

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)

EINFÜHRUNG UND ÜBERBLICK	1
TEIL A HERAUSFORDERUNG DATENERSCHLIEßUNG	5
1 Vorklärungen	6
2 Die Integration heterogener Daten.....	15
3 Hypothesenfreie Datenauswertung – Data Mining	49
4 Computergestützte Datenvisualisierung.....	68
5 Internet, World Wide Web und Grids	93
TEIL B HINLEITUNG ZUR FRAGESTELLUNG	107
6 Das Potsdam-Institut für Klimafolgenforschung (PIK)	108
7 Zu erschließende Datenräume	118
8 Analyse der Ausgangslage.....	122
9 Voraussetzungen und Herausforderungen.....	128
10 Zusammenfassung der Aufgabenstellung.....	131
TEIL C LÖSUNGSSTRATEGIEN UND KONZEPTION DER SCHNITTSTELLE	133
11 Eingesetzte Lösungsstrategien	134
12 Systemanforderungen	149
13 Grobkonzept	151
14 Feinkonzept	160
15 Entwurf der Programmlogik	166
16 Client und Server.....	178
17 IDA – Interactive Digital Atlas.....	182
TEIL D ERGEBNISSE.....	199
18 Ausgestaltung der zugänglichen Datenschicht	200
19 Fensterstruktur und Hauptfenster	218
20 Filtermodule	223
21 Auswertungsmodule	246
22 Zeitreihenzugriff	258
TEIL E ERREICHTER STAND, BEWERTUNG UND AUSBLICK.....	267
23 Betrieb	268
24 Akzeptanz.....	272
25 Nachnutzungen	279
26 Bewertung des erreichten Standes und Ausblick.....	283
SCHLUSSBEMERKUNG.....	297
ANHANG.....	i
Tabellenverzeichnis	ii
Abbildungsverzeichnis	iv
Literaturverzeichnis	xii

INHALT

EINFÜHRUNG UND ÜBERBLICK	1
TEIL A HERAUSFORDERUNG DATENERSCHLIEßUNG.....	5
1 Vorklärungen.....	6
1.1 Zum Daten-Begriff	6
1.1.1 Abgrenzung von Wissen, Information und Daten	6
1.1.2 Daten und Metadaten	7
1.1.3 Datenräume	7
1.1.4 Größenordnungen.....	8
1.2 Klassifizierung nach Struktur und Auswertbarkeit	8
1.2.1 Strukturierte Daten.....	8
1.2.2 Semistrukturierte Daten	9
1.2.3 Unstrukturierte Daten.....	10
1.2.4 Zuordnung zu Speicherformen.....	11
1.3 Klassifizierung anhand von Attributeigenschaften.....	11
1.3.1 Unterscheidungen in der Statistik	11
1.3.2 Unterscheidungen in der Visualisierung.....	12
1.4 Multidimensionalität.....	13
1.5 Heterogenität.....	14
2 Die Integration heterogener Daten	15
2.1 Einführung.....	15
2.1.1 Formen der Heterogenität am Beispiel strukturierter Daten	15
2.1.2 Nachteile der Heterogenität am Beispiel von Unternehmensdaten.....	19
2.2 Das Konzept des Data Warehouse	21
2.2.1 Definition des Data Warehouse	21
2.2.2 Organisationsformen des Data Warehouse	23
2.2.3 Datenhaltung im Data Warehouse	26
2.2.4 Datenversorgung	29
2.2.5 Herausforderungen	33
2.3 Exkurs – Modellierung multidimensionaler Daten durch OLAP.....	34
2.3.1 Motivation – eine dimensionenbezogene Datensicht	34
2.3.2 Vergleich von OLAP und OLTP	35
2.3.3 Multidimensionales OLAP (MOLAP).....	36
2.3.4 Relationales OLAP (ROLAP)	36
2.3.5 OLAP-Funktionalität.....	39
2.4 Alternative Ansätze zur Datenintegration	40
2.4.1 Virtuelle oder materielle Datenintegration	40
2.4.2 Vermittler im Sinne der Anfrage – Mediatoren	41
2.4.3 Integration durch Migration	42
2.4.4 Teilverzicht auf Autonomie – Föderierte Datenbanksysteme	43
2.5 Fazit.....	47
3 Hypothesenfreie Datenauswertung – Data Mining	49
3.1 Einführung	49
3.1.1 Einflussfaktoren	49
3.1.2 Knowledge Discovery in Databases (KDD).....	50
3.1.3 Der KDD-Prozess	50
3.1.4 Varianten des Data Mining.....	52
3.2 Verfahren und Methoden.....	53
3.2.1 Segmentierung.....	53
3.2.2 Klassifikation	54
3.2.3 Assoziation.....	56
3.2.4 Anmerkungen.....	56
3.3 Anwendungsbeispiele	57
3.4 Web Mining	60

