An der Schwelle zum Vierten Paradigma – Datenmanagement in der Klimaplattform

1 Das Vierte Paradigma

Das „Vierte Paradigma“ bezeichnet nach Jim Gray von Microsoft Research die Vision von einem neuen Wissenschaftsparadigma: nach Empirie, Theorie und Simulation befinden wir uns am Übergang zu einer „Daten getriebenen Wissenschaft“.(Gray, 2009. S. XVIII f.) Die informationstechnologische Verknüpfung der heterogenen Datenmengen eröffnet neue Dimensionen wissenschaftlicher Erkenntnis. Veranschaulichen lässt sich dies anhand „Virtueller Observatorien“: Astronomische Messdaten aus unterschiedlichsten Quellen werden unter einer Oberfläche zusammengeführt und dienen für sich als Datenbasis weiterer Forschung. Voraussetzung für diese eScience (enhanced Science) ist ein systematisches Datenmanagement: die Datenkuratierung.(Vgl. Bell, Hey, & Szalay, 2009; Lynch, 2009. S. 181.) Das heißt, die im Forschungsprozess angefallenen relevanten Daten werden gesichert, archiviert und nachnutzbar gemacht. Neben den Aspekten der Auffindbarkeit und Wiederverwertbarkeit von Datensätzen ist die Sicherung der Nachweiskette ein zentrales Motiv für diese Anstrengungen.(PARSE.Insight, 2009, S. 7.)

2 Datenmanagement in der Klimaplattform

2.1 Forschungsdomäne Klimawandel – Projekt Wibaklidama

An der Fachhochschule Potsdam wurde im Projekt Wibaklidama (Wissensbasiertes Klimadatenmanagement) die Praxis des Forschungsdatenmanagements untersucht, um anhand der Ergebnisse den Wandel mit zu gestalten. Im Fokus standen dabei die Einrichtungen der „Forschungsplattform Klimawandel“, kurz: Klimaplattform. Sie ist ein eingetragener Verein, bestehend aus 23 Wissenschaftseinrichtungen, mit dem Ziel, „Brandenburg-Berlin als Modellregion für das wissenschaftliche Verständnis und den Umgang mit den Folgen des Klimawandels im nationalen und internationalen Kontext zu platzieren.“ (http://www.klimaplattform.de/motivation-zur-klimaplattform.html) Die beteiligten Einrichtungen setzen sich zusammen aus

Forschungsinstituten, Hochschulen aber auch dem Deutschen Wetterdienst als staatliche Behörde. Das Themenspektrum der Forschungs- und Arbeitsfelder ist ähnlich breit gefächert wie die Art der Institutionen: von Agrarforschung über Gewässerforschung bis zu globalen Fragestellungen der Klimafolgenforschung. Allen Einrichtungen ist gemeinsam, dass sie klimaforschungsrelevante Daten erfassen, über sie verfügen oder mit ihnen arbeiten. Die Fachhochschule Potsdam betreut in der Klimaplattform das Querschnittsforum „Datenaustausch und -verfügbarmachung." (http://www.klimaplattform.de/foren/datenaustausch.html)

Mit qualitativen Erhebungen und der Durchführung von Workshops wurde im Projekt Wibaklidama die Praxis des Datenmenagements und der Bedarf an Forschungsdaten ermittelt. Im Dialog mit Datenmanagern wurden Ansätze zum Ausbau einer Dateninfrastruktur entwickelt. (Grossmann, 2010. S. 3.)

2.2 Daten in den Einrichtungen Klimaplattform

Bereits in dem begrenzten Forschungsfeld „Klimawandel" gibt es eine große Vielfalt an Forschungsdaten. Das Spektrum reicht von Daten aus limnologischen Instrumenten über Gravitationsmessungen durch Satelliten bis hin zu Wetteraufzeichnungen. Manche Einrichtungen erheben selbst keine Daten, sondern greifen auf die Bestände anderer zu und berechnen auf dieser Basis komplexe Modelle.(Vgl. Interviews Nr. 1,2,3,4,9,10) Wesentliches Kennzeichen klimaforschungsrelevanter Daten ist, dass sie großvolumig und heterogen sind. (Data-Pracitices-Befragung, Frage 11. Interview Nr. 2)

Bei der Analyse der Klimaplattform-Einrichtungen haben sich vier charakteristische Formen im Umgang mit Forschungsdaten herauskristallisiert (Müller, 2010. S. 12f.)

