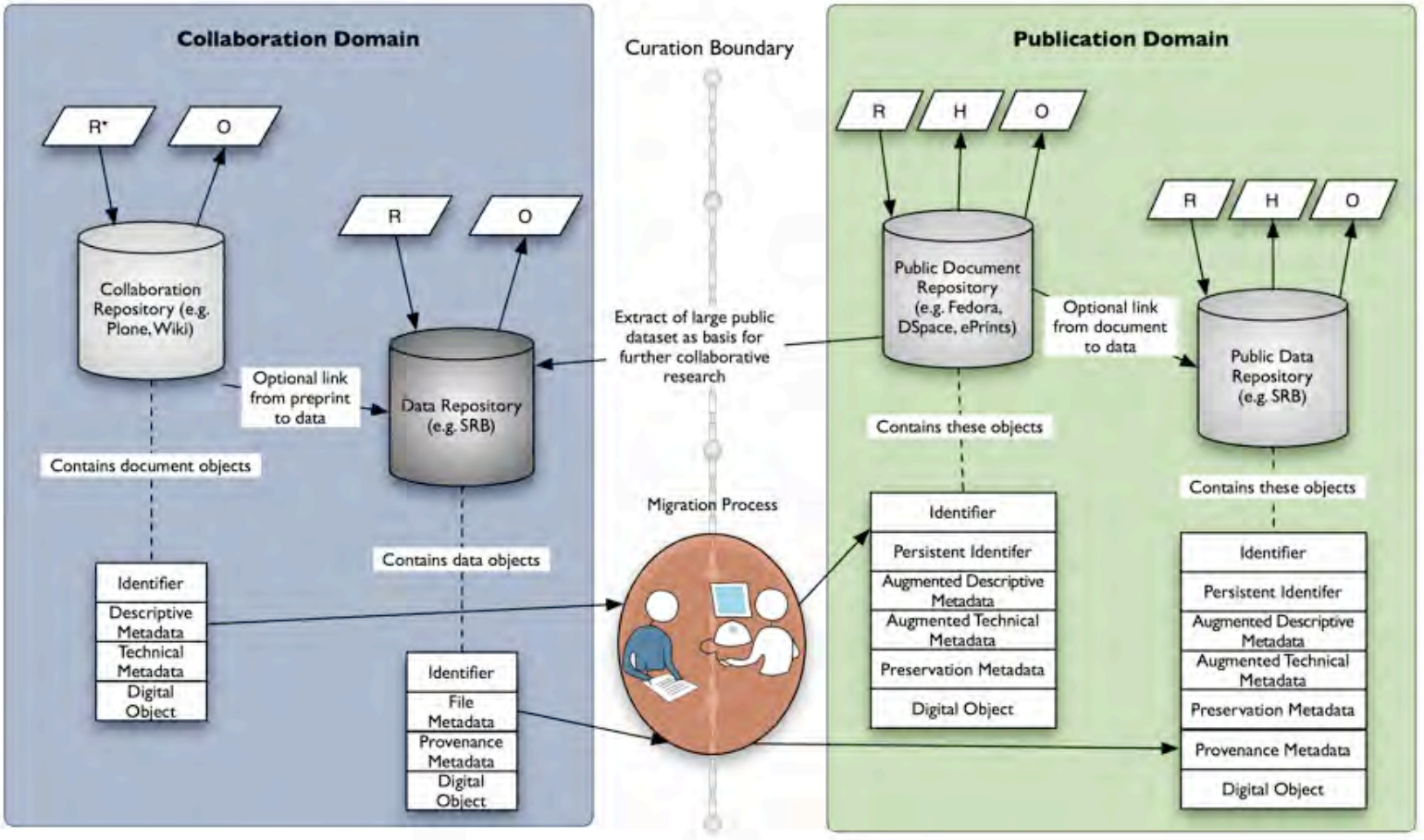
# Research Data Management Requires more than Storing Bits and Bytes

A researcher in digital arts says he needs “something to handle **all the media and metadata**.”

“**Data access and formatting** are not standardised, even within one lab.”

*A Slice of Research Life: Information Support for Research in the US*

# Collaboration, Publication, and Curation Boundary




Treloar, A., Groenewegen, D. und Harboe-Ree, C., 2007. The Data Curation Continuum. *D-Lib Magazine,* 13 (9/10)