3.4.1 Ziele und Untergebiete	60
3.4.2 Web Content Mining	61
3.4.3 Web Structure Mining	61
3.4.4 Web Usage Mining	62
3.5 Data Mining, Data Warehouse und Datenschutz	64
3.5.1 Data Warehouse als Ausgangsbasis für Data Mining	64
3.5.2 Interessenkollisionen	64
3.5.3 Stellungnahme der Datenschutzbeauftragten	65
3.5.4 Klassifizierung potentieller Terroristen – Total Information Awareness	66
3.6 Fazit	67
4 Computergestützte Datenvisualisierung.....	68
4.1 Einführung	68
4.1.1 Visualisierung	68
4.1.2 Die Visualisierung wissenschaftlicher Daten – Scientific Visualization	71
4.1.3 Die Visualisierung abstrakter Daten – Informationsvisualisierung	73
4.2 Anforderungen an eine Datenvisualisierung	73
4.2.1 Expressivität	73
4.2.2 Effektivität	74
4.2.3 Angemessenheit	74
4.3 Visualisierungsprozess und Unterscheidungsmerkmale von Visualisierungstechniken	75
4.3.1 Der Visualisierungsprozess	75
4.3.2 Zwei- oder dreidimensionale Darstellungen	75
4.3.3 Statische oder dynamische Darstellungen	76
4.3.4 Vollständige oder unvollständige Darstellungen	76
4.4 Dimensionalität und Visualisierung	76
4.4.1 Begriffsklärungen	76
4.4.2 Herausforderung Dimensionalität und Wertebereich	78
4.5 Basistechniken	78
4.5.1 Balken- und Säulendiagramme	78
4.5.2 Histogramme	79
4.5.3 Kreisdiagramme	79
4.5.4 Linien- und Kurvendiagramme	80
4.5.5 Scatterplots	80
4.5.6 Verbunddiagramme	80
4.6 Visualisierung multivariater Daten	80
4.6.1 Panel-Matrizen	80
4.6.2 Streckenzüge	81
4.6.3 Ikonen	82
4.6.4 Pixelbasierte Techniken	82
4.6.5 Hierarchische Techniken	83
4.7 Visualisierung von Zeit- und Raumbezug	83
4.7.1 Visualisierung des Zeitbezuges	83
4.7.2 Visualisierung des Raumbezuges	84
4.8 Beispiele aus der Informationsvisualisierung	85
4.8.1 Baumstrukturen	86
4.8.2 Gleichzeitige Darstellung von Orientierungs- und Detailinformationen	87
4.8.3 Fokus und Kontext	88
4.8.4 Visualisierung von Dokumenten und Dokumentensammlungen	88
4.9 Fazit	91
5 Internet, World Wide Web und Grids	93
5.1 Einführung	93
5.1.1 Ursprünge und Konzept des Internet	94
5.1.2 Ursprünge und Konzept des World Wide Web	94
5.2 Internet und World Wide Web als Infrastruktur	95
5.2.1 Größenordnungen	95
5.2.2 Nutzung	97
5.2.3 Anbindung bestehender Ressourcen	99

