# MULTIDIMENSIONALE, HETEROGENE, VISUALISIERBARE DATENRÄUME

Anforderungen, Entwurf und Implementierung
einer adaptiven und interaktiven Schnittstelle
für transdisziplinäre wissenschaftliche Daten
im Kontext der Erdsystemanalyse

**Dissertation**
**zur Erlangung des Grades des Doktors der Philosophie**
**am Fachbereich Politik- und Sozialwissenschaften**
**der Freien Universität Berlin**

**vorgelegt von**
**Markus Wrobel**

**Berlin 2004**

**Erstgutachter:** Prof. Dr.-Ing. habil. Ralf-Dirk Hennings

**Zweitgutachter:** Univ.-Prof. Dr.-Ing. Rupert Klein

**Datum der Disputation: 26.10.2004**

# Inhalt (Kurzübersicht)

# INHALT

# EINFÜHRUNG UND ÜBERBLICK

Dem Austausch von Wissen über die Welt kommt eine zentrale Bedeutung zu. Die hierfür im Lauf der Menschheitsgeschichte ersonnenen Hilfsmittel haben - von der Höhlenzeichnung bis zum Internet - die Effizienz hinsichtlich Zahl der erreichbaren Adressaten, Bandbreite der übertragbaren Informationen sowie Zeit- und Ortsunabhängigkeit ihrer Nutzung bis heute dramatisch erweitert (vgl. [Wersig 1983b] [Wersig 2000]). Die Berliner Informationswissenschaft hat bereits frühzeitig auf die durch das Vordringen der Computertechnologie ausgelösten tiefgreifenden Veränderungen (Informatisierung) hingewiesen [Wersig 1983a]. Sie konstatiert, dass einzelne wissenschaftliche Disziplinen - seien es nun Informatik, Ökonomie oder Geisteswissenschaften - jeweils nur spezifische Aspekte des Phänomens wahrnehmen [Wersig 1983c] und dass zur Bewältigung der durch Informations- und Kommunikationstechnologien ausgelösten dramatischen Umgestaltungen ein neuer wissenschaftlicher Ansatz erforderlich ist, der sich quer zu den bestehenden Disziplinen positioniert [Wersig 1993] [Wersig 1996].

Bereits 1982 wurde von der Informationswissenschaft der Anspruch formuliert, eine *Brückenfunktion* auszuüben, die vermittelnd zwischen der Informatik und den von dieser hervorgebrachten Technologien sowie den Menschen, die diese Technologien zur Lösung von Problemen einsetzen, auftritt [Wersig et al. 1982]. Die vorliegende Arbeit stellt sich in diese Tradition. Sie versteht sich als Beitrag zur Überwindung der immanenten Schwierigkeiten, die sich aus der Nutzung der immer komplexer werdenden Informations- und Kommunikationstechnologien zur Unterstützung moderner wissenschaftlicher Forschung ergeben.

Die heute verfügbaren Technologien erlauben es, bereits vorhandenes Wissen in Form digitaler Daten zu kodieren sowie gigantische Mengen neuer digitaler Daten zu erheben, zu sammeln und dauerhaft vorzuhalten, sie mit Computern immer höherer Leistungsfähigkeit zu verarbeiten sowie potentiell weltweit verfügbar zu machen. Diese technologischen Fortschritte korrespondieren mit einem permanenten Anstieg des potentiell verfügbaren Datenvolumens. Eine kürzlich veröffentlichte Studie der University of California in Berkeley schätzt den Anstieg des 2002 weltweit gespeicherten Datenvolumens gegenüber 1999 auf 30 Prozent; das Gesamtvolumen des 2002 neu gespeicherten Datenaufkommens wird auf rund 5 Exabyte ($5 * 10^{18}$ Byte) beziffert [UC SIMS 2003a].

Die neuen informationstechnologischen Potentiale haben auch der Forschung und Erkenntnisgewinnung neue Möglichkeiten eröffnet und werden entsprechend von nahezu allen Wissenschaftsdisziplinen genutzt [Drenth 2001]. Der Computertechnologie kommt dabei eine Schlüsselrolle bei der Koordination von Wissen und seiner Umsetzung in Innovationen zur Bewältigung der zentralen Probleme - seien es nun nachhaltige Entwicklungen, medizinische Fragen oder Verkehrs- und Transportprobleme - im 21. Jahrhundert zu [Mainzer 1999]. Mit dem heute erreichten Stand von Informations- und Kommunikationstechnologien besteht die einzigartige Option, zur Klärung drängender wissenschaftlicher Fragestellungen auf die jeweils besten hierzu verfügbaren Datenressourcen zurückzugreifen; die zur Erschließung von immer mehr und immer komplexeren Daten immer schneller hervorgebrachten Technologien sind hingegen von Einzelnen kaum noch zu überschauen.