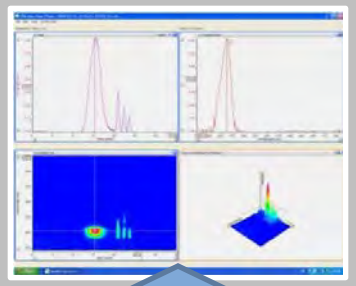
FIZ Karlsruhe

# BW-eLabs: Collecting Data throughout the Research Process

Preparation

Synthesis and Online Analysis

Offline Analysis and Visualisization



Publication

Common Data Infrastructure

eSciDoc PROJECT

Experiment

Experiment

Data Objects

Experiment

Data Objects

Experiment

Data Objects

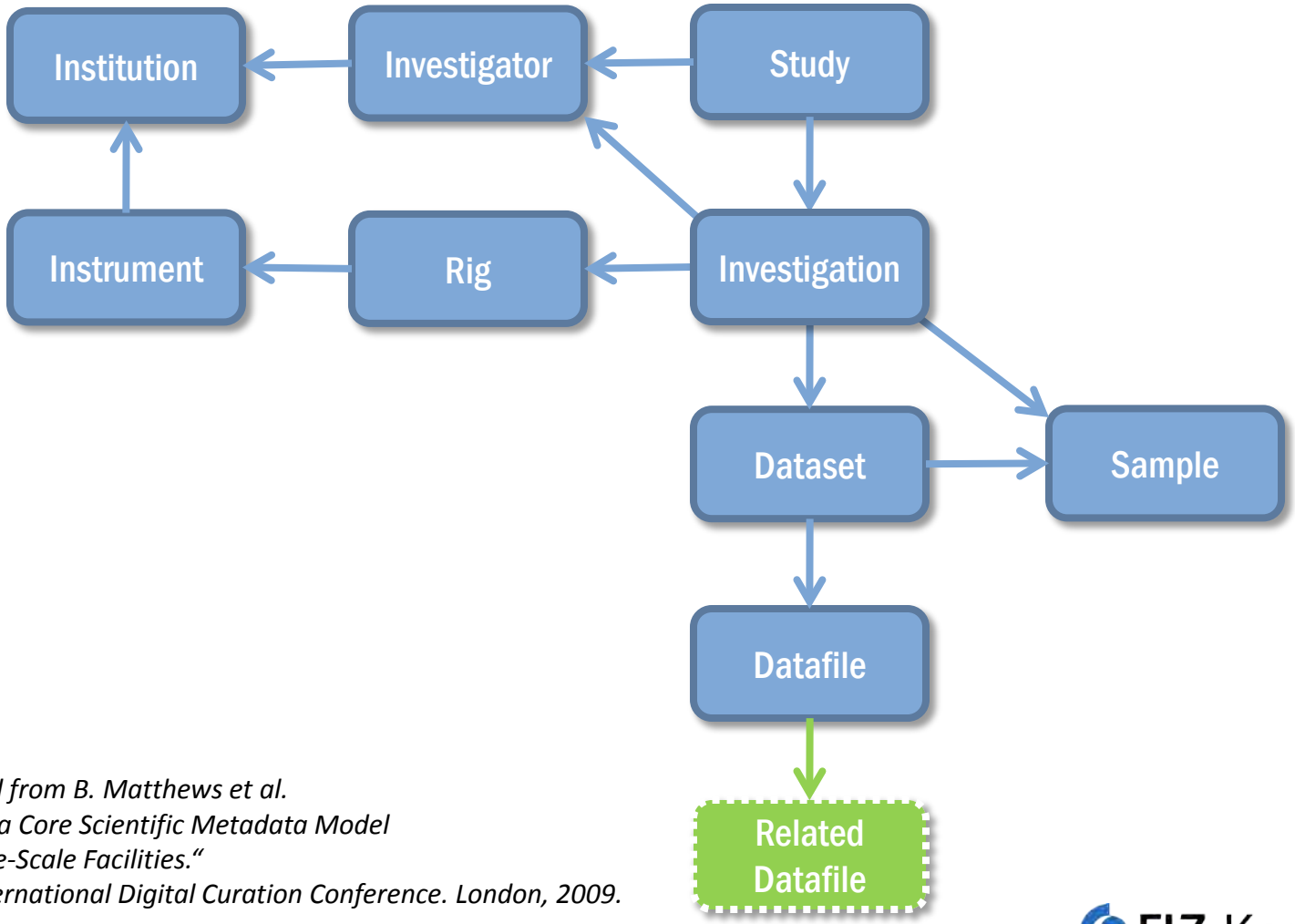


FIZ Karlsruhe

# More About the Research Question

FIZ Karlsruhe

# BW-eLabs: Metadata Model for Contextual Information

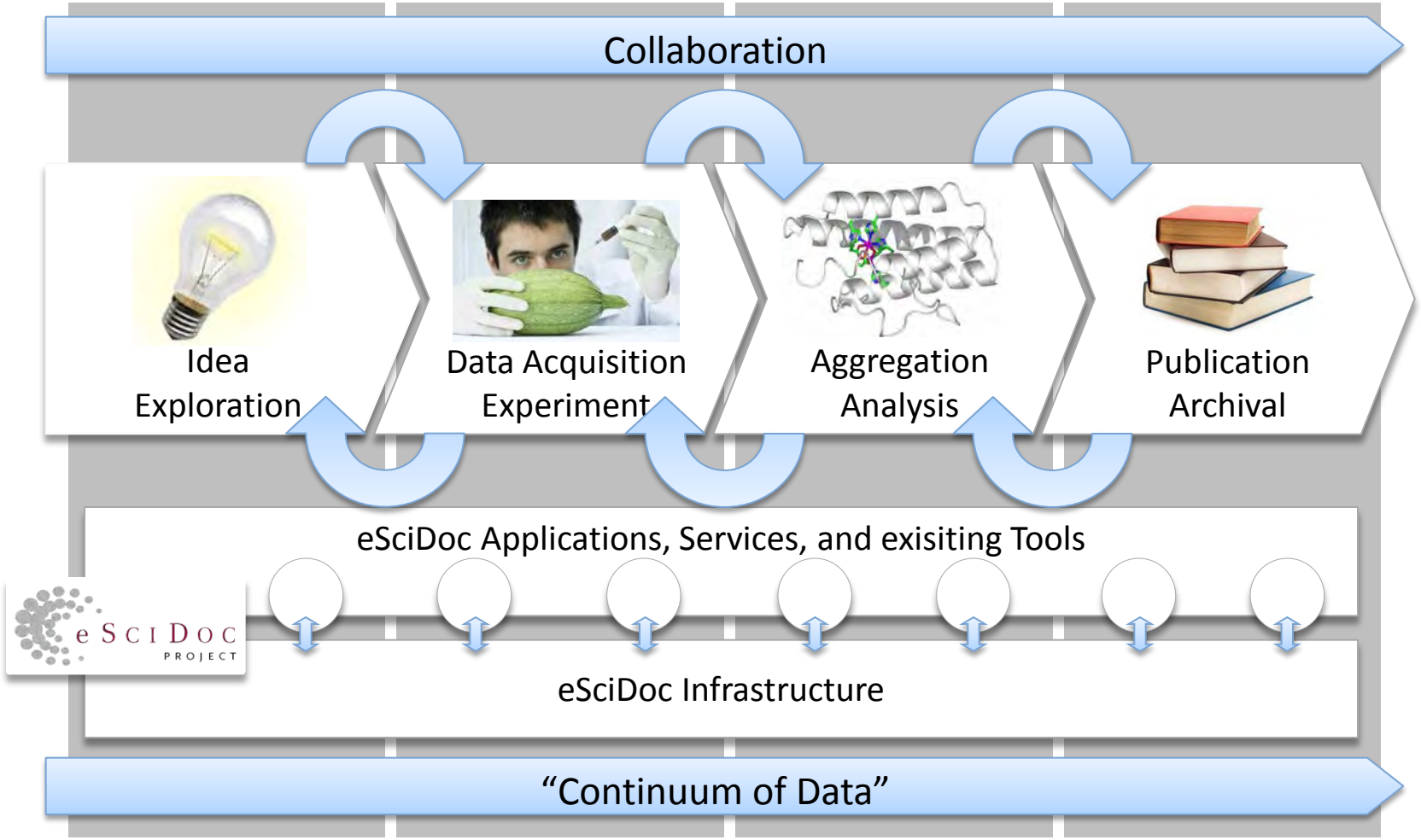



*Derived from B. Matthews et al.*
*„Using a Core Scientific Metadata Model*
*In Large-Scale Facilities."*
*5th International Digital Curation Conference. London, 2009.*



FIZ Karlsruhe

# eSciDoc Vision






FIZ Karlsruhe

# eSciDoc Services

Data Acquisition Service
Control of Named Entities
Deposit (SWORD)
Citation Style Service
Transformation
Validation
Search and Indexing
Technical Metadata Extraction
Duplicate Detection
PID Service
Digilib
OAI-PMH
Admin Tool
Deposit Service

Policy Enforcement Point

Security
- Group Handler
- Role Handler
- User Account Handler

Admin Service

Object Manager
- Organizational Unit Service
- Context Service
- Container Service
- Item Service
- Content Model Service

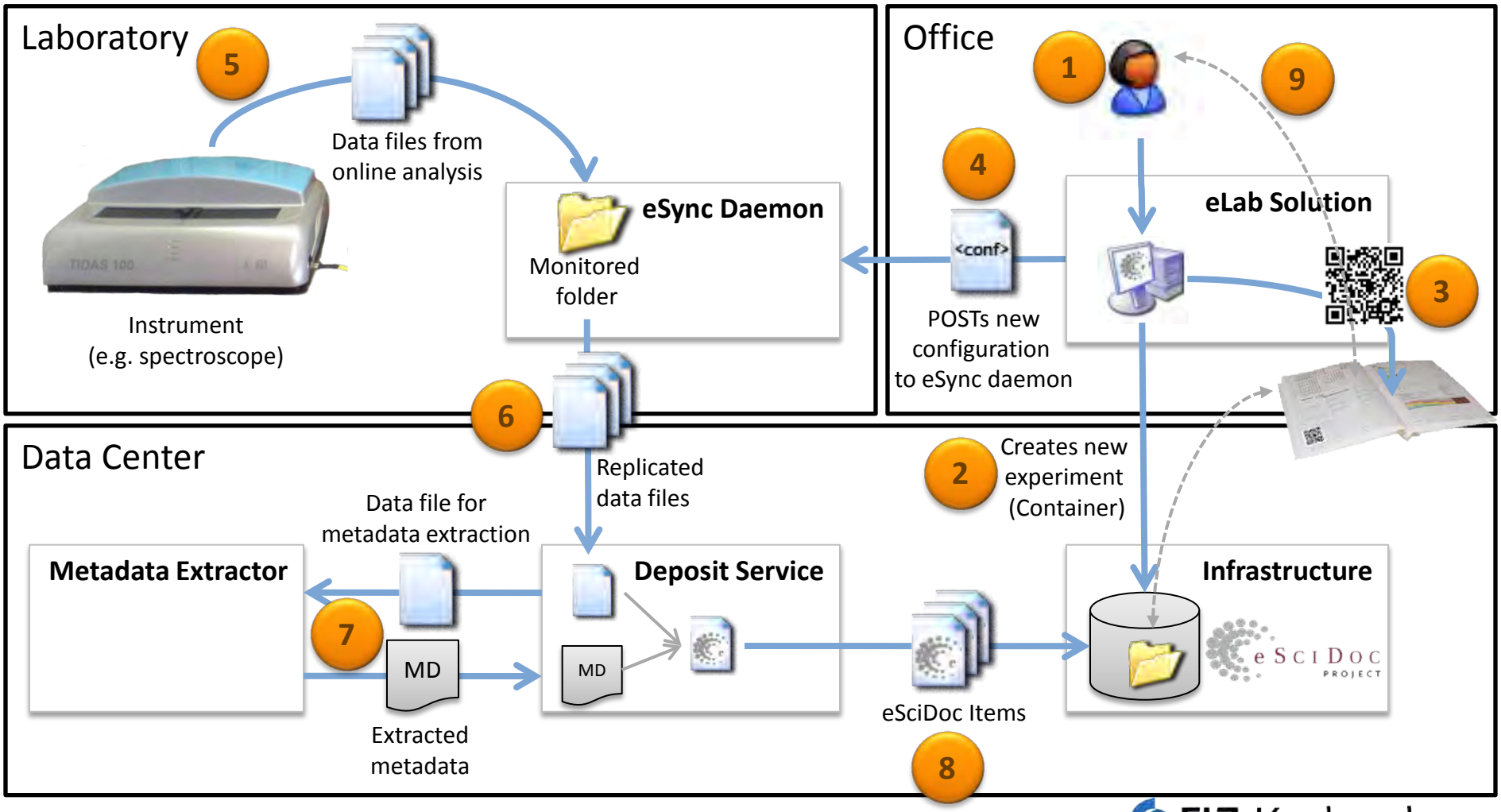
Set Service

Statistics Manager
- Scope Service
- Statistics Data Service
- Aggregation Definition Service
- Report Definition Service
- Report Service