5.3 Grids.....	99
5.3.1 Ansatz.....	99
5.3.2 Definition.....	100
5.3.3 Architektur.....	101
5.3.4 Anwendungsgebiete.....	102
5.3.5 Bisheriger Stand.....	105
5.4 Fazit.....	106
TEIL B HINLEITUNG ZUR FRAGESTELLUNG.....	107
6 Das Potsdam-Institut für Klimafolgenforschung (PIK).....	108
6.1 Kontext.....	108
6.2 Auftrag und Ansatz.....	109
6.2.1 Kernauftrag des Institutes.....	109
6.2.2 Transdisziplinarität.....	109
6.2.3 Die zentrale Rolle von Modellen, Simulation und Daten.....	110
6.3 Forschungsstruktur.....	111
6.3.1 Die fünf Abteilungen.....	111
6.3.2 Von Kernprojekten zu TOPIKs.....	113
6.4 Herausforderung Datenbasis.....	113
6.4.1 Notwendigkeit der Kompilierung.....	114
6.4.2 Immanente Komplexität und Heterogenität.....	115
6.4.3 Notwendigkeit der lokalen Bereitstellung.....	115
6.4.4 Notwendigkeit der Spezialisierung.....	116
6.4.5 Entstehen getrennter Datenräume.....	116
6.5 Wissenschaftliches Datenmanagement.....	117
6.6 Ziel – Eine geeignete Schnittstelle zur autonomen Datenerschließung.....	117
7 Zu erschließende Datenräume.....	118
7.1 Allgemeine Metadaten.....	118
7.1.1 Charakterisierung.....	118
7.1.2 Bedarf.....	119
7.2 Zeitreihenmetadaten.....	119
7.2.1 Charakterisierung.....	119
7.2.2 Bedarf.....	119
7.3 Punktverortete Zeitreihen.....	120
7.3.1 Charakterisierung.....	120
7.3.2 Bedarf.....	121
8 Analyse der Ausgangslage.....	122
8.1 Allgemeine Metadaten – CERA-2.....	122
8.2 Zeitreihenmetadaten und punktverortete Zeitreihen.....	124
8.3 Resultierende Defizite.....	125
8.3.1 Mangelnde Möglichkeiten der Orientierung.....	126
8.3.2 Mangelnde allgemeine Verfügbarkeit.....	126
8.3.3 Mangelnde direkte / individuelle Auswertbarkeit.....	126
8.3.4 Mangelnde Integration.....	126
8.3.5 Gleichbleibend hoher Aufwand.....	126
8.4 Folgerungen.....	127
8.4.1 Unzureichende Ausschöpfung.....	127
8.4.2 Geringe individuelle Autonomie.....	127
9 Voraussetzungen und Herausforderungen.....	128
9.1 Datenräume.....	128
9.1.1 CERA-2.....	128
9.1.2 Zeitreihenmetadaten und punktverortete Zeitreihen.....	128
9.2 Anforderungsbeitragende.....	128
9.2.1 Immanente Vielschichtigkeit und Heterogenität.....	128
9.2.2 Notwendigkeit einer übergreifenden Akzeptanz.....	129
9.2.3 Notwendigkeit einer Integration in den laufenden Forschungsbetrieb.....	129
9.3 Schwierigkeiten einer vollständigen Vorabdefinition.....	130