Wersig betont, dass Daten als transformierbare Repräsentationen der Welt „[...] *letztlich der Sinnenwelt von Menschen zugänglich gemacht werden müssen*" [Wersig 2000, 14]. Diese Sichtweise auf Daten kann gleichsam als Leitmotiv der hier vorliegenden Arbeit gelten, die sich mit den Herausforderungen befasst, die sich aus der Erschließung multidimensionaler

und heterogener Datenräume ergeben. Sie ist motiviert durch die Einsichten, die der Autor als Informationswissenschaftler und Informatiker seit 1996 an einer neuartigen, transdisziplinären Forschungseinrichtung zur Erdsystemanalyse gewinnen konnte. Das 1992 gegründete Potsdam-Institut für Klimafolgenforschung, in dessen Kontext diese Arbeit entstand, integriert in einem holistischen Ansatz Disziplinen aus Natur-, Sozial- und Geisteswissenschaften, um den enormen wissenschaftlichen Herausforderungen zu begegnen, die durch den sich abzeichnenden Klimawandel aufgeworfen werden. Zentrale Voraussetzung hierfür ist der Einsatz modernster Computertechnologie sowie die Einbeziehung der weltweit besten verfügbaren Daten aus verschiedenen wissenschaftlichen Themengebieten, die im Institut vor Ort von jeweiligen Experten zusammengeführt und bereitgestellt werden. Die Komplexität der hieraus resultierenden Herausforderungen, die als Muster für moderne wissenschaftliche Anforderungen gelten können, bestärken den Autor in seiner Ansicht, dass die von der Informationswissenschaft geforderte Brückenfunktion nach wie vor unabdingbar ist.

Die vorliegende Arbeit strebt in diesem Sinne eine spürbare Verringerung konkret bestehender informationsbezogener Defizite durch Errichtung einer funktionierenden „Brücke" zwischen Menschen und Technologien an. Ihr Ziel ist die Konzeption und Realisierung einer geeigneten, flexiblen Schnittstelle, die der vielschichtigen Wissenschaftlergemeinde des Institutes - und potentiell jedem seiner weltweit verteilten Kooperationspartner - einen autonomen, komfortablen und funktionalen Zugriff auf wesentliche Bestandteile der dort zusammengeführten und bereitgestellten komplexen Datenräume eröffnet. Vielgestaltigkeit und Komplexität sowohl der anvisierten Anwendergemeinde wie der adressierten Datenräume erfordern hierzu sowohl eine Kombination moderner Technologien wie Datenbanken und Internet, die Realisierung intuitiver graphischer Nutzerschnittstellen und unterstützender Formen interaktiver Visualisierung sowie insbesondere die Entwicklung geeigneter Strategien zur Einbeziehung und Abbildung der vielfältigen und dynamischen Anforderungen.

Diese Arbeit folgt damit dem Handlungs-Imperativ, dem sich die Informationswissenschaft verpflichtet hat: Diese „*muß dafür sorgen, daß sich im informationellen Aktionsfeld die Zuführung benötigten Wissens interessen- und problemgeeignet vollzieht*"[1] [Wersig et al. 1982, 206]. Eine wichtige Metapher in diesem Zusammenhang sind die „Brücken", die zwischen Menschen und Technologien geschaffen werden müssen: „*Was wir brauchen, sind ‚Brücken' zwischen den Technologien und den Menschen, die humanrational, d.h. von der subjektiven und sozialen Problemsituation der Menschen aus, Technologien als deren Verlängerung gestalten* [...]" [Wersig 1983c, 299]. Solche Brückenschläge können nicht zuletzt auch dazu beitragen, Frustrationen sowohl bei Entwicklern von Technologien - die untergenutzt bleiben, weil sie an den Bedürfnissen der Anwender vorbeigehen - wie bei deren Anwendern - denen es an Hilfestellungen fehlt, die Technologien zur Bewältigung ihrer Probleme einzusetzen - abzubauen (vgl. [Wersig 1993, 157]).

Die Informationswissenschaft kann hier eine Brückenfunktion zwischen Ingenieur- und Sozialwissenschaften ausüben, „*indem sie ‚symbiotische Systeme' (die Problemlösungen durch Ineinandergreifen von maschinellen und menschlichen Komponenten anzielen) und Mechanismen der Hilfe zur kommunikativen und informativen ‚Selbsthilfe' ausbildet*" [Wersig et al. 1982, 4]. Dabei wird von der Auffassung ausgegangen, dass der Einsatz von Informations- und Kommunikationstechnologien in der Regel einen als System beschreibbaren Organisationszusammenhang einnimmt. Ferner wird zugrundegelegt, dass eine übergreifende Struktur von Problemen feststellbar ist, die aus einem „harten" Problemkern - der

[1] Schreibweise im Original.

zumeist technologisch lösbar ist - und einer diesen umlagernden, „weichen" Problemschale besteht, die in der Regel den menschlichen Faktor des Problems darstellt. Die Informationswissenschaft fordert hier die explizite Berücksichtigung der menschlichen Aspekte bspw. durch Problemlösungen, die nicht von den Bedingungen der Technologie, sondern von den Bedingungen der Menschen ausgehen und bei denen Menschen und Technologie aufeinander abgestimmte Lösungsanteile übernehmen [Wersig et al. 1982, 207]. Ferner soll die Informationswissenschaft sich „*in einer helfenden Rolle verstehen, die nicht das Ziel verfolgt, Abhängigkeit zu vergrößern, sondern die Betroffenen in ihren Möglichkeiten stärkt, sich selber zu organisieren*" [Wersig et al. 1982, 207].