FedoraCommons™

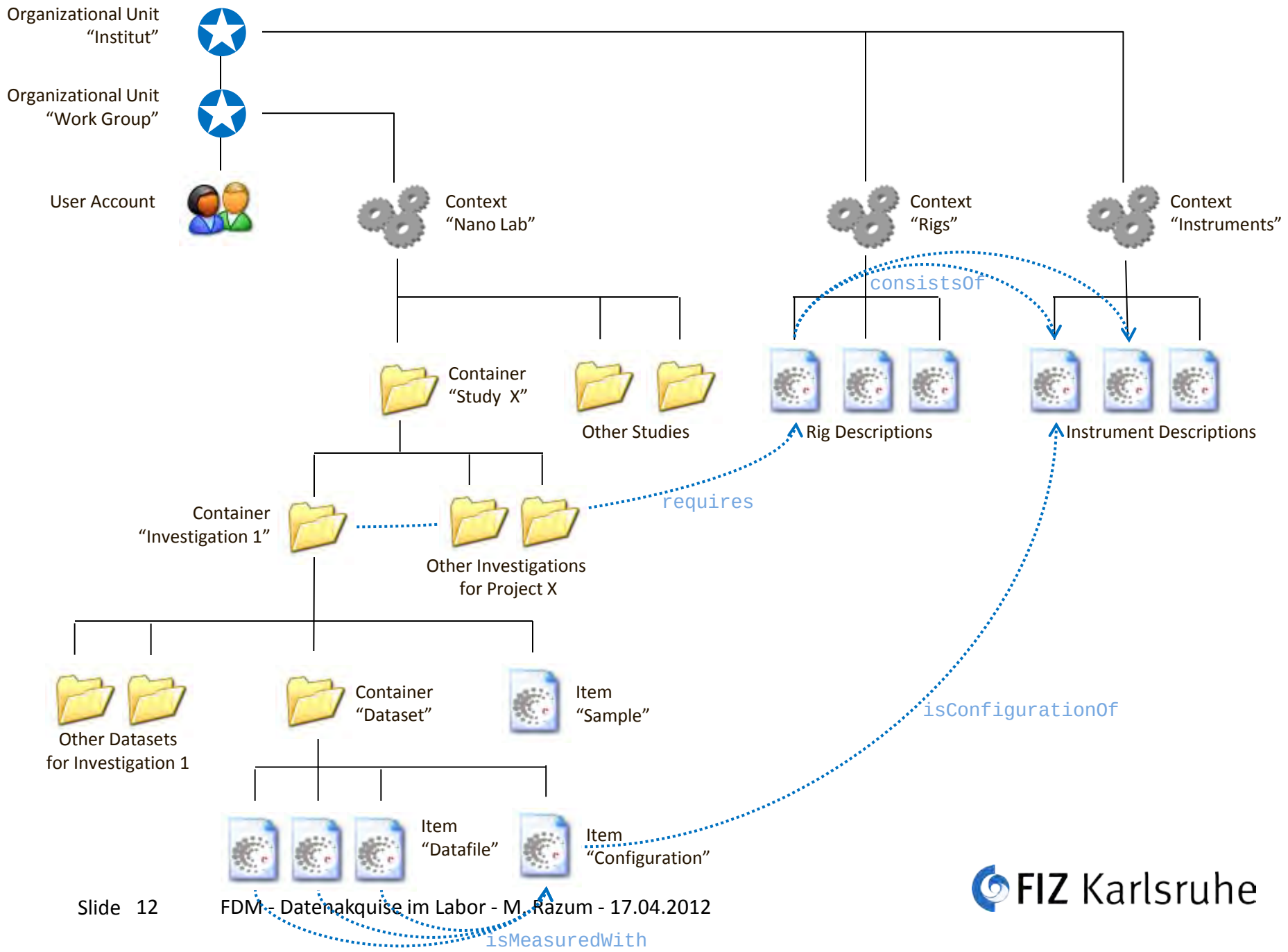
FIZ Karlsruhe

# BW-eLabs: Data Acquisition in the Lab






FIZ Karlsruhe






FIZ Karlsruhe

# Fragen?

matthias.razum@fiz-karlsruhe.de

www.escidoc.org
www.bw-elabs.org

CC BY NC

FIZ Karlsruhe

# Zentral oder dezentral?
# Internationale Infrastrukturen für Primärdaten aus der Biodiversitätsforschung

Anton Güntsch

Freie Universität Berlin

Botanischer Garten und Botanisches Museum Berlin-Dahlem

BiodiversityInformatics@bgbm.org

# Primary biodiversity data

- Specimens
  - „Traditional" preserved material (dried and pressed plants, feathers, bones, fossils, wood, …).
  - But also DNA, seed-banks, culture collections, etc.
- Observations, DNA sequence data, multimedia objects, etc.
- Distributed storage: Museums, Universities, private holdings.
- Broad coverage: organismic, temporal, spatial.

Biodiversity data document the history of life on earth in space and time.

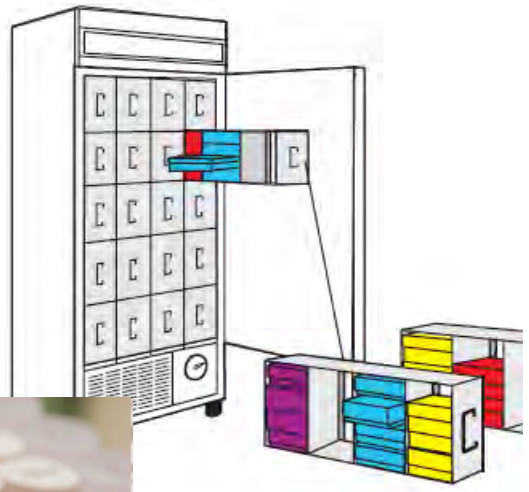
# Example: insect collection

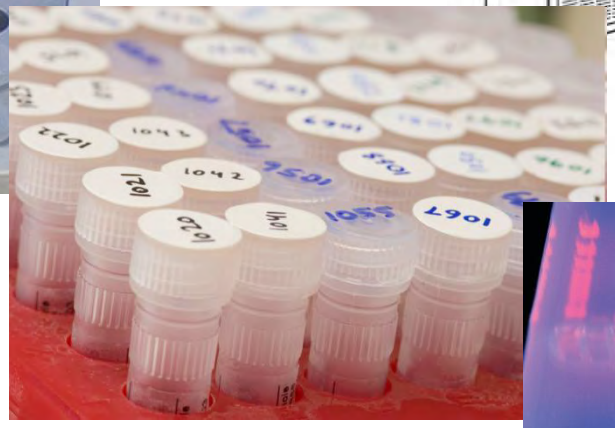
© J. Holstein et al.

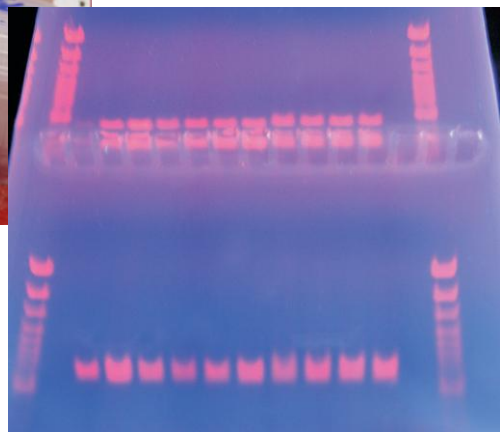
# Example: Herbarium

© J. Holstein et al.

# Example: DNA Collection







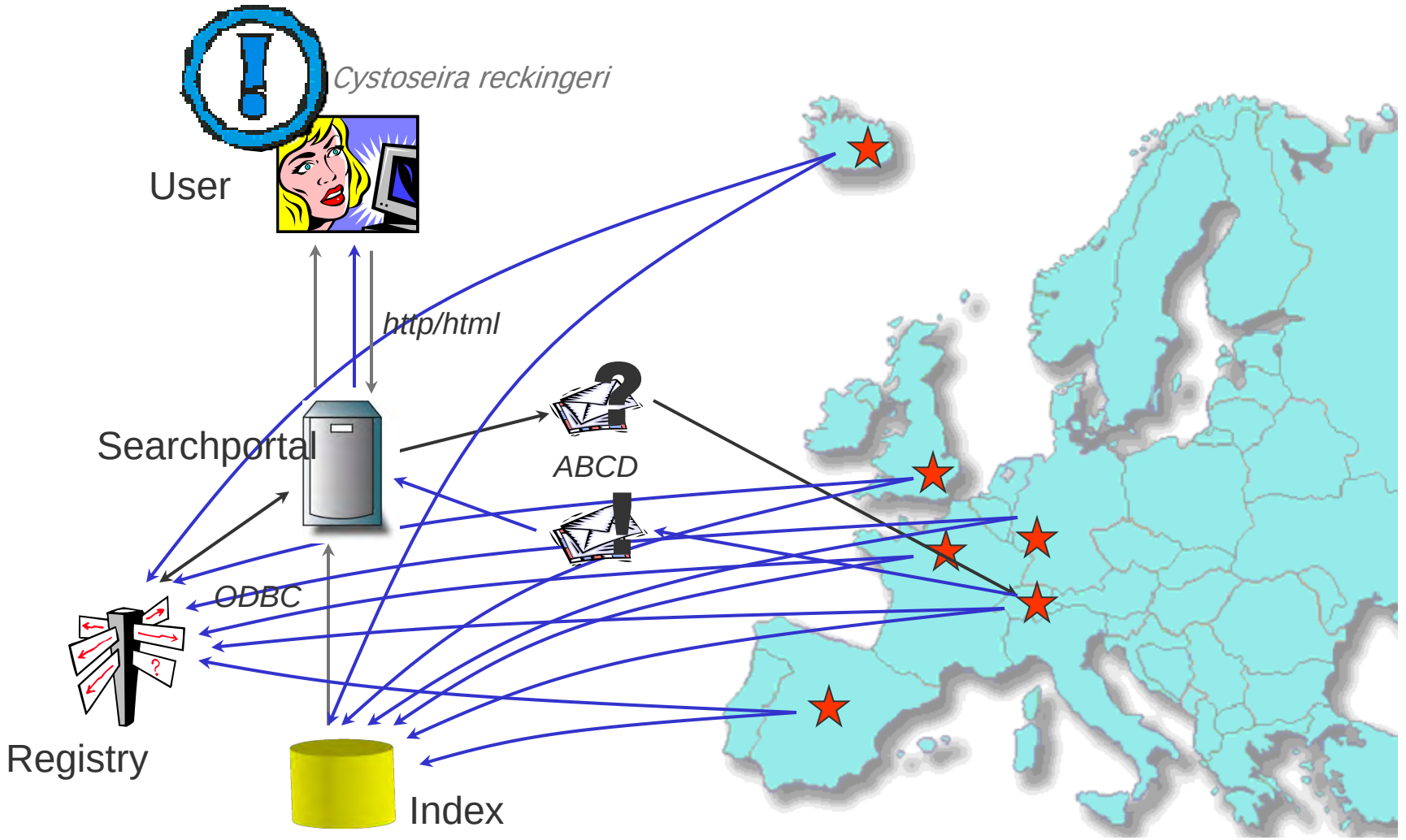
# A physical database of the natural world

- 2-3 Billion physical collection objects in natural history collections.
- A valuable resource for science and education.
- Example applications:
  - species inventories,
  - studies on biodiversity changes,
  - distribution modelling, effect of climate change on biodiversity and prediction of invasion of species.

# Networking primary biodiversity data

- Joint effort of several international initiatives (e.g. GBIF and BioCASE).
- Based on data standards (DarwinCore, ABCD, ...).
- Agreed access protocols (DiGIR, TAPIR, BioCASE).
- Today: simultaneous access to 321M records provided by 379 data publishers world wide.