9.3.1 Einzubeziehende Kenntnisse und erforderlicher Zeitaufwand	130
9.3.2 Verschiedene fachspezifische Blickwinkel	130
9.3.3 Unterschiedliches Hintergrundwissen und Fachvokabular	130
10 Zusammenfassung der Aufgabenstellung	131
TEIL C LÖSUNGSSTRATEGIEN UND KONZEPTION DER SCHNITTSTELLE	133
11 Eingesetzte Lösungsstrategien	134
11.1 Einführung	134
11.1.1 Unterschiede zwischen Menschen	135
11.1.2 Unterschiede in Nutzergruppen	135
11.1.3 Faktoren für die Schnittstellengestaltung	136
11.1.4 Das Object-Action Interface Model	137
11.1.5 Mentale Modelle	138
11.1.6 Das unabdingbare Verständnis für die Anwender	139
11.2 Iterativer Entwicklungsprozess	140
11.2.1 Evolutionäres und inkrementelles Prozessmodell	140
11.2.2 Prototypen-Prozessmodell	141
11.2.3 Iterative Herausbildung einer zugänglichen Datenschicht	141
11.2.4 Die Schnittstelle als Katalysator	142
11.2.5 Vorgesehene Entwicklungszyklen	142
11.3 Unterstützung des Anwenders	144
11.3.1 Graphisch-interaktive Selektion von Raumbezügen	144
11.3.2 Interaktive Visualisierung selektierter Stationen und Zeitreihen	145
11.4 Reduzierung der Bedienungskomplexität	145
11.4.1 Modulare Datenselektion	145
11.4.2 Modulare Auswertung von Selektionsergebnissen	146
11.4.3 Vorteile	146
11.5 Flexibilität und Adaptivität	147
12 Systemanforderungen	149
13 Grobkonzept	151
13.1 Visuell-interaktive Bereitstellung von Raumbezügen	151
13.2 Internet-basierte Client-Server-Architektur	151
13.3 Realisierung mit Java	152
13.3.1 Exkurs: Vorteile objektorientierter Programmierung	153
13.3.2 Exkurs: Plattformunabhängige Bereitstellung	155
13.3.3 Exkurs: Applets vs. Applikationen	156
13.4 Anwenderseitig ablauffähiger ‚Fat‘ Client	158
13.5 Datenbank-Anbindung über JDBC	158
13.6 Abbildung von Datenräumen auf Tabellen	159
14 Feinkonzept	160
14.1 Aufteilung in drei Kernkomponenten	160
14.1.1 Kernkomponente <i>Selektion</i>	160
14.1.2 Kernkomponente <i>Abbildung</i>	160
14.1.3 Kernkomponente <i>Auswertung</i>	161
14.1.4 Iteration und Adaption	161
14.2 Modulare Datenselektion und Ergebnisauswertung	161
14.2.1 Filtermodule und Teilbedingungen	161
14.2.2 Auswertungsmodule	162
14.3 Konfigurierbare Generierung der Nutzerschnittstellen zur Laufzeit	162
14.3.1 Konfiguration der Kernkomponente <i>Selektion</i>	163
14.3.2 Konfiguration der Kernkomponente <i>Auswertung</i>	164
15 Entwurf der Programmlogik	166
15.1 Entwurf der Kernkomponente <i>Selektion</i>	166
15.1.1 Filtermodule	166
15.1.2 Architektur	169
15.1.3 Von der Konfiguration zur Generierung der Gesamtanfrage	171
15.2 Entwurf der Kernkomponente <i>Abbildung</i>	172