Beide Forderungen werden in der vorliegenden Arbeit verfolgt. Zum einen kann die zu realisierende Schnittstelle, die Wissenschaftlern einen intuitiven und autonomen Zugriff auf die komplexen Datenbestände des Institutes eröffnen soll, als die Herausbildung eines „symbiotischen Systems" aus vielschichtigen menschlichen Aspekten und technologischen Komponenten aufgefasst werden, die geeignet zusammenfinden müssen. Dieses System kann nur dann erfolgreich sein, wenn es gelingt, die Bedürfnisse einer ihrerseits komplexen Nutzergruppe zu erfassen und in geeigneten Einsatz von Technologie abzubilden. Ob es dabei gelingt, eine funktionierende „Symbiose" herzustellen, wird unmittelbar anhand der Akzeptanz des Systems durch die Wissenschaftler abzulesen sein. Ferner kann die zu realisierende Schnittstelle auch als „Hilfe zur Selbsthilfe" aufgefasst werden: Sie dient dem Ziel, dem einzelnen Wissenschaftler größere Handlungsspielräume bei seiner individuellen Datenversorgung aus den Datenbanken des Institutes zu eröffnen und ihn dabei trotz der gegebenen, unabdingbaren Komplexität von einzubeziehenden Daten und Technologien vom technologiespezifischen Expertenwissen und der Mithilfe Dritter weitgehend unabhängig zu machen.

Die hierfür eingesetzten Technologien werden dabei im Sinne der Informationswissenschaft nicht als Selbstzweck, sondern als Hilfsmittel des Menschen angesehen: „*Auch wenn Technologien die zentrale Thematik bilden, sind diese doch nur in bezug auf die Problembewältigung von Individuen, Gruppen und Gesellschaft von Interesse*" [Wersig et al. 1982, 205].

Die Arbeit folgt folgender Struktur:

Ihr erster, allgemeiner Teil (Teil A) dient zur Einführung in die komplexe Thematik der Datenerschließung. Nach der Klärung zentraler Begriffe werden gegenwärtige Entwicklungstendenzen, Potentiale und Herausforderungen der Erschließung heterogener und multidimensionaler Datenräume anhand ausgewählter Konzepte und Technologien dargestellt. Zunächst wird die Problematik der Integration heterogener Daten am Beispiel des Data Warehousing behandelt, wobei auch auf das oft in Verbindung mit diesem zur Modellierung multidimensionaler Datenräume eingesetzte Online Analytical Processing (OLAP) sowie auf weitere Ansätze zu einer materiellen oder virtuellen Datenintegration eingegangen wird. Weitere Kapitel haben die hypothesenfreie Datenauswertung durch Data Mining sowie die computergestützte Visualisierung von Daten zum Gegenstand; einige Ausführungen zu Internet und World Wide Web sowie zu den bisher eher prototypisch realisierten Ansätzen des auf dem Internet aufsetzenden Grid-Computing beschließen diesen ersten Teil.

Teil B dient der Hinleitung zur Fragestellung. Zunächst wird das Potsdam-Institut für Klimafolgenforschung vorgestellt; daran anschließend erfolgt eine Analyse der Ausgangslage, eine Darstellung der spezifischen Voraussetzungen und Herausforderungen sowie die Konkretisierung der Aufgabenstellung.

In Teil C werden zunächst die eingesetzten Lösungsstrategien zur Erstellung der anvisierten geeigneten und flexiblen Schnittstelle zur Datenerschließung vorgestellt. Um der Viel-

schichtigkeit und Dynamik sowohl von Anwendern wie Datenräumen Rechnung zu tragen, wurde für die Entwicklung der Schnittstelle ein iterativer Prozess unter Einbeziehung von Prototypen gewählt, der es erlaubt, den jeweils erreichten Stand mit Anwendern zu erörtern und darauf basierend sowohl die erforderliche Software wie die über diese bereitgestellte Datenbasis schrittweise zu erweitern. Ferner dient dieser Teil zur Darstellung der Konzeption des gewählten Softwareentwurfes zur Umsetzung der Lösungsstrategien.

Teil D stellt die erzielten Ergebnisse dar. Dies umfasst die Beschreibung der Struktur der heute über die Schnittstelle zugänglichen Datenschicht ebenso wie die Darstellung der im iterativen Prozess in Wechselwirkung mit den Anwendern entwickelten Funktionalität und Ausgestaltung der graphischen Oberfläche.

Teil E dient zur zusammenfassenden Darstellung des erreichten Standes sowie seiner Bewertung. Ein Ausblick auf die hierauf aufbauende Vorgehensweise für eine Ausweitung der Datenerschließung im Rahmen eines 2004 beginnenden Projektes des Institutes beschließt diese Arbeit.