- Datenrepositorien für Messinstrumente
  Daten werden automatisch in Repositorien eingespielt und über Internetportale zugänglich gemacht.
- Disziplinspezifische Repositorien-Projekte und Datenzentren
  Zu speziellen Teildisziplinen werden Datenrepositorien unterhalten.
- Projektbezogene Datenpublikation (Publikation auf Website)

Eine Zwischenform der Datenpublikation stellt die Praxis dar, auf Internetseiten von Forschungsprojekten zugehörige Forschungsdaten zu publizieren. Diese sind damit zugänglich, aber nicht im Sinne von zuverlässigen Publikationsservern archiviert und auch nicht systematisch mit standardisierten Metadaten erschlossen.

- Einzelprojekte mit interner Ablage/Speicherung
  Die anfallenden Daten werden auf internen Speichersystemen abgelegt, ein Zugang von außen ist nicht vorgesehen.

Vereinzelt existieren in den Einrichtungen Abteilungen oder Arbeitsgruppen, mit der Aufgabe Datenservices für ihre Einrichtung zu Entwickeln und anzubieten.

2.3 Teilen von Daten: Bedarf und Vorbehalte

Eindeutig zeigt sich ein breiter Bedarf an Zugangsmöglichkeiten zu vorhandenen Forschungsdaten. „Wir erzeugen eigentlich gar nicht selber Daten, sondern nehmen die Daten Anderer“, (Interview Nr. 4.) so ein Datenmanager. Wichtig für die Nutzung fremder Daten sind „Geeichte und korrigierte, vergleichbare Messwerte (Saubere Rohdaten ohne Fehler und prozessierte Daten).“ (Befragung Informationsbedarf) Um gezielt und unmittelbar auf Forschungsdaten zugreifen zu können,wurde in der Online-Befragung mehrfach der Wunsch nach Verzeichnissen zu Datenbeständen sowie zu datenrelevanten Projekten und Publikationen geäußert.Der Bedarf an solchen Verzeichnissen verweist wiederum auf die Bedeutung hochwertiger Metadaten. Der mehrfach geäußerten Wunsch nach einem Expertenverzeichnis macht deutlich, dass der über Personen vermittelte Weg zu den Daten von großer Relevanz ist. Dies deckt sich auch mit den Ergebnissen aus unseren anderen Befragungen, dass Zugriff auf fremde Forschungsdaten oft über die persönliche Ansprache der Wissenschaftler/innen führt, die über „ihre“ Daten verfügen. (Vgl. Grossmann, 2010. S. 7.)

Die Bereitschaft, eigene Daten zu publizieren ist trotz des bekannten Bedarfs nicht sehr ausgeprägt. Folgende Aussage ist kein Einzelfall: „Metadaten werden nicht automatisch generiert, sondern das muss händisch gemacht werden.“ (Interview Nr. 4) Die Publikation von Daten wäre mit erheblichem Aufwand für die Forscher/innen verbunden. Zudem wurde die Furcht vor Missinterpretation publizierter Daten mehrfach hervorgehoben. Forscher/innen wollen sich also die

Interpretationshoheit sichern und damit wohl auch Wettbewerbsnachteile vermeiden.(Vgl. Interviews Nr. 1, 4, 6)

Auch auf Seiten der Datennutzer gibt es große Vorbehalte gegenüber fremden Daten. Dem grundsätzlichen Bedarf steht die Skepsis hinsichtlich der Qualität dieser Daten gegenüber: „Ich selber bin vorsichtig mit Daten, die ich nicht kenne“ (Interview Nr. 1) Für die Nachnutzung von Daten ist der Persönliche Kontakt zu den Datenerzeugern bzw. -inhabern von großer Bedeutung bei der Qualitätsbewertung. Somit gilt für Datenerzeuger wie -nutzer: „der Zugriff auf Daten muss im Dialog erfolgen“ (Interview Nr. 3., vgl. auch Grossmann, 2010. 7f.)

Es ist deutlich geworden, dass *gegenseitiges* Vertrauen eine Grundbedingung für die Nachnutzung von Daten ist. Die Herausbildung von allgemein anerkannten Publikationsformen für Forschungsdaten, die den üblichen Maßstäben für wissenschaftliche Veröffentlichungen entsprechen, ist deshalb fundamental wichtig, weil die qualitätsgeprüfte Publikation das Vertrauensverhältnis zwischen Datenerzeuger und Nutzer vermitteln kann.