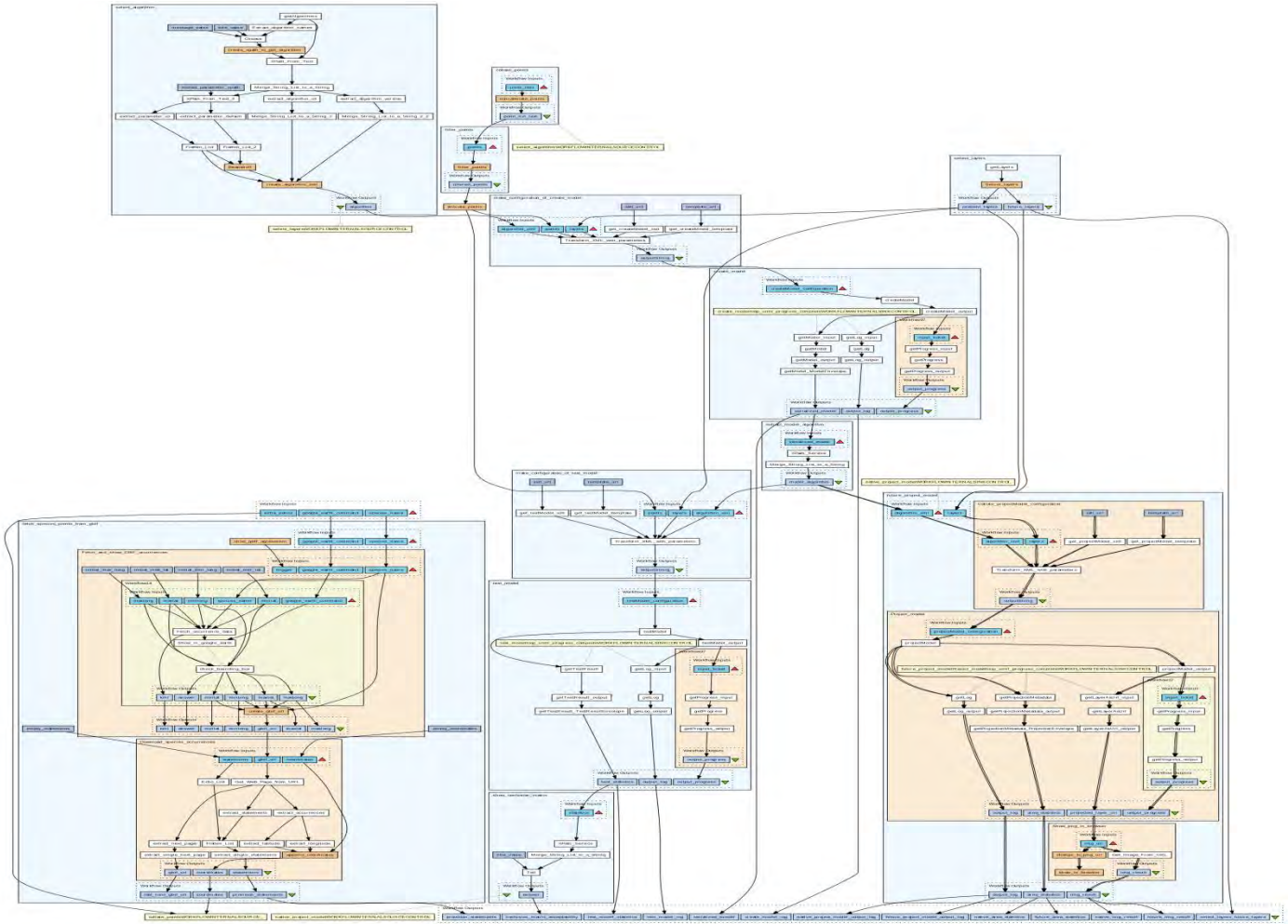
# Simplified information flow



# History

- **2000 – : mobilisation and basic networking**
  - Standards
  - Software components
  - Capacity building
- **2005 – : services, services, services**
  - Specimens & observations
  - Taxon-level data
  - Literature
  - Tools
  - …
- **2010 – : integration**
  - Service registries
  - Workflows

# Example: niche modelling workflow



# Warum dezentral?

- Datenhalter können leichter überzeugt werden, ihre Daten zur Verfügung zu stellen, wenn sie jederzeit die **Kontrolle** über die Daten behalten.
- Datenhalter, die Teil der technischen Infrastruktur sind, begleiten in der Regel auch den **Standardisierungsprozess** aktiver.
- Vermeidung von **Synchronisation**sproblemen.

# Warum doch lieber zentral?

- Ein substantieller Teil der Datenanbieter ist nicht in der Lage, **stabile Services** anzubieten.
- Die **Integrität** der Daten lässt sich besser in einem zentralen System sicherstellen oder verbessern.
- In Zukunft wird es wahrscheinlich nicht mehr möglich sein, die für bestimmte **Analysen** benötigten Daten über das Netz zu transportieren. Die Analysen müssen dann also zu den Daten kommen.

# Oder doch eher gemischt?

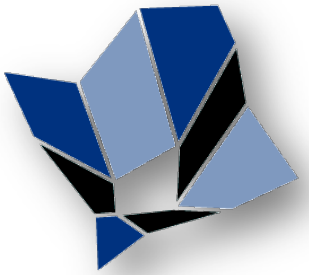
- **Weiterhin dezentrale** Infrastruktur mit zunehmender Spezialisierung und **Schwerpunktbildung**.
- "**Hosting Institutionen**" helfen Datenanbietern, die selbst nicht über eine adäquate technische Infrastruktur verfügen.
- "**Aggregatoren**" bieten performanten Zugriff auf Massendaten und zentralisierte Datenanlysen.

**Anton Güntsch**
Freie Universität Berlin
Botanic Garden & Bot. Museum Berlin-Dahlem
Dept. Biodiversity Informatics & Laboratories
Königin-Luise-Straße 6-8
D-14195 Berlin-Dahlem

a.guentsch@bgbm.org

Biodiversity informatics @bgbm.org

www.bgbm.org/biodivinf



# DFG-Projekt RADIESCHEN

Workshop Materialien
des Experten-Workshops
vom 17.April 2012

## Session 3 – Generische vs. Disziplinspezifische Dienste:
Was sind die Erfolgskriterien
disziplinübergreifender Dienste?

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

da|ra Services
da-ra.de
da|ra Services
da|ra
DataCite Services
The DataCite Metadata Store (MDS)
DataCite Metadata Search
DataCite OAI Provider
Best practice
Struktur
und Verantwortlichkeiten
Upgrade
da|ra: DOI-Registrierung für
Sozial- und Wirtschaftsdaten
da|ra Services
gesis.org/dora
Mitgliedschaft
da|ra
Pilotprojekt
1. DataCite
2. Das da|ra Projekt
3. Warum da|ra?
DataCite
1. Dezentrale Datenlage
Warum da|ra?
5. Policy Framework zur
Unterstützung der Nutzer
2. Spezifische Anforderungen
an die Metadaten
4.Optimierung der Findbarkeit
3. Entwicklung von
Zusatzdiensten
auf der Basis umfangreicher
Metadaten
da|ra Metadatenschema v2.2.1

# da|ra: DOI-Registrierung für Sozial- und Wirtschaftsdaten

Brigitte Hausstein
GESIS Leibniz Institut für Sozialwissenschaften, Berlin

Radieschen Workshop GFZ Potsdam, 17 April 2012

ZBW

da|ra Registration agency for social and economic data

gesis

Prezi

1. DataCite

2. Das da|ra Projekt

3. Warum da|ra?

Prezi

# 1. DataCite

Prezi

# DataCite

- Internationales Konsortium, gegründet in London 1. Dezember 2009
- Mission: Einfacher Zugang zu Forschungsdaten und Erhöhung deren Akzeptanz
- Unterstützung für die Archivierung der Daten (Verifikation und Re-Use)
- Persistente Identifikation der Daten durch die Nutzung des DOI-Systems

Prezi

# Mitgliedschaft

## Gegenwärtig: 16 Mitglieder aus 11 Ländern

**Members**

| | |
|---|---|
| AUS | Australian National Data Service (ANDS) |
| CAN | Canada Institute for Scientific and Technical Information |
| CH | Library of the ETH Zurich |
| DK | Technical Information Center of Denmark |
| FR | Institute for Scientific and Technical Information |
| GER | German National Library of Science and Technology (TIB) |
| | German National Library of Medicine (ZB MED) |
| | GESIS - Leibniz Institute for Social Science |
| NL | TU Delft Library |
| UK | The British Library |
| USA | California Digital Library (CDL) |
| | Purdue University Libraries |

**New members 2011**

| | |
|---|---|
| GER | German National Library of Economics (ZBW) |
| S | The Swedish National Data Service (SNDS) |
| USA | Office of Scientific and Technical Information (OSTI), US Department of Energy |

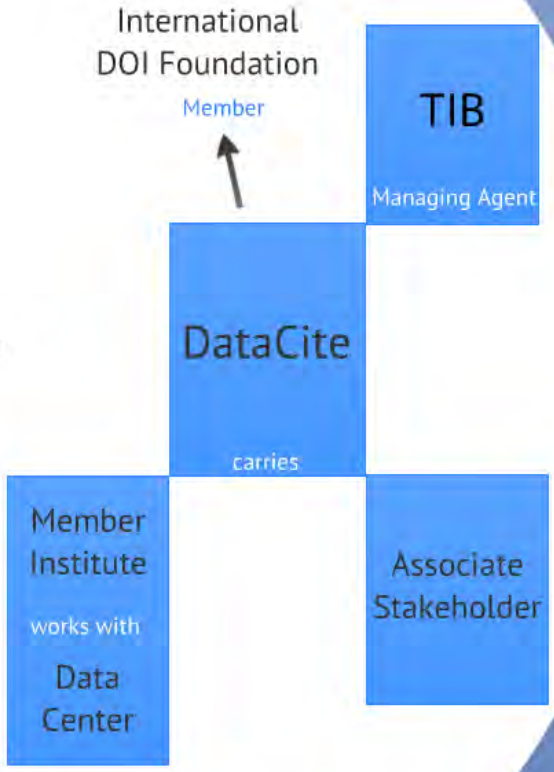
**Associated Members**

| | |
|---|---|
| UK | Digital Curation Centre |
| USA | Microsoft Research |

mehr als 1.000.000 registrierte DOI-Namen

Prezi

# Struktur und Verantwortlichkeiten




DataCite (DOI-Registrierungsagentur):

- Resolution Infrastruktur und Datenbank
- Langzeit-Management der DOI-Namen
- Best practice

Allocation agencies (DC Mitgliedsinstitute):