15.2.1 Konfiguration.....	172
15.2.2 Architektur.....	173
15.2.3 Adaptive interne Ergebnisrepräsentation	174
15.3 Entwurf der Kernkomponente <i>Auswertung</i>	176
15.3.1 Kommunikation zwischen unabhängigen Auswertungsmodulen.....	176
15.3.2 Architektur.....	177
16 Client und Server	178
16.1 Aufteilung und Kommunikation.....	178
16.2 Performance-Tuning.....	178
16.2.1 Unterstützung von Tabellenkaskaden	179
16.2.2 Dynamische Abbildung von Ergebniswerten.....	180
17 IDA – Interactive Digital Atlas	182
17.1 Systemidee und Anforderungen.....	182
17.2 Grobkonzept.....	183
17.2.1 Webfähiger Java-Client.....	183
17.2.2 Zweidimensionale Kartendarstellung	183
17.2.3 Verwendung von Vektordaten.....	184
17.2.4 Clientseitige Interaktionsverarbeitung und Bildgenerierung	184
17.3 Feinkonzept.....	184
17.3.1 Herausforderungen	184
17.3.2 Sukzessives Laden und clientseitiges Caching.....	184
17.3.3 Selbstbeschreibende Karten	185
17.3.4 Karten unterschiedlicher Detailgenauigkeit.....	186
17.3.5 Dreistufige Datenreduktion.....	187
17.4 Entwurf der Programmlogik.....	188
17.4.1 Interne Repräsentation der aktiven Karte	188
17.4.2 Dynamische Kartenbereitstellung	190
17.4.3 Transformation zwischen Koordinatensystemen.....	191
17.4.4 Visualisierung.....	192
17.4.5 Umsetzung der Nutzerinteraktionen.....	193
17.5 Interaktion.....	194
17.5.1 Vollständige Mausbedienbarkeit	194
17.5.2 Feedbackmodell.....	194
17.5.3 Interaktionsmodell.....	195
17.5.4 Konfigurierbare Raumauswahl.....	196
TEIL D ERGEBNISSE	199
18 Ausgestaltung der zugänglichen Datenschicht.....	200
18.1 Iterative Erweiterung	200
18.1.1 Betriebsphase I – Metadaten	200
18.1.2 Systemexterne Integrationsprozesse.....	201
18.1.3 Betriebsphase II – Metadaten und Zeitreihen	202
18.2 Allgemeine Metadaten.....	204
18.2.1 Von CERA-2 zu PIK CERA-2.....	204
18.2.2 Entitäten zur Selektion und Präsentation	205
18.3 Zeitreihenmetadaten	207
18.3.1 Hierarchische räumliche Klassifikation.....	207
18.3.2 Entitäten zur Selektion und Präsentation	210
18.4 Zeitreihen	216
18.4.1 Auswahl von Zeitreihen.....	216
18.4.2 Auswahl der zeitlichen Aggregation	216
19 Fensterstruktur und Hauptfenster.....	218
19.1 Fensterstruktur	218
19.2 Selektion einer Datenbankgruppe	219
19.3 Selektion eines Datenraumes	220
19.4 Definition und Auslösen der Anfrage.....	221
19.5 Optische Hervorhebung des Stationstyps	222