2.4 Small Science

In den Untersuchten Einrichtungen dominiert nach wie vor die „Small Science.“ (Vgl. Grossmann, 2010) Überschaubare Forschungsprojekte werden zu spezifischen Fragen von wenigen Einzelpersonen durchgeführt. Die anfallenden Daten befinden sich dann am Arbeitsplatz der Wissenschaftler/innen, einen Austausch gibt es innerhalb der Arbeitsgruppe, darüber hinaus findet nur wenig Austausch statt. Nach Projektabschluss ist der Verbleib der Daten ungewiss, oft gehen sie verloren. Policies und Datenmanagementpläne existieren nicht, sind nicht bekannt oder werden nicht eingehalten.[1] Selbstverständlich findet in den Arbeitsgruppen ein spezifisches Datenmanagement statt. Allerdings dominieren projektbezogene Lösungen, wie die Ablage von Daten in einem einfachen Dateisystem. Faktisch sind bei „selbstgestrickten“ Lösungen die Daten ohne das Wissen der beteiligten Forscher/innen wertlos. Selbst wenn zentrale Services wie ein Datenrepository in der Institution existieren, werden sie nicht zwangsläufig genutzt. Als besonders

1 Ergebnis des Wibaklidama-Sommerworkshop, 22.6.2010; vgl. hierzu auch die Präsentation von von Jens Klump: http://wibaklidama.fh-potsdam.de/fileadmin/downloads/klump.2010-06022-wibaklidama.pdf

problematisch erweist sich die sorgfältige Erschließung mit Metadaten, die für Nachnutzung von Forschungsdaten unabdingbar, aber in der Praxis nur schwer umzusetzen ist, denn sie ist mühsam und zeitintensiv.

Dreh- und Angelpunkt für Nachnutzung von Daten sind die Datenerzeuger/innen selbst. Sie prüfen, ob sie Zugang gewähren können bzw. dürfen und verfügen über das erforderliches Kontextwissen. Das Auffinden erfolgt mittelbar über Projektinformationen und Publikationen.(Grossmann, 2010. S. 7.)

2.5 Big Science

Die Situation sieht in datenintensiven Großprojekten deutlich anders aus. „Data Center“ werden von einigen Klimaplattform-Mitgliedern selbst oder unter ihrer Beteiligung betrieben. Sie sind disziplinär ausgerichtet, hoch entwickelt und ermöglichen Erfassung nach Metadatenstandards, Online-Recherche und Download von Datensets. Die Daten werden laufend in die Systeme eingespielt und stehen der wissenschaftlichen Analyse zur Verfügung. In den Organisationsbereichen, in denen diese Großprojekte durchgeführt werden, existieren ausgearbeitete Datenpolicies, die Metadatenerfassung findet statt und erfolgt unter Einhaltung internationaler Standards. Der Zugang für externe Forscher/innen zu den Daten ist meist eindeutig geregelt und wird vielfach auch ohne Einschränkungen über Online-Portale ermöglicht. Angesichts hoher Kosten bei der Datengewinnung bzw. der Unwiederholbarkeit von Messungen wurde von Anfang an darauf geachtet, geeignete Standards und Formate zur Speicherung zu nutzen. Auch die digitale Langzeiterhaltung der Daten wurde geplant und ist weitgehend gesichert. Allerdings handelt es sich in vielen Fällen um Insellösungen und der Ausbau zu einer institutionen- und disziplinübergreifenden Infrastruktur, die auch die Small Science einschließt, steht aus. (Grossmann, 2010, S. 4. Müller 2010, S. 11f.)

2.6 Einrichtungstypen

Projekte im Sinne systematischer Forschungsdatenkuratierung gibt es innerhalb der Klimaplattform nur in den reinen Forschungseinrichtungen. Dort existiert zudem ein hohes Maß an internationaler

Vernetzung, Datenrepositorien werden entwickelt und betrieben. In der Regel gibt es auch Richtlinien zum Umgang mit Forschungsdaten. Allerdings gilt auch in diesen Forschungszentren die Unterscheidung zwischen Big Science und Small Science: Die Small Science existiert parallel zu den den großen Daten-Infrastrukturprojekten in den selben Einrichtungen. Es zeigt sich dennoch ein deutlicher Unterschied zwischen diesen großen und fachlich hoch spezialisierten Forschungsinstituten und den Hochschulen.