- Erzeuegen der DOI-Namen
- Qualitätssicherung
- Suchbare Datenbank der Metadaten
- Best practice

Publikationsagenten (Datenzentren, Daten-Publishers):

- Datenspeicherung und Zugang
- Persistenz der Objekte
- Erstellen und Updating der Metadaten

Prezi

# DataCite Services

(April 2012)

- The DataCite Metadata Store (MDS)
- DataCite Metadata Search
- DataCite OAI Provider
- Best practice

Prezi

1. DataCite

2. Das da|ra Projekt

3. Warum da|ra?

Prezi

1. DataCite

2. Das da|ra Projekt

3. Warum da|ra?

Prezi

da|ra

Registration agency for social and economic data

Pilotprojekt

POLICIES v1.0

Registry service und Informationssystem

da|ra Metadatenschema v1.0

Prezi

# da|ra Services
# gesis.org/dara

Prezi

# Upgrade

- Kooperation mit der ZBW
- DOI service auch für Wirtschaftsdaten
- DFG Förderung (2011-2014)

04|11 - 12|11

POLICIES v2.0

Ugrading, add-on services

da|ra Metadatenschema v2.2.1

SLA Template

Prezi

# da|ra Services
# da-ra.de

Prezi

# da|ra Services

04|2012: 5500 registrierte DOI Namen

Prezi

1. DataCite

2. Das da|ra Projekt

3. Warum da|ra?

Prezi

# 3. Warum da|ra?

Prezi

# Warum da|ra?

1. Datenlage in den Sozial- und Wirtschaftswisschenschaften

2. Community spezifische Anforderungen an die Metadaten

3. Entwicklung add-on Services

4. Optimierung der Findbarkeit

5. Policy Framework zur Unterstützung der Nutzer

Prezi

# Warum da|ra?

1. Datenlage in den Sozial- und Wirtschaftswisschenschaften

2. Community spezifische

Prezi

# 1. Dezentrale Datenlage

# Warum da|ra?

1. Datenlage in den Sozial- und Wirtschaftswisschenschaften

2. Community spezifische Anforderungen an die Metadaten

3. Entwicklung add-on Services

4. Optimierung der Findbarkeit

5. Policy Framework zur Unterstützung der Nutzer

Prezi

# 2. Community spezifische Anforderungen an die Metadaten

# 3. Entwicklung add-on Ser

# 2. Spezifische Anforderungen an die Metadaten

- komplexe Datenstruktur

  Längsschnittdaten, Querschnittdaten, Mehrländerstudien, Kohorten, Aggregatdaten, Microdaten etc.
  ->Versionierungen, Erratenverwaltung etc.

- Rolle der Metadaten bei der Sekundäranalyse
- Studienbeschreibungen seit den 60er Jahren
- Entwicklung des DDI Standards

Prezi

# da|ra Metadatenschema v2.2.1

36 Elemente (8 + 4 obligatorisch, 24 optional)

(DC Metadata Schema: 17 Elemente, 5 mandatory)

spezifische Beschreibungselemente:

- Auswahl, Grundgesamtheit
- Geographischer Raum
- Ergebungsverfahren
- Publikationen
- Logo des PA
- ...

Prezi

# Warum da|ra?

1. Datenlage in den Sozial- und Wirtschaftswisschenschaften

2. Community spezifische Anforderungen an die Metadaten

3. Entwicklung add-on Services

4. Optimierung der Findbarkeit

5. Policy Framework zur Unterstützung der Nutzer

Prezi

# Anforderungen an die Metadaten

3. Entwicklung add-on Services

4. Optimierung der Findbarkeit

Prezi

# 3. Entwicklung von Zusatzdiensten auf der Basis umfangreicher Metadaten

- DOI und link checker
- Version linking
- Zitationsausgabe
- Verlinkung Daten und Literatur (sowiport)
- Content negotiation
- Registrierung weiterer Datentypen
- Nutzerstatistik
- Unterstützung peer review und impact factors

Prezi

# Warum da|ra?

1. Datenlage in den Sozial- und Wirtschaftswisschenschaften

2. Community spezifische Anforderungen an die Metadaten

3. Entwicklung add-on Services

4. Optimierung der Findbarkeit

5. Policy Framework zur Unterstützung der Nutzer

Prezi

3. Entwicklung add-on Services

4. Optimierung der Findbarkeit

5. Policy Framework zur Unterst
der Nutzer

Prezi

# 4.Optimierung der Findbarkeit

- Nutzung des fachspezifischen Knowhows in GESIS und ZBW
- Kategorisierung nach Wissenschaftsklassifikationen
- wissenschaftliches Navigieren
- Verbindung von Literatur und Daten, LOD

Prezi

# Warum da|ra?

1. Datenlage in den Sozial- und Wirtschaftswisschenschaften

2. Community spezifische Anforderungen an die Metadaten

3. Entwicklung add-on Services

4. Optimierung der Findbarkeit

5. Policy Framework zur Unterstützung der Nutzer

Prezi

# 4. Optimierung der Findbarkeit

# 5. Policy Framework zur Unterstützung der Nutzer

Prezi

# 5. Policy Framework zur Unterstützung der Nutzer

- Support bei der DOI Registrierung und Metadatenerstellung
- Berücksichtigung spezieller Nutzerbedürfnisse
- Verwaltung und Qualitätssicherung

Prezi

# da|ra: DOI-Registrierung für Sozial- und Wirtschaftsdaten

Brigitte Hausstein
GESIS Leibniz Institut für Sozialwissenschaften, Berlin

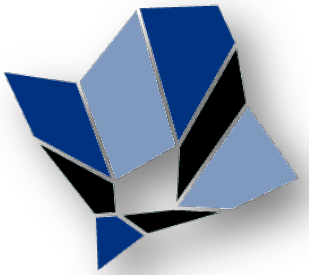
Radieschen Workshop GFZ Potsdam, 17 April 2012

ZBW

da|ra Registration agency for social and economic data

gesis

Prezi



# DFG-Projekt RADIESCHEN

Workshop Materialien

des Experten-Workshops

vom 17.April 2012

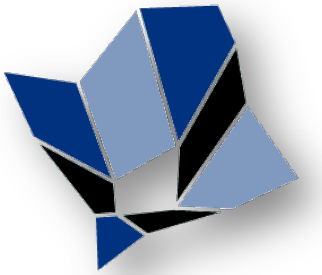
## Session 4 – Möglichkeiten und Grenzen der Auslagerung und Zentralisierung von Diensten