20 Filtermodule	223
20.1 Übergreifende Gestaltungskriterien	223
20.1.1 Eigene Dialogfenster	223
20.1.2 Übergreifende Funktionalität	224
20.1.3 Automatische Validierung beim Aktivieren	225
20.2 Beliebige Einzelattribute – SingleAttributeFilter	225
20.2.1 Motivation und Anforderungen	225
20.2.2 Umsetzung	226
20.3 Raumauswahl – SpatialFilter mit IDA	227
20.3.1 Motivation und Anforderungen	227
20.3.2 Umsetzung	228
20.4 Zeitfenster – TimeFrameSelector	231
20.4.1 Motivation und Anforderungen	231
20.4.2 Umsetzung	231
20.5 Vordefinierte Werte – ValueCombinator	232
20.5.1 Motivation und Anforderungen	232
20.5.2 Umsetzung	233
20.6 Transparente Suche – GlobalSearchFilter	234
20.6.1 Motivation und Anforderungen	234
20.6.2 Umsetzung	234
20.7 Hierarchische Thesauri – HierarchyBrowser	235
20.7.1 Motivation und Anforderungen	235
20.7.2 Umsetzung	235
20.8 Stationsklassifizierung – StationClassifier	237
20.8.1 Motivation und Anforderungen	237
20.8.2 Umsetzung	238
20.9 Statistische Anfragen – StatisticFilter	240
20.9.1 Motivation und Anforderungen	240
20.9.2 Umsetzung	241
20.10 Attribute für die Ergebnispräsentation – AttributeSelector	242
20.10.1 Motivation und Anforderungen	242
20.10.2 Umsetzung	243
20.11 Generieren und Ausführen einer Anfrage	245
21 Auswertungsmodule	246
21.1 Allgemeine Metadaten	246
21.1.1 Motivation und Anforderungen	246
21.1.2 Umsetzung	247
21.1.3 Überblick über die Anfrageergebnisse – ResultNavigator	248
21.1.4 Datensatzspezifische Details – EntryViewer	248
21.1.5 Visualisierung des Raumbezuges – BoundingBoxVisualizer	249
21.1.6 Online-Zugriffe	249
21.2 Zeitreihenmetadaten	251
21.2.1 Motivation und Anforderungen	251
21.2.2 Umsetzung	251
21.2.3 Interaktive tabellarische Präsentation – TableView	251
21.2.4 Visualisierung – StationVisualizer mit IDA	253
21.2.5 PostProcessor	256
22 Zeitreihenzugriff	258
22.1 Selektion	258
22.2 Die Schnittstelle zum Data Warehouse	258
22.3 Interaktive Visualisierung	259
22.3.1 Motivation und Anforderungen	259
22.3.2 Umsetzung	259
22.3.3 TimeSeriesSelector	259
22.3.4 TimeSeriesVisualizer	260
22.4 Export von Zeitreihen zum Anwender	262
22.4.1 Motivation und Anforderungen	262
22.4.2 Umsetzung	263

TEIL E ERREICHTER STAND, BEWERTUNG UND AUSBLICK	267
23 Betrieb.....	268
23.1 Aktuelle Konfiguration und Größe des Client	268
23.2 Bereitstellung des Client.....	268
23.3 Adaptierbarkeit	270
23.3.1 Software.....	270
23.3.2 Einbindung neuer Kartendaten	270
23.3.3 Einbindung neuer Datenräume	270
23.3.4 Erweiterung eingebundener Datenräume	271
23.3.5 Strukturelle Änderung eingebundener Datenräume	271
24 Akzeptanz	272
24.1 Zugängliche Daten	272
24.2 Zahl und Art der Zugriffe	274
24.3 Institutsexterne Sichtbarkeit	277
25 Nachnutzungen	279
25.1 CERA-Befüllung	279
25.2 ClimateDiagramGenerator	279
25.3 Webbasierte Präsentation von Forschungsergebnissen.....	281
26 Bewertung des erreichten Standes und Ausblick	283
26.1 Erzielter Nutzen für das Institut	283
26.1.1 Behebung der Ausgangsdefizite	283
26.1.2 Verbesserte Ausgangsbasis für Kooperationen	284
26.1.3 Offenheit für weiterführende Konzepte	285
26.1.4 Fazit	288
26.2 Bewertung der gewählten Vorgehensweise	288
26.3 Ausblick: Ausweitung auf gegitterte und polygonverortete Zeitreihen.....	289
26.3.1 Zeitreihen auf regelmäßigen Gittern	289
26.3.2 Sozioökonomische Zeitreihen auf Polygonstrukturen	290
26.3.3 Prototyp 1 – Gegitterte Zeitreihen	291
26.3.4 Prototyp 2 – Zeitreihen auf Polygonstrukturen.....	292
26.3.5 Die nächsten drei Jahre – das neue TOPIK-Projekt PlxDat.....	294
 SCHLUSSBEMERKUNG	 297
 ANHANG	 i
Tabellenverzeichnis	ii
Abbildungsverzeichnis.....	iv
Literaturverzeichnis	xii

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 Informatikwissenschaft 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 Wissenschaftlergemeinschaft 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 Anwendergemeinschaft 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.