8 Analyse der Ausgangslage

Kap. 8 gibt einen Überblick über die Ausgangslage zu Beginn dieser Arbeit. Die Betrachtung erfolgt zunächst getrennt nach allgemeinen Metadaten (Kap. 8.1) sowie nach Zeitreihenmetadaten und punktverortete Zeitreihen (Kap. 8.2). Anschließend werden die resultierenden Defizite dargestellt (Kap. 8.3) und entsprechende Folgerungen abgeleitet (Kap. 8.4).

8.1 Allgemeine Metadaten – CERA-2

Zur Verwaltung und Bereitstellung allgemeiner Metadaten beteiligte sich das PIK an der institutsübergreifenden Entwicklung des *Climate and Environmental Data Retrieval and Archive System 2* (CERA-2) [Toussaint et al. 1999] [Lautenschlager et al. 1998]. CERA-2 ist ein relationales Datenbankmodell, das ab 1996 unter Federführung des PIK in Zusammenarbeit mit dem Deutschen Klimarechenzentrum in Hamburg (DKRZ)[221] - jetzt Model and Data Group des Max-Planck-Institut für Meteorologie in Hamburg[222] -, dem Alfred-Wegener-Institut für Polarforschung (AWI)[223] in Bremerhaven und anfänglich dem Forschungszentrum Karlsruhe (FZK)[224] entwickelt wurde.