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

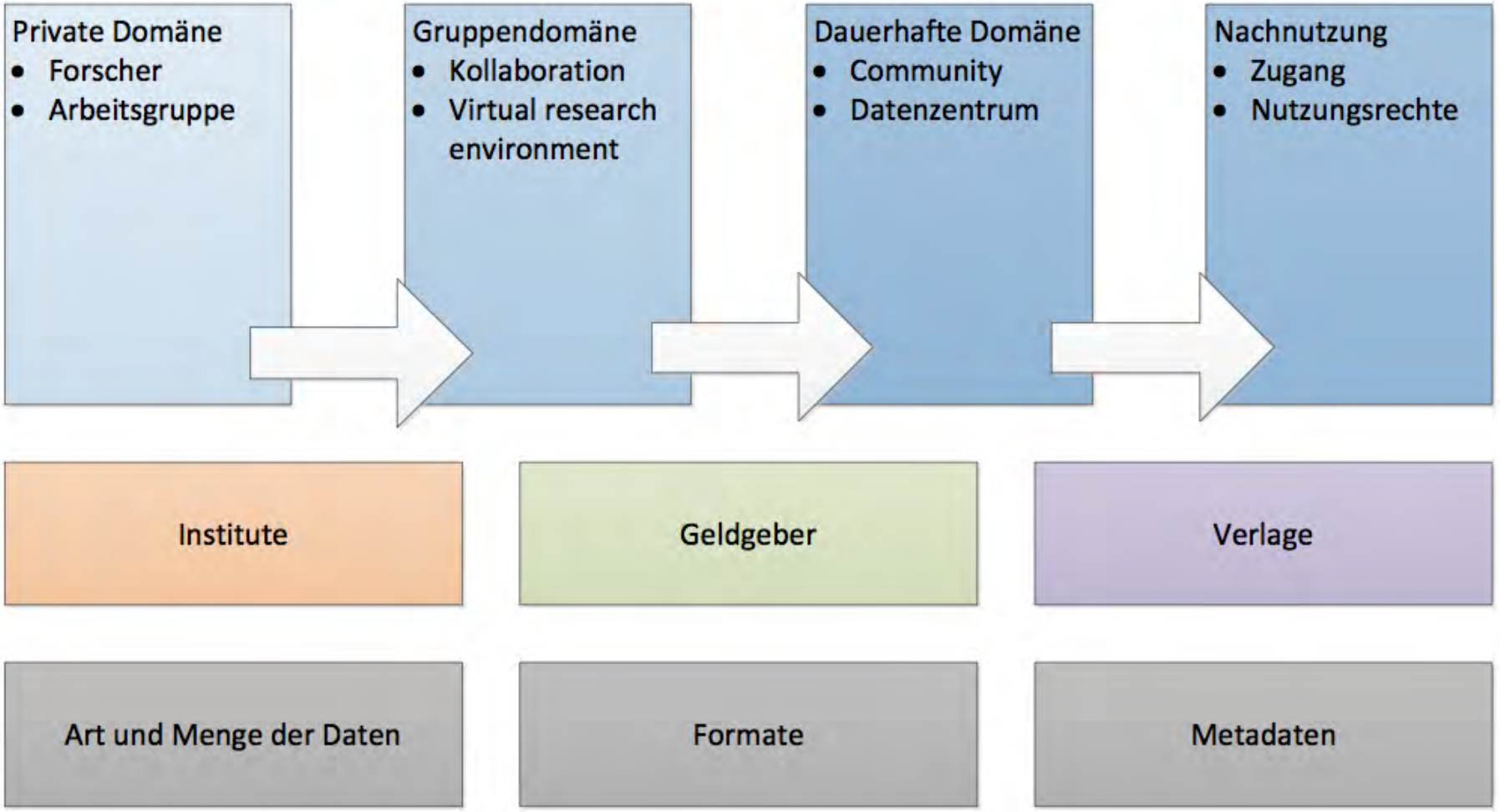
Radieschen Workshop

Breakout Session IV

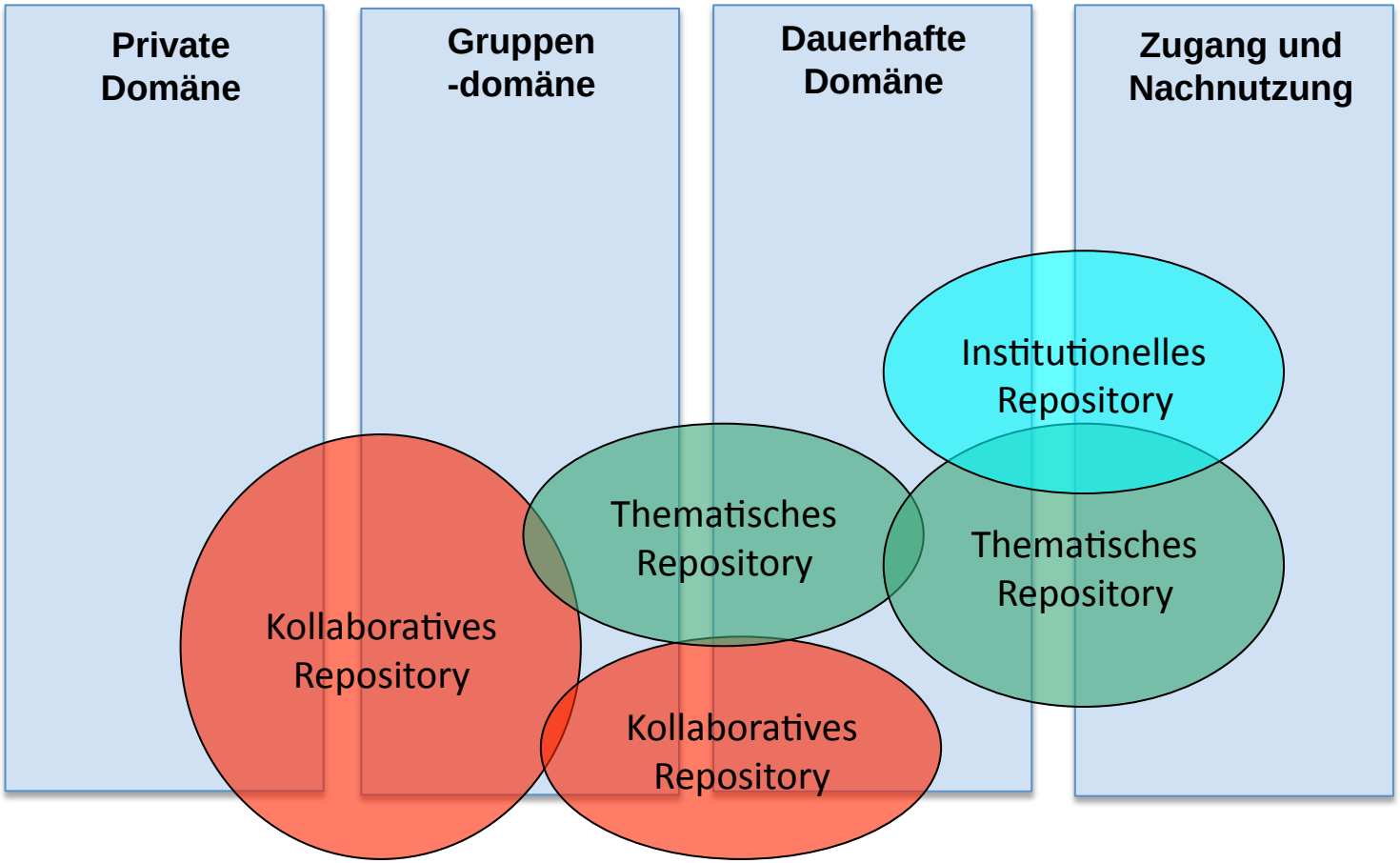
# Möglichkeiten und Grenzen der Auslagerung und Zentralisierung von Diensten



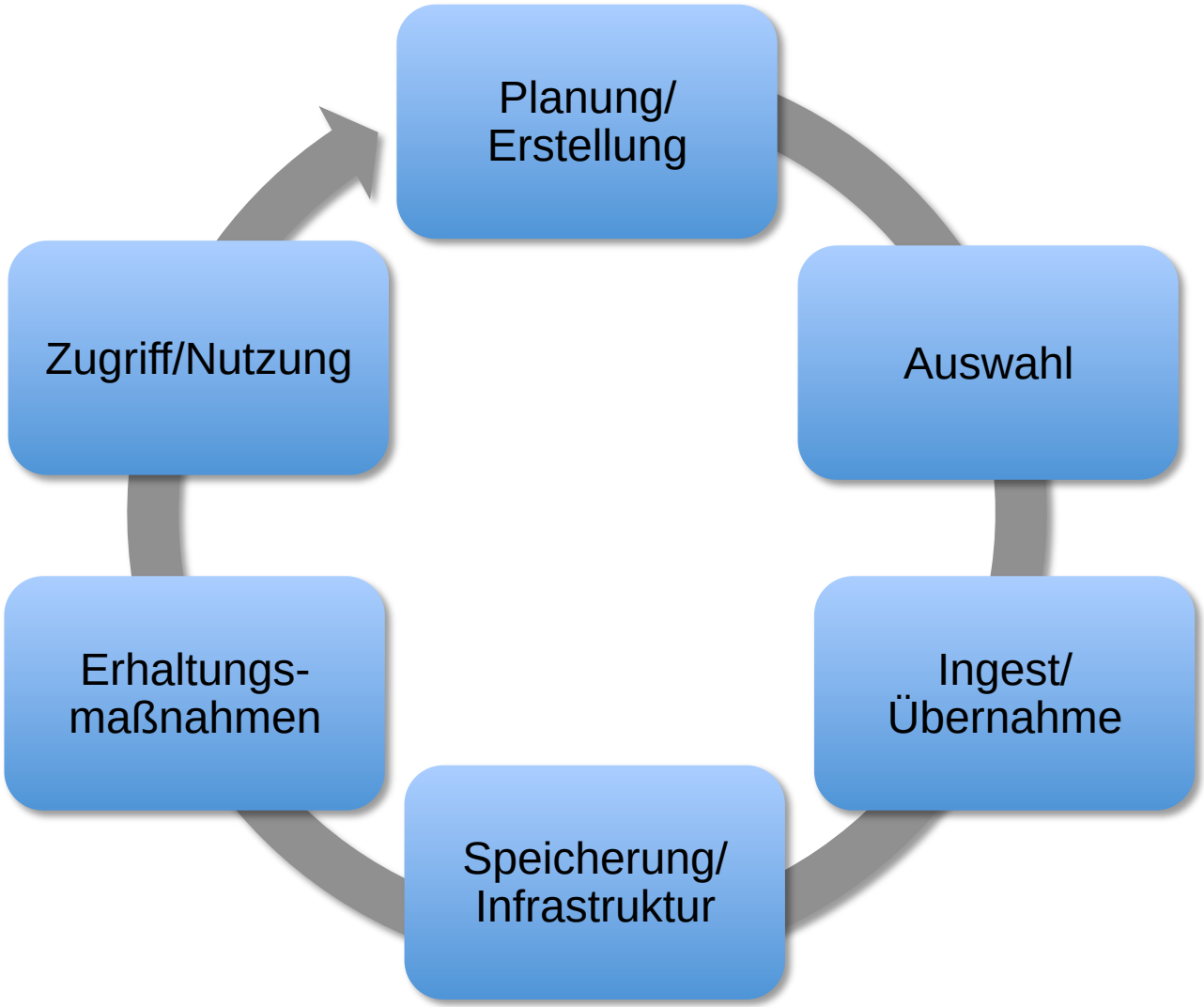
# Domain


Private Domäne
• Forscher
• Arbeitsgruppe
Gruppendomäne
• Kollaboration
• Virtual research environment
Dauerhafte Domäne
• Community
• Datenzentrum
Nachnutzung
• Zugang
• Nutzungsrechte
Institute
Geldgeber
Verlage
Art und Menge der Daten
Formate
Metadaten