Forschungsdaten als Gegenstand oder Publikationsform eigener Arbeit haben wir an Hochschulen nicht gefunden. Zentrale Dienstleistungen der Rechenzentren beschränken sich dort auf Bereitstellung von IT-Infrastruktur und Gewährleistung der technischen Datensicherung. Da Forschungsdatenrepositorien disziplinspezifisch aufgebaut und betrieben werden, können die fachlich breit ausgerichteten Hochschulen diesen Service nicht leisten.

Ein Sonderfall in unserer Untersuchung war der Deutsche Wetterdienst. Er unterscheidet sich im Umgang mit seinen Daten sowohl von Forschungseinrichtungen als auch Hochschulen und wird deshalb als eigener Einrichtungstyp betrachtet. Die Daten, ihre Speicherung und Weitergabe sind das Kerngeschäft dieser Behörde. Die Handhabung der Daten leitet sich aus ihrem gesetzlichen Auftrag ab und ist genau geregelt. Datenspeicherung und Metadatenerfassung erfolgt gemäß internationaler Standards und stellt kein organisatorisches Problem dar. Auch Weitergabe und Verkauf von Daten ist eindeutig geregelt. (Vgl. Müller, 2010. S. 14f.) Bei der Entwicklung von Forschungsdateninfrastrukturen müssen folglich nicht nur die Wissenschaftsdisziplinen sondern auch die institutionellen Bedingungen unter denen sie ausgebaut werden, Beachtung finden.

3 Ergebnisse

3.1 Vier Brücken zur Forschungsdateninfrastruktur

Die Entwicklung einer Infrastruktur für die Data-Driven-Science im Sinne des „4. Paradigmas“ steht am Anfang. Es existieren zahlreiche Insellösungen. Für die Ausdehnung dieser Ansätze bedarf es weiterer Anstrengungen. Bezogen auf die Untersuchung der Klimaplattform lässt sich sagen, dass

in vier Feldern Brücken geschlagen werden müssen, um Forschungsdatenaustausch und -versorgung nachhaltig zu verbessern:

1. Zwischen den Disziplinen
   Gerade im interdisziplinär geprägten Forschungsfeld „Klimawandel“ bilden Daten aus Nachbardisziplinen häufig die wesentliche Basis für neue Arbeiten. Übergreifende technische wie qualitative Standards werden die Integration von Daten aus unterschiedlichen Systemen verbessern. Die Verwendung von verbreiteten Metadatenstandards muss den disziplinübergreifenden Transfer von Forschungsdaten unterstützen.
2. Von Big Science zu Small Science
   Große, projektübergreifende Systeme müssen auch die Forschungsdaten aus den Small Sciences aufnehmen und ganze Wissenschaftsdisziplinen abdecken.
3. Zwischen den Institutionen
   Aktive Kooperationen zwischen den drei Einrichtungstypen sind erforderlich, um die gesamte Wissenschaftslandschaft mit einer guten Forschungsdateninfrastruktur zu versorgen.
4. Zwischen technischen Systemen
   Die Wissenschaftlerarbeitsplätze müssen mit den Forschungsdatenspeichern/-repositorien unmittelbar gekoppelt werden können. Das gilt gleichermaßen für Datenerzeuger wie -nutzer. Es geht dabei darum, bereits im Entstehungsprozess Daten systematisch zu verwalten, automatisch mit Metadaten zu versehen und in archivtauglichen Formaten zu speichern. Für die Nachnutzung heißt dies, den direkten Zugriff mit Analyse- oder Visualisierungstools auf die Repositorien zu ermöglichen.