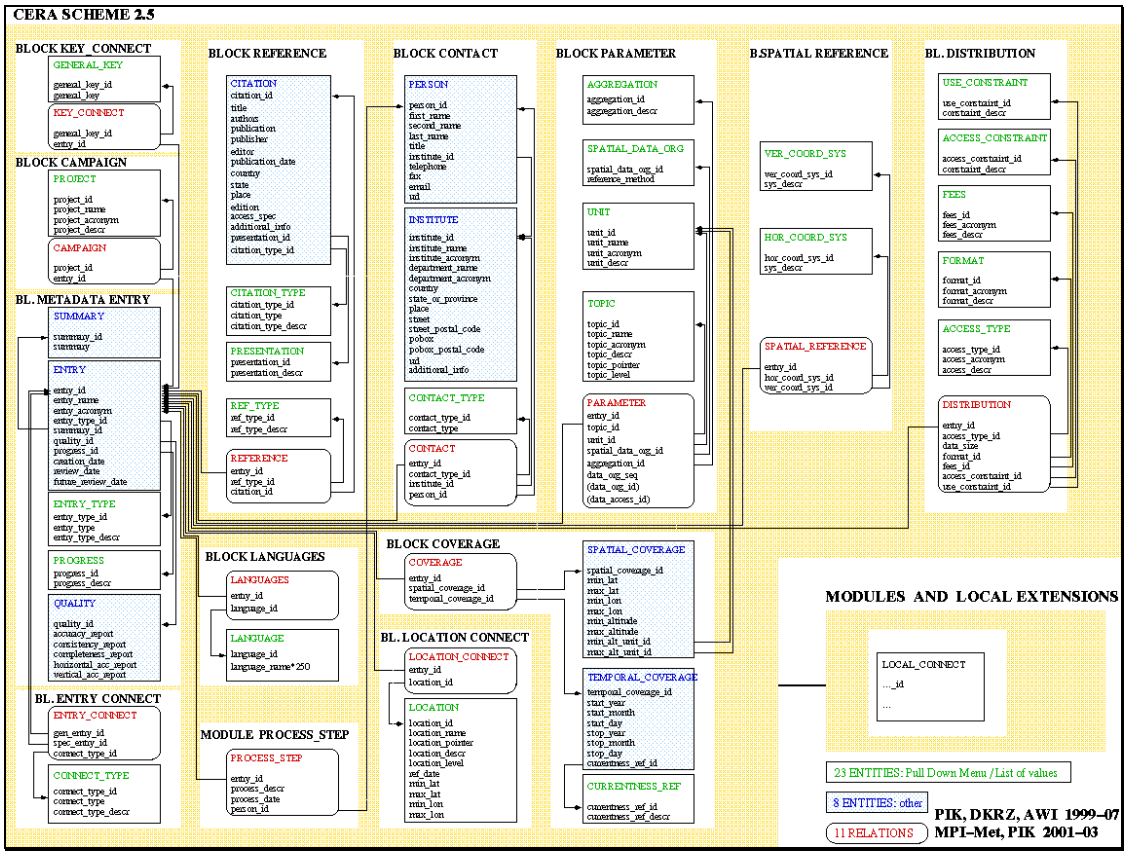

Abb. 8.1 - Der Kern von CERA-2, Version 2.5 [225].

Da diese geowissenschaftlichen Forschungseinrichtungen mit jeweils extrem inhomogenen Datenbeständen umgehen müssen, sollte mit CERA-2 ein konkretes Datenbankmodell zu deren Beschreibung und Austausch entstehen. Dabei wurde eine möglichst hohe Kompatibilität zu bestehenden inhaltlichen Standards - und hier insbesondere zum Directory Interchange Format (DIF) des Global Change Master Directory sowie dem Content Standard for

[221] http://www.dkrz.de/
[222] http://www.mad.zmaw.de/
[223] http://www.awi-bremerhaven.de/
[224] http://www.fzk.de/
[225] Abbildung übernommen von der CERA Central Page (http://www.pik-potsdam.de/cera/).

Digital Geospatial Metadata (CSDGM) des Federal Geographic Data Committee - angestrebt. Während das Vorläufermodell CERA-1 [Höck et al. 1995] insbesondere für die Beschreibung großer Mengen von Modelldaten ausgelegt war, wurde mit CERA-2 ein weitaus flexibleres Datenmodell entwickelt, um nun zusätzlich auch beliebige andere georeferenzierte Daten wie bspw. Messdaten, kartographische Daten etc. einheitlich beschreiben zu können.

Die Vielschichtigkeit der Daten, zu deren Beschreibung CERA-2 entworfen wurde, spiegelt sich in der Komplexität seiner Struktur (vgl. Abb. 8.1). CERA-2 besteht aus einem sog. Kern (CERA Core), fakultativ hinzuzunehmenden Modulen sowie Möglichkeiten für lokale Erweiterungen zur Abbildung jeweils institutsspezifischer Anforderungen. Der eigentliche Kern, der im Idealfall von allen CERA-2-Betreibern implementiert werden soll, um einen übergreifenden Datenaustausch zu vereinfachen, wurde aus Gründen der Übersichtlichkeit in Gruppen, den sog. Blöcken, organisiert. Jeder dieser Blöcke dient jeweils zur Dokumentation unterschiedlicher semantischer Teilaspekte eines Metadatensatzes:

Block *Metadata Entry*
- Ablage der grundlegenden Informationen zu jedem Eintrag. Hierzu zählen bspw. die Vergabe eines Titels, die thematische Einordnung anhand entsprechender Schlagwörter, Kurzbeschreibungen oder die Zuordnung zu internen Projekten.

Block *Coverage*
- Dokumentation der raumzeitlichen Überdeckung der zu beschreibenden Daten. Hierunter fallen bspw. die geographischen Extrema anhand von vier geographischen Koordinaten (*Boundingbox*), Höhe, Tiefe und der Name einer zugeordneten geographischen Region. Die zeitliche Abdeckung der Daten wird nicht über Datumsvariablen, sondern durch einen Satz natürlicher Zahlen beschrieben, um so auch prähistorische Zeiträume - etwa für Daten aus der Klimamodellierung - dokumentieren zu können.

Block *Spatial Information*
- Beschreibung weiterer Rauminformationen. Hierzu zählen verwendete Koordinatensysteme sowie zugrundeliegende räumliche Organisation wie Punkte, Gitter oder Vektoren.

Block *Attribute*
- Dokumentation der durch die Daten beschrieben Größen. Hier können physikalische, biologische oder chemische Messgrößen abgelegt werden; ferner Kenngrößen aus den Bereichen Wirtschaft, Soziales oder Politik sowie Angaben über die räumliche und zeitliche Aggregation der Daten.