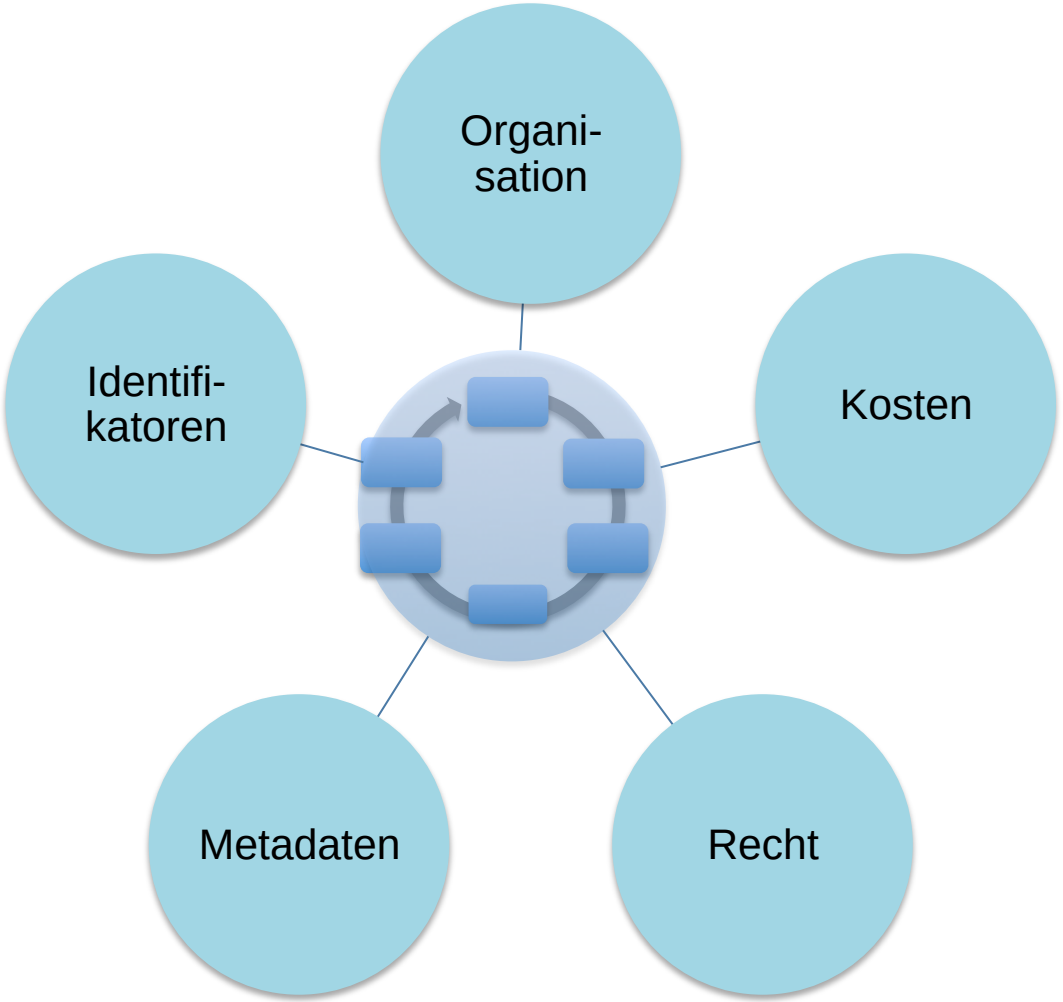
# Domain


Private Domäne
Gruppen -domäne
Dauerhafte Domäne
Zugang und Nachnutzung
Institutionelles Repository
Thematisches Repository
Thematisches Repository
Kollaboratives Repository
Kollaboratives Repository

# Aufgabenbereiche im Lebenszyklus



# Übergreifende Aufgabenbereiche



TECHNISCHE UNIVERSITÄT DRESDEN

Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH)

# Möglichkeiten und Grenzen der Auslagerung und Zentralisierung von Diensten

Zellescher Weg 12

Willers-Bau A207

Tel. +49 351 - 463 - 35450

Wolfgang E. Nagel (wolfgang.nagel@tu-dresden.de)

ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen

# Zukünftige Anforderungen in Universitätslandschaften

- **Integration**
  - **durchgängig IT-gestützte integrierte Prozesse**
  - **durchgängige organisatorische und technische Integration**
  - **multimediale und mobile Kommunikation, neue Lehr-, Lern- und Arbeitsformen**
- **Virtualisierung von Daten-, Informations- und Rechendiensten**
  - **global transparenter Zugriff auf Kapazitäten und Ressourcen**
  - **global kollaboratives Forschen, Entwickeln, Planen und Handeln**
- **Integrierte wissenschaftliche Arbeitsumgebung**
  - **neue Formen der Organisation, Kollaboration, Kooperation und Kommunikation mit einer umfassenden Wissensverarbeitung**
  - **transparente Nutzung eines breiten Spektrums von Diensten sowie Bereitstellung von Ressourcen und Know-how**
  - **erhöhte Anforderungen aus den Grundsätzen zur Sicherung guter wissenschaftlicher Praxis (Laborbuch, Sicherung der Experimentdaten etc.)**

# Zukünftige Anforderungen in Universitätslandschaften

- Service-orientierte Architekturen
- Forderung nach erhöhter Qualität
  - bedarfsgerechte Verfügbarkeit von Informationen und Ressourcen
  - Service Levels, Verlässlichkeit und Robustheit
  - Mobilität, Ubiquität („zu jeder Zeit, an jedem Ort“)
- Sicherheit
  - Transfer, Zugang und Management von Daten
  - Privacy, Protection, Integrität
  - Erweiterte rechtliche Rahmenbedingungen

# Forschungsdaten verwalten und sichern – Wie?

- Wissenschaftler können Forschungsdaten
  - selbst verwalten
    - Eigene Hardware, eigene Software, Beschaffungszyklen
    - Innerhalb der Community (meist fachspezifisch, Synergien)
  - an wissenschaftliche Rechenzentren geben
    - Lokale RZ (Universität, Forschungseinrichtung, -verbund)
    - Gauß-Allianz-Zentren, Community-Zentren
  - an kommerziellen Anbieter auslagern
    - Speicherdienste, Backup-Dienste, Archivierung …
    - Cloud-Dienste

- Was ist die beste Wahl?

# Selbstverwaltung: “Ich kann es selbst am besten!“

- Unter unmittelbarer Kontrolle einer Einrichtung oder der Community
- Volle Selbstbestimmung
- Konzepte können an die eigenen Bedürfnissen flexibel angepasst werden
- Service muss selbst erbracht werden
  - Zeitlicher Aufwand
  - Kosten
    - Hardware
    - Software
    - Personal
    - Unterhaltung
- Wie kann Nachhaltigkeit erreicht werden?

  z.B. für Langzeitarchivierung

# Wo sind wir in anderen Bereichen bisher “Selbstverwaltet”

- Much of the innovation now comes from software, not hardware!
- Software creates the added value
- Software allows to sell services
- Industry landscape shifting
  - from device-centric
  - to ecosystem-centric

# Auslagerung an wiss. Rechenzentren: “Wir sind die Profis!“

- Expertise und Erfahrung vorhanden
- Nähe zu wissenschaftlichen Anwendern, bündeln Anwenderinteressen
- Bisher: zumeist auf der Basis von „best effort“
  - Häufig noch keine Service Level Agreements
- Flexibilität begrenzt
  - Einheitliche Lösungen bevorzugt, da viele Anwender und knappe Ressourcen
- Meist zentral finanziert
  - Universität, Land, Bund
  - Nachhaltigkeit in gewissem Rahmen gesichert, da im Interesse einer größeren Gemeinschaft (alle anderen Nutzer)
  - Kosten bisher nicht vollständig transparent
- Auslagerung rechtlich meist einfach möglich – gleiche Domäne

# 100GE-Testbed Dresden – Freiberg (60/200/400km) / Teilprojekte

- **Synthetische Lasttests**
  - Messung und Optimierung der Leistungskennwerte aller Einzelkomponenten bei unregelmäßigen Datenverkehr und Überlastsituationen auf Transport- und Anwendungsebene
  - Durchgängige Auslastung zu 99.9% möglich
- **Cluster Virtueller Maschinen**
  - Stabilität und Verfügbarkeit 100%
  - Vmotion: Verschiebung von 100 Servern von DD nach FG in 7 Minuten (60km)
  - StorageVmotion: Verschiebung von 100 Servern von Storage in DD nach Storage in FG in 40 Minuten (60km)
  - Einfluss der Streckenlänge auf Latenz / Optimierungen notwendig
- **Network Attached Storage**
  - Testsystem BlueArc Titan-3200 (30 TB Sata)
  - Einfluss der Streckenlänge auf Latenz/Datendurchsatz von 75MB/s (60km) =>15 MB/s (400km) / Einfluss der Latenz – Optimierungen notwendig (lfd.)
- **Videokonferenz**
  - Auslastungsunabhängige Videokonferenz (auch im HD-Bereich) ohne Störeinflüsse!
- **3D-Visualisierung**

# Parallele Dateisysteme

- Evaluierung von drei etablierten Dateisystemen
- Hauptfragen:
  - kann die 100G-Leitung gesättigt werden
  - Einfluss der Latenz auf kleine und große Dateizugriffe
- Ergebnisse:
  - Performance-Werte z.T. über der Spezifikation des Herstellers
  - Lustre 21,8 GB/s Bandbreite (Maximum ist 25 GB/s)
  - FhGFS 22,58 GB/s
  - GPFS 19,4 GB/s
  - Latenz hat fast keinen Einfluss auf kleine Zugriffe
  - Interaktive Nutzung auch für auch für große Bandbreiten möglich

T-Systems
T-Systems Solutions for Research GmbH
IBM
GPFS

ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen
whamcloud
lustre

Fraunhofer
ITWM
Fraunhofer FS

# Beauftragung kommerzieller Anbieter: “Wir sind am günstigsten!“

- Expertise und Erfahrung vorhanden
- Service Level Agreements vorhanden
  - Kosten steigen aber mit den Anforderungen
- Kosten müssen vom Anwender aufgebracht werden
- Anwender sollte wissen, was er will
  - Geschäftsmodelle meist fixiert auf bestimmte Anwendungsszenarien
  - Consulting und Erweiterung eines Dienstes ist meist teuer
- Gefahr des “Vendor-Lock-in“
- Ist Auslagerung überhaupt möglich? - Datenschutzbestimmungen
- Ist mein (ausgelagertes) geistiges Eigentum sicher?