3.2 Ansatzpunkte

Für den Ausbau einer Daten-Infrastruktur zeichnen sich drei zentrale Aufgabenfelder ab.
1. Eine breite Einführung von Forschungsdatenpolicies muss auf die Institutsorganisation sowie die Förderung einer Publikationskultur für Forschungsdaten zielen. Damit werden die Sensibilisierung für das Thema Datenmanagement vorangetrieben und die institutionellen Rahmenbedingungen

verbessert. Es geht dabei beispielsweise um die Festlegung von verbindlichen Standardformaten, um wissenschaftliche Anerkennung für publizierte Datensätze und nicht zuletzt um ausreichende technische wie personelle Ausstattung des Datenmanagements. Auf dieser organisatorischen Basis kann die technische Infrastruktur aufsetzen. 2. Die Einführung von Semantic-Web-Technologien zum Datenmanagement und -austausch zeichnet sich als geeignet ab, um bestehende Einzellösungen zu einem Forschungsdatennetz zu verweben. Vielfalt und Ungleichzeitigkeit von Systementwicklungen erlauben kein zentral administriertes System. Durch die Verwendung eines offenen Standards wie Linked Data und möglichst offenen Schnittstellen sollen sich Verknüpfungen und Übergänge zwischen Disziplinen und Systemen selbst generieren. 3. Es bedarf der Entwicklung von Servicestrukturen, die Datenanbieter wie Nutzer von Verwaltungsaufgaben entlasten.

Angesichts der eingangs erläuterten Vision vom Vierten Paradigma wirken diese Maßnahmen freilich profan. Es zeigt sich, dass „Data-driven-Science“ ein Infrastrukturprojekt auf organisatorischer wie technischer Ebene ist. Die Entwicklungen stehen am Anfang und sind eine komplexe und langfristiges Aufgabe. Dennoch: Es gibt beeindruckende Technik und tragfähige Konzepte für datenbasierte Wissenschaften, die ein großes Potential für die Schaffung neuen Wissens bergen. Wir stehen an der Schwelle, diese Ansätze zu einer umfassenden Infrastruktur auszubauen.

Interviews

Die qualitativen Interviews wurden mit Datenmanager/inne/n aus jeweils unterschiedlichen Einrichtung der Klimaplattform geführt. Der Interviewleitfaden beinhaltete Art, Handhabung und Nutzung der Daten in der Einrichtung sowie künftigen Bedarf.

Interview-Nummer	Datum	Interviewer/in
Interview Nr. 1	3.6. 2009	Grossmann, Silke
Interview Nr. 2	5.6.2009	Grossmann, Silke
Interview Nr. 2b	24.7.2009	Grossmann, Silke
Interview Nr. 3	15.6.2009	Grossmann, Silke
Interview Nr. 4	17.6.2009 / 2.7.2009	Grossmann, Silke; Büttner, Stephan; Hobohm, Hans-Christoph
Interview Nr. 5	6.7.2009	Grossmann, Silke
Interview Nr. 6	13.7.2000	Grossmann, Silke; Büttner, Stephan
Interview Nr. 7	13.7.2009	Grossmann, Silke
Interview Nr. 8	30.7.2009	Grossmann, Silke
Interview Nr. 9	28.9.2009	Grossmann, Silke; Büttner, Stephan
Interview Nr. 10	14.9.2009	Grossmann, Silke

Literatur

Bell, G., Hey, T., & Szalay, A. (2009). Computer Science: Beyond the Data Deluge. Science, 323(5919), 1297–1298. doi:10.1126/science.1170411

Gray, J. (2009). Jim Gray on eScience: a Transformed Scientific Method. In T. Hey, S. Tansley, & K. Tolle (Eds.), The Fourth Paradigm, Data-Intensive Scientific Discovery (pp. xvii–xxxi). Redmond: Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf.

Grossmann, S. (2010). Datenmanagement in der Klimaforschung; Akteure, Bedarfe, Practices.: Projektbericht B1. http://wibaklidama.fh-potsdam.de/fileadmin/publikationen/wibaklidama-B1V01-04.pdf.

Lynch, C. (2009). Jim Gray's Fourth Paradigm and the Construction of the Scientific Record. In T.

Hey, S. Tansley, & K. Tolle (Eds.), The Fourth Paradigm, Data-Intensive Scientific Discovery (pp. 177–183). Redmond: Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part4_lynch.pdf.

Müller, L. (2010). Umfeldanalyse zum Klimadaten-Management: Projektbericht B2. http://wibaklidama.fh-potsdam.de/fileadmin/publikationen/wibaklidama-B2-V02-02.pdf.

PARSE.Insight (2009). Road Map. Deliverable D2.1. from http://www.parse-insight.eu/downloads/PARSE-Insight_D2-1_DraftRoadmap_v1-1_final.pdf.