Block *Status*
- Dokumentation des Zustandes der Daten, ihrer Qualität sowie von Informationen über eine mögliche Bearbeitung.

Block *Contact*
- Beschreibung von Kontaktstellen zu den Daten in Form von Personen und Institutionen. Hierzu zählt insbesondere die Quelle der Daten, ferner Angaben über Personen, die Auskunft über Entstehung oder Anwendung der Daten geben können, sowie über Inhaber von Nutzungsrechten.

Block *Distribution*
- Beschreibung von Zugangsmöglichkeiten zu den Daten. Dies umfasst bspw. Angaben zu Zugriffs- und Nutzungsbeschränkungen, Kosten, Format, Umfang der Daten und verwendetem Datenträger.

Block *Publication*
- Dokumentation von Publikationen, die mit den beschriebenen Daten in Zusammenhang stehen.

Um auch solche Metadaten flexibel integrieren zu können, die nicht von allen Anwendern benötigt werden, enthält CERA-2 zusätzlich weitere Tabellengruppen (sog. Module), die als fakultative, aber standardisierte Erweiterungen zu betrachten sind:

Modul *Data Access*
- Beschreibung von Zugriffsspezifikationen für einen automatisierten Zugriff auf die beschriebenen Daten. Hier können bspw. Dateinamen, Pfadnamen, Datenbanknamen o.ä. abgelegt werden.

Modul *Data Organization*
- Unterstützung der flexiblen Dokumentation der in der Praxis vorkommenden Kombinationen aus vierdimensionalen raumzeitlichen Gittern und beliebigen Punkten in der Raumzeit.

Zusätzlich können lokale Erweiterungen vorgenommen werden und andere Datenbanken integriert werden[226].

Zu Beginn dieser Arbeit war die Entwicklung des CERA-2-Datenbankmodells weitgehend abgeschlossen. Es bestanden allerdings *keine* geeigneten Zugriffsmöglichkeiten, die Anwendern eine flexible und komfortable Selektion von Metadaten aus dieser Datenbank gestattet hätten. Damit war am PIK zwar die Grundvoraussetzung zum Aufbau einer homogenen Datenbank zur einheitlichen Dokumentation seiner heterogenen Datenbestände gegeben; eine übergreifende, intuitive und effiziente Erschließung der komplexen allgemeinen Metadaten durch einzelne Wissenschaftler war jedoch durch das Fehlen einer funktionalen Schnittstelle zu diesem Zeitpunkt faktisch nicht möglich.

8.2 Zeitreihenmetadaten und punktverortete Zeitreihen

Im Gegensatz zu CERA-2 lagen die Datenräume der Zeitreihenmetadaten und der punktverorteten Zeitreihen im Institut in ebenfalls komplexer, zusätzlich jedoch *heterogener* Form vor. So befanden sich zu Beginn dieser Arbeit im Institut bereits punktverortete Zeitreihen einiger tausend Stationen, verteilt auf getrennte, von einzelnen Projekten und Abteilungen aufgebaute heterogene Datenbanken zu unterschiedlichen Themengebieten und räumlichen Ausschnitten. Weitere Zeitreihen lagen in Dateien oder in gedruckter Form vor. Ferner waren ebenfalls mehrere getrennte heterogene Datenbanken mit Zeitreihenmetadaten entstanden, die zur Dokumentation intern bereitgestellter sowie extern bei diversen Datengebern vorgehaltener Zeitreihen aufgebaut wurden.

(a)

Station_ID	Date	Var_1	Var_2	…
0815	01.01.1966	14.77	34	...
0815	02.01.1966	14.79	32	...
0815	03.01.1966	15.88	28	...
…	…	…	…	...