# Amazon, 21. April 2011 (Interrupt on April 21st)

**Computerwoche vom 26. April 2011**

http://www.computerwoche.de/management/cloud-computing/2370907/

**Nach massiven Problemen mit den Amazon Web Services (AWS) in den USA Ende letzter Woche läuft die Cloud-Infrastruktur mittlerweile größtenteils wieder.**

**Vor allem am Gründonnerstag und Karfreitag gingen allerlei auch durchaus namhafte EC2-Kunden - darunter die Webfirmen Foursquare, Reddit und Quora - zwangsweise vom Netz, nachdem bei AWS unter anderem Datenbankinstanzen ausgefallen waren. Die Amazon-Panne war der bislang größte Ausfall bei einem Cloud-Anbieter und unter anderem deswegen besonders bedenklich, weil sie mehrere eigentlich voneinander isolierte "Availability Zones" betraf.**

**Die Analyse von Amazon findet sich unter** http://aws.amazon.com/message/65648/

**Analysten hatten bisher geschätzt, dass Amazon in diesem Jahr mit AWS mehr als 500 Millionen Dollar erlösen könnte. Der E-Commerce-Riese aus Seattle sichert seinen Cloud-Kunden vertraglich eine Verfügbarkeit von 99,95 Prozent zu (< 6 Std. Ausfall pro Jahr).**

# Amazon, 12. April 2011 (Interrupt on April 21st)

*Email von Amazon an die betroffenen Kunden:*

*Hello,*

*A few days ago we sent you an email letting you know that we were working on recovering an inconsistent data snapshot of one or more of your Amazon EBS volumes. We are very sorry, but ultimately our efforts to manually recover your volume were unsuccessful.* ***The hardware failed in such a way that we could not forensically restore the data****.*

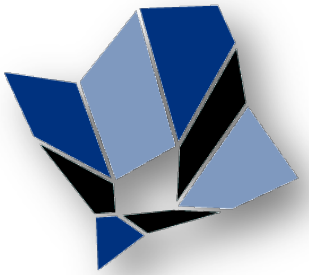
*What we were able to recover has been made available via a snapshot,* ***although the data is in such a state that it may have little to no utility...***

*If you have no need for this snapshot, please delete it to avoid incurring storage charges.*
***We apologize for this volume loss and any impact to your business.***

*Sincerely,*
*Amazon Web Services, EBS Support*

# Zusammenfassung

- IT-Versorgungskonzepte müssen heute nachhaltig in die Organisation von Prozessen eingreifen, um zukunftsfähige Strukturen zu schaffen
- Rationalisierungseffekte durch Kompetenzbündelung
  - Es gibt sie, gleichzeitig wachsen aber die Anforderungen an die Dienste und an die Qualität
  - Ziel: wirtschaftlicherer Personaleinsatz in den Versorgungsbereichen
- Die Kommission der IT-Infrastruktur der DFG (KfR) hat sich in diesem Kontext eindeutig positioniert
- Die Business-Modelle der Cloud-Anbieter für den Wissenschaftsbereich werden sich noch anpassen (Poltische Randbedingungen, Langfrist-beziehungen, Daten-Affinität)
- **Wir müssen bei unseren Veränderungsprozessen im IT-Bereich und in der organisatorischen Umsetzung sorgfältig – und auch vorsichtig – sein!**



# DFG-Projekt RADIESCHEN

Workshop Materialien

des Experten-Workshops

vom 17.April 2012

## Teilnehmerliste

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

# Teilnehmerliste Experten - Workshop "Elemente einer übergreifenden Forschungsdaten- Infrastruktur: Eine für Alle?"

| Titel | Vorname | Nachname | Affiliation |
|---|---|---|---|
| Dr. | Zaven | Akopov | DESY |
| | Henrike | Barkamann | Deutsches Zentrum für Luft- und Raumfahrt (DLR) |
| | Stefan | Bender | Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- u. Berufsforschung (FDZ) |
| | Roland | Bertelmann | Deutsches GeoForschungsZentrum Potsdam (GFZ), Bibliothek und Informationsdienste (LIS) |
| | Peer | Brauer | Christian-Albrechts-Universität zu Kiel |
| Dr. | Christoph | Bruch | Helmholtz Open Access Koordinationsbüro |
| Dr. | Andreas | Degkwitz | Universitätsbibliothek der Humboldt-Universität zu Berlin |
| | Frank | Dickmann | Universitätsmedizin Göttingen (UMG) |
| Dr. | Michael | Diepenbroek | World Data Center for Marine Environmental Sciences - PANGAEA (WDC-MARE) |
| | Kersten | Döring | Albert-Ludwigs-Universität Freiburg |
| | Friederike | Dressel | Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) |
| | Jürgen | Eils | Deutsches Krebsforschungszentrum Heidelberg (DKFZ) |
| | Thomas | Engel | Fachhochschule Mainz |
| | Claudia | Engelhardt | Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) |
| Dr. | Harry | Enke | Leibniz-Institut für Astrophysik Potsdam (AIP) |
| | David | Fichtmueller | Botanischer Garten und Botanisches Museum Berlin (BGBM) |
| | Dirk | Fleischer | Helmholtz-Zentrum für Ozeanforschung Kiel (GEOMAR) |
| | Petra | Gebauer | Freie Universität Berlin - Institut für Meteorologie |
| Prof. Dr. | Frank Oliver | Glöckner | Max-Planck-Institut für Marine Mikrobiologie (MPI) |
| | Anne | Grauholz | Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| | Anton | Güntsch | Botanischer Garten und Botanisches Museum Berlin (BGBM) |
| Dr. | Silke | Halstenberg | Karlsruher Institut für Technologie - Steinbuch Centre for Computing (KIT – SCC) |
| Dr. | Ulrich | Harms | Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| | Brigitte | Hausstein | GESIS, da\|ra |
| | Maurice | Heinrich | Deutsches Archäologisches Institut (DAI) |
| | Frank | Henze | Technische Universität Cottbus |

| | | | |
|---|---|---|---|
| Dr. | Rainer | Jonas | Helmholtz-Zentrum für Infektionsforschung (HZI) |
| | Brigitte | Jörg | Current Research Information Systems (euroCRIS) |
| | Maxi | Kindling | HU Berlin - Institut für Bibliotheks- und Informationswissenschaft (IBI) |
| Dr. | Jochen | Klar | Leibniz-Institut für Astrophysik Potsdam (AIP) |
| Dr. | Jens | Klump | Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| Dr. | Klaus | Köhler | Technische Universität Dresden - Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) |
| Prof. Dr. | Birgitta | König-Ries | Friedrich-Schiller-Universität Jena |
| Dr. | Daniela | Koudela | Technische Universität Dresden - Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) |
| | Jens | Krumpe | Friedrich-Schiller-Universität Jena |
| Dr. | Peter | Lampen | Leibniz-Institut für Analytische Wissenschaften (ISAS) |
| Dr. | Michael | Lautenschlager | Deutsches Klimarechenzentrum (DKRZ) |
| Dr. | Jörn | Lauterjung | Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| | Jens | Ludwig | Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) |
| Dr. | Wolfgang | Ludwig | Technische Universität München - ARB |
| | Daniel | Mallmann | Forschungszentrum Jülich |
| | Lars | Müller | Fachhochschule Potsdam |
| Dr. | Ralph | Müller-Pfefferkorn | Technische Universität Dresden - Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) |
| | Peter | Mutschke | GESIS – Leibniz-Institut für Sozialwissenschaften |
| Prof. Dr. | Wolfgang | Nagel | Technische Universität Dresden - Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) |
| | Gerald | Neumann | Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) |
| Dr. | Heike | Neuroth | Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) |
| | Claudia | Oellers | Rat für Sozial- und Wirtschaftsdaten (RatSWD) |
| Dr. | Sonja | Palfner | TU Berlin - Zentrum für Technik und Gesellschaft (ZTG) |
| | Heinz | Pampel | Helmholtz Open Access Koordinationsbüro |
| Dr. | Wolfgang | Peters-Kottig | Zuse-Institut Berlin (ZIB) |
| Dr. | Hans | Pfeiffenberger | Alfred-Wegener-Institut für Polar- und Meeresforschung (AWI) |
| | Holger | Przibytzin | Universitäts- und Landesbibliothek Münster (ULB Münster) |
| Dr. | Torsten | Rathmann | Deutsches Klimarechenzentrum (DKRZ) |

| | | | |
|---|---|---|---|
| | Matthias | Razum | FIZ Karlsruhe |
| Dr. | Bernd | Ritschel | Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| | Beate | Rusch | Zuse-Institut Berlin (ZIB) |
| Dr. | Leonie | Schäfer | Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| Dr. | Felix | Schäfer | Deutsches Archäologisches Institut (DAI) |
| Dr. | Sandra | Schaffner | Rheinisch-Westfälisches Institut für Wirtschaftsforschung Essen (RWI) |
| Prof. Dr. | Peter | Schirmbacher | HU Berlin - Institut für Bibliotheks- und Informationswissenschaft (IBI) |
| | Mirko | Schmidt | Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| Dr. | Stefan | Schmunk | Leibniz-Institut für Europäische Geschichte |
| | René | Schneider | Haute École de Gestion, Genf |
| | Thomas | Schnicke | Helmholtz-Zentrum für Umweltforschung (UFZ) |
| Dr. | Paul | Schultze-Motel | Helmholtz Open Access Koordinationsbüro |
| Dr. | Christian | Senger | Albert-Ludwigs-Universität Freiburg |
| | Elena | Simukovic | HU Berlin - Institut für Bibliotheks- und Informationswissenschaft (IBI) |
| | Claus | Spiecker | Fraunhofer-Informationszentrum Raum und Bau Stuttgart (IRB) |
| Dr. | Harald | Stachelscheid | Berlin-Brandenburg Center for Regenerative Therapies (BCRT) |
| | Stefan | Strathmann | Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) |
| | Ulla | Tschida | TU Berlin - Zentrum Technik und Gesellschaft (ZTG) |
| | Damian | Ulbricht | Deutsches GeoForschungsZentrum Potsdam (GFZ) |
| | Plaga | Vlado | Max-Planck-Institut für Psycholinguistik (MPI) |
| | Sven | Vlaeminck | Leibniz-Informationszentrum Wirtschaft (ZBW) |
| Dr. | Dirk | Walther | Max-Planck-Institut für Molekulare Pflanzenphysiologie (MPI - MP) |
| Dr. | Stefan | Winkler-Nees | Deutsche Forschungsgemeinschaft (DFG) |
| Dr. | Andreas | Witt | Institut für Deutsche Sprache Mannheim (IDS) |
| | Peter | Wittenburg | Max-Planck-Institut für Psycholinguistik (MPI) |
| | Thorsten | Wuest | Bremer Institut für Produktion und Logistik (BIBA) |
| | Benjamin | Zapilko | GESIS Leibniz-Institut für Sozialwissenschaften |
| | Saskia | Kirsch | Deutsches GeoForschungsZentrum Potsdam (GFZ) |