(b)

Stat_ID	Date	Var_ID	Value
4711	01.01.1966	1	50.00
4711	01.01.1966	2	12.03
4711	02.01.1966	1	46.44
…	…	…	…

Abb. 8.2 - Beispiel für die heterogene Modellierungen von Zeitreihen in zwei Datenbanken des PIK (vereinfachte Darstellung): Unterscheidung von Variablen über Attributnamen (a) und über Attributwerte (b).

Während Unterschiede durch eine Verwendung unterschiedlicher Datenmodelle sowie syntaktische Heterogenität (vgl. Kap. 2.1.1) durch Verwendung eines einheitlichen relationalen Datenbankmanagementsystems (Oracle) für die einzelnen Datenbanken des Institutes nicht gegeben waren, waren heterogene Formen der Datenorganisation entstanden. Dies soll kurz am Beispiel zweier getrennt aufgebauter Datenbanken des Instituts verdeutlicht werden, die beide zur Verwaltung punktverorteter Zeitreihen verwendet werden. So dient die von der Abteilung Climate System entwickelte globale Meteorologiedatenbank des

[226] So hält bspw. das DKRZ Modellergebnisse in Form von Binary Large Objects (BLOBs) in einem Relationalen Datenbankmanagesystem vor und dokumentiert die Zugriffsinformationen im CERA-Schema.

Institutes [Österle et al. 1999] zur Dokumentation von Zeitreihen einer festgelegten Zahl von 10 meteorologischen Variablen je Station, so dass es hier naheliegend war, die einzelnen Variablen anhand von *Attributnamen* zu unterscheiden (vgl. Abb. 8.2a). Die Zeitreihendatenbank des PIK-Kernprojektes RAGTIME[227] wurde hingegen für die Aufnahme von Zeitreihen einer großen Zahl unterschiedlicher Variablen konzipiert, so dass dort die Unterscheidung der einzelnen Variablen anhand von *Attributwerten* realisiert wurde (vgl. Abb. 8.2b). Zwischen beiden Datenbanken besteht damit ein Modellierungskonflikt (Attributname vs. Attributwert, vgl. Kap. 2.1.1).

Aufgrund der unterschiedlichen Anwendungsschwerpunkte beider Datenbanken ergaben sich ferner auch abweichende Modellierungen der Zeitreihenmetadaten. Tab. 8.1 fasst einige der so auftretenden Heterogenitäten zusammen.

Art der jeweils dokumentierten Information	**Zeitreihenmetadaten**		**Abweichungen von Datenbank 2 im Vergleich zu Datenbank 1**
	Datenbank 1: PIK-Kernprojekt RAGTIME	**Datenbank 2: globale Meteorologie-datenbank des PIK**	
Namen der Stationen	Attribut STAT_NAME	Attribut STATION	Synonymer Attributname
Typ der Stationen	Attribut STAT_TYPE	nicht erforderlich, da nur Meteorologie-Stationen	fehlendes, jedoch implizites Attribut
Beginn des Erhebungszeitraumes	Attribut BEGIN_DATE	Attribut TBEG	Synonymer Attributname
Ende des Erhebungszeitraumes	Attribut END_DATE	Attribut TEND	Synonymer Attributname
An den Stationen jeweils erhobene Variablen	Attribut VARIABLE	nicht erforderlich, da fixer Satz von 10 Variablen an jeder Station	fehlendes, jedoch implizites Attribut
Zeitliche Auflösung der Variablen	Attribut TEMP_RESOL	nicht erforderlich, da nur Variablen mit täglicher Auflösung dokumentiert	fehlendes, jedoch implizites Attribut

Tab. 8.1 - Heterogenitäten bei der Modellierung von Zeitreihenmetadaten in der Datenbank des PIK- Kernprojekts RAGTIME und der globalen Meteorologiedatenbank des PIK.

Für den Zugriff auf die einzelnen Datenbanken standen zu Beginn dieser Arbeit SQL, die Standardschnittstelle zur Abfrage relationaler Datenbanken, sowie einzelne proprietäre Programme und Skripte zur Verfügung. Es existierte jedoch keine geeignete Schnittstelle, um einzelnen Wissenschaftlern eine übergreifende, komfortable und autonome Erschließung dieser komplexen und heterogenen Datenräume zu ermöglichen. Für einen Zugriff waren entsprechende Spezialkenntnisse erforderlich, die bei der Mehrzahl der Wissenschaftler des Institutes nicht vorausgesetzt werden konnten.

8.3 Resultierende Defizite

Aufgrund ihrer Reichhaltigkeit, Komplexität und Heterogenität steht ein effizienter Zugang zu den am PIK vorgehaltenen Datenräumen vor besonderen Herausforderungen. Zu Beginn dieser Arbeit bestanden daher wesentliche Einschränkungen bezüglich einer autonomen individuellen Datenversorgung. Herauszuheben sind dabei insbesondere Beeinträchtigungen durch die nachfolgend beschriebenen Defizite.

[227] RAGTIME (Regional Assessment of Global Change Impacts Through Integrated Modelling in the Elbe River Basin) repräsentierte als eines der Kernprojekte des Institutes die *regionale*, flusseinzugsgebiets-orientierte Forschung des PIK. Hauptziel von RAGTIME war die Erforschung der Auswirkungen von Klimawandel, Landnutzung und anderer menschlicher Einflüsse auf hydrologische und ökologische Charakteristika [PIK 1996/97, 35ff.].

8.3.1 Mangelnde Möglichkeiten der Orientierung

Aus der Vielfalt von Daten, Speicherformen und jeweils zuständigen Ansprechpartnern resultierten in direkter Folge Schwierigkeiten, entsprechend individueller Bedürfnisse zu einem profunden Überblick über die lokal vorgehaltenen Daten zu gelangen. Da mangels einer geeigneten einheitlichen Zugriffsschnittstelle noch kein zentrales Informationsmedium zur komfortablen Auswertung der verfügbaren Datenräume bereitstand, war es häufig zeitaufwendig, herauszufinden, ob aktuell benötigte Daten bereits lokal verfügbar waren oder erst von entsprechenden Datengebern angefordert werden mussten.

8.3.2 Mangelnde allgemeine Verfügbarkeit

Die Erschließung der verschiedenen Datenbanken des Institutes konnte nur von einem geringen Teil der dort arbeitenden Wissenschaftler völlig autonom durchgeführt werden. So setzt ein effizienter Zugriff über SQL zunächst eine hinreichende Beherrschung dieser Sprache und genügend Praxis in ihrer Anwendung voraus. Ferner sind jeweils spezifische Hintergrundkenntnisse über den logischen Aufbau einer Datenbank, die Bedeutung von Tabellen, Attributnamen, verwendeten Bezeichnern für die Beschreibung der Daten etc. erforderlich. Auch die Verwendung vorhandener Programme und Skripte für die Automatisierung spezifischer Datenbankanfragen erfordert Hintergrundkenntnisse, die nicht generell vorausgesetzt werden können. Entsprechend konnte der Zugriff auf diese zentralen Ressourcen durch die Mehrzahl der Wissenschaftler des Institutes nicht selbständig, sondern in der Regel nur mit persönlicher Unterstützung entsprechender Spezialisten durchgeführt werden.

8.3.3 Mangelnde direkte / individuelle Auswertbarkeit

Eine Auswertung der Zeitreihenmetadaten zur Identifikation von Zeitreihen gemäss individueller Anforderungen erforderte in der Regel mehrstufige und aufwendige Prozesse. So konnte zumeist über die vorhandenen Datenbankschnittstellen nur eine grobe Vorauswahl getroffen werden, die anschließend durch Einbeziehung weiterer spezieller Werkzeuge wie Geoinformationssystemen, Statistik- oder Visualisierungssoftware - für die gegebenenfalls zunächst Datentransformationen durchzuführen waren - verfeinert werden musste. Stellte sich dabei heraus, dass die aus der Datenbank extrahierten Daten den gegebenen Anforderungen nicht genügen, musste der gesamte aufwendige Ablauf erneut durchgeführt werden. Eine schnelle, iterative Erschließung der Daten durch Selektion, Auswertung und entsprechende Anpassung der Selektionskriterien war aufgrund solcher zeitintensiver Prozesse deutlich erschwert.

8.3.4 Mangelnde Integration

Die für eine gegebene Anforderung potentiell relevanten Zeitreihenmetadaten und Zeitreihen konnten sich jeweils auf verschiedene heterogene Datenbanken mit unterschiedlichem Aufbau und unterschiedlichen Schnittstellen verteilen. Eine flexible integrierte Auswertung der verteilten Daten war entsprechend nur in eingeschränkter Form und mit erheblichem Aufwand möglich.

8.3.5 Gleichbleibend hoher Aufwand

Insbesondere die individuellen Prozesse einer *integrierten* Datenerschließung der verteilten Zeitreihenmetadaten und Zeitreihen waren aufgrund der Vielschichtigkeit möglicher Anforderungen an die Datenversorgung kaum aufeinander übertragbar. Entsprechend konnten Ansätze einer integrierten Nutzung kaum von dem Aufwand profitieren, der bereits in vorangegangene Zugriffe investiert worden war, so dass jeweils erneut zeit- und arbeitsaufwendige Prozesse durchzuführen waren.

8.4 Folgerungen

Aus den beschriebenen Defiziten ergaben sich mehrere Folgerungen, die als negativ für die Erschließung der verfügbaren Datenbasis durch die einzelnen Wissenschaftler zu bewerten sind. Zu nennen sind hier insbesondere unzureichende Möglichkeiten zur Ausschöpfung der vorhandenen Datenbasis sowie ein eindeutiger Mangel an individueller Autonomie bei der Datenerschließung.

8.4.1 Unzureichende Ausschöpfung

Die effiziente Nutzung der im Institut vorgehaltenen Datenbasis war aufgrund der gegebenen Heterogenität und Komplexität limitiert. Insbesondere konnte der potentiell verfügbare, sich auf mehrere Datenbanken verteilende Informationsgehalt von Zeitreihenmetadaten und Zeitreihen nur unzureichend oder nur mit sehr hohem Aufwand erschlossen werden.

8.4.2 Geringe individuelle Autonomie

Ein individueller und autonomer Zugriff von Wissenschaftlern innerhalb wie außerhalb des Instituts auf wesentliche Teile des dort vorgehaltenen Datenpools entsprechend individueller Anforderungen war nur in eingeschränktem Maße möglich. Als Ursachen hierfür sind sowohl Personenabhängigkeit, Zeitabhängigkeit wie Ortsabhängigkeit der Datennutzung zu nennen:

Personenabhängigkeit
- Wesentliche Teile der im Institut vorgehaltenen Datenbasis konnten von einer großen Zahl von Wissenschaftlern *nur unter Einbeziehung und direkter Mitwirkung Dritter*, die über die für Zugriff und Selektion jeweils erforderlichen Kenntnisse verfügen, durchgeführt werden.

Zeitabhängigkeit
- Bedingt durch die gegebene Abhängigkeit von der Mithilfe dritter Personen konnte der Zugriff auf wesentliche Teile der im Institut vorgehaltenen Datenbasis von einer großen Zahl von Wissenschaftlern *nicht zu beliebigen Zeiten* durchgeführt werden.

Ortsabhängigkeit
- Ebenfalls bedingt durch die gegebene Abhängigkeit von der Mithilfe dritter Personen konnte der Zugriff auf wesentliche Teile der im Institut vorgehaltenen Datenbasis von einer großen Zahl von Wissenschaftlern zumeist *nicht von beliebigen Orten* aus durchgeführt werden.

Insbesondere vor dem Hintergrund der transdiziplinären Ausrichtung des Institutes, die nicht zuletzt eine zunehmend disziplinübergreifende Datennutzung erfordert, bestand damit dringender Bedarf nach Konzeption, Realisierung und Bereitstellung geeigneter Möglichkeiten zur effizienten Unterstützung der Wissenschaftler bei der individuellen Erschließung der gegebenen, heterogenen und multidimensionalen Datenbasis.