7 Zu erschließende Datenräume

Entsprechend der Bedarfslage des Institutes wurden von diesem drei komplexe Datenräume als relevante Ziele für eine gemeinsame Erschließung vorgegeben. Da diese Daten sämtlich in Form von Datenbanken bereitgestellt werden, waren hier ausschließlich strukturierte Daten (vgl. Kap. 1.2.1) zu behandeln. Zu erschließen waren allgemeine Metadaten, die zur Beschreibung beliebiger Daten des Institutes verwendet werden (Kap. 7.1), ferner die am PIK für die Beschreibung punktverorteter Zeitreihen eingesetzten Zeitreihenmetadaten (Kap. 7.2) sowie die am Institut zusammengeführten punktverorteten Zeitreihen aus unterschiedlichen Wissenschaftsgebieten (Kap. 7.3). Nachfolgend werden diese drei Datenräume jeweils kurz charakterisiert und der zu Beginn der Entwicklung der Schnittstelle gegebene diesbezügliche Erschließungsbedarf formuliert.

7.1 Allgemeine Metadaten

7.1.1 Charakterisierung

Als Metadaten (vgl. Kap. 1.1.2) werden solche Daten bezeichnet, die Informationen über andere Daten bereitstellen. Da die Datenbestände der Erdsystemanalyse durch extreme Vielgestaltigkeit charakterisiert sind, die ihre einheitliche Speicherung in einem universellen Datenmodell vorn vorneherein ausschließen, gewinnt der Einsatz von Metadaten zur einheitlichen Beschreibung der heterogenen und komplexen Daten für eine effiziente Erschließung dieser Bestände besondere Relevanz. In Abgrenzung zu speziellen Arten von Metadaten (vgl. Kap. 7.2, Zeitreihenmetadaten) werden hier solche Arten von Metadaten, die zur Dokumentation beliebiger Daten der Erdsystemanalyse verwendet werden, als *allgemeine Metadaten* bezeichnet. Dass die Entwicklung eines universell geeigneten Satzes allgemeiner Metadaten in diesem Kontext keineswegs ein triviales Unterfangen ist, dokumentiert die Vielzahl nationaler und internationaler Standardisierungsbestrebungen, die in den letzten Jahren entstanden sind; zu nennen sind hier:

- das Directory Interchange Format (DIF) des Global Change Master Directory (GCMD) der NASA[212],
- der Content Standard for Digital Geospatial Metadata (CSDGM) des Federal Geographic Data Committee (FGDC)[213],
- INFOCLIMA der World Meteorological Organization (WMO)[214],
- der Standard[215] des Australia New Zealand Land Information Council (ANZLIC)[216] sowie
- der Standard ISO 19115 (Geographic Information-Metadata) der International Organization for Standardization (ISO)[217].

Allgemeine Metadaten im Kontext der Erdsystemanalyse dienen zur homogenen Beschreibung extrem heterogener Daten. Sie müssen entsprechend eine Vielzahl flexibler Beschreibungskriterien bereitstellen, die eine möglichst einheitliche Dokumentation dieser Daten erlauben. Hierzu zählen typischerweise Angaben über den thematischen Bezug der Daten, jeweilige Raum- und Zeitbezüge, textuelle Beschreibungen der Daten, vorliegende Speicherformen, Kontaktpersonen, Zugriffsinformationen und vieles mehr.

[212] http://gcmd.gsfc.nasa.gov/
[213] http://www.fgdc.gov/metadata/contstan.html
[214] http://www.wmo.ch
[215] http://www.anzlic.org.au/get/2358011755
[216] http://www.anzlic.org.au/index.html; vgl. auch [ANZLIC 2001].
[217] http://www.iso.ch

7.1.2 Bedarf

Zu Beginn dieser Arbeit bestand akuter und übergreifender Bedarf nach geeigneten Möglichkeiten zur Orientierung über die heterogene Datenbasis des Institutes anhand allgemeiner Metadaten. Es ist von wesentlicher Bedeutung, dass sich *potentiell jeder* Wissenschaftler innerhalb und außerhalb des PIK zu jeder Zeit und an jedem Ort autonom und effizient entsprechend seiner individuellen Anforderungen anhand der am PIK vorgehaltenen allgemeinen Metadaten über die so dokumentierten heterogenen Daten orientieren kann. Zentrale Voraussetzung hierfür sind geeignete Möglichkeiten für eine nutzerdefinierbare Selektion von allgemeinen Metadaten anhand von Anforderungskriterien, die die von diesen beschriebenen heterogenen Daten erfüllen müssen. Zudem sind die selektierten Metadaten so zu präsentieren, dass dem einzelnen Wissenschaftler ihre effiziente Auswertung gemäss seiner jeweiligen Anforderungen ermöglicht wird.

7.2 Zeitreihenmetadaten

7.2.1 Charakterisierung

Eine zentrale Datenressource für die Forschungsarbeit des PIK bilden multidisziplinäre punktverortete Zeitreihen (vgl. Kap. 7.3), die von einer Vielzahl von Erhebungsstationen - im folgenden auch kurz als *Stationen* bezeichnet - an unterschiedlichen Orten der Erde gewonnen werden. Zugangsvoraussetzung für die Erschließung solcher Daten ist die effiziente Auswertung von Informationen, die eine geeignete Auswahl einzelner Stationen erlauben. Solche Informationen werden hier als *Zeitreihenmetadaten* bezeichnet; sie umfassen dabei nicht die eigentlichen Zeitreihen, sondern enthalten vielmehr Angaben über die Stationen sowie über die diesen zugeordneten Zeitreihen. Die hier betrachteten Zeitreihenmetadaten dokumentieren typischerweise Stationen u.a. anhand ihres Namens und ihrer geographischen Position und enthalten insbesondere weitere Angaben, die eine Charakterisierung der den Stationen jeweils zugeordneten Zeitreihen erlauben.

Die am Institut dokumentierten Zeitreihenmetadaten können dabei in zwei Gruppen aufgeteilt werden. Die erste Gruppe bilden Zeitreihenmetadaten zur Dokumentation von Stationen, für die noch keine Zeitreihen im Institut vorgehalten werden. Sie dienen zur schnellen Information über extern verfügbare Zeitreihen, die dann bei Bedarf bei den jeweiligen Datengebern angefordert werden können. Die zweite Gruppe wird von Zeitreihenmetadaten gebildet, die solche Stationen dokumentieren, denen lokal im Institut vorgehaltene Zeitreihen zugeordnet sind. Sie dienen zur schnellen Information über intern verfügbare Zeitreihen.

7.2.2 Bedarf

Zu Beginn dieser Arbeit bestand akuter und übergreifender Bedarf nach einem autonomen und effizienten Zugang zu den am PIK vorliegenden Informationen über Stationen, die für eine Orientierung über und die Identifikation von extern wie intern vorgehaltenen punktverorteten Zeitreihen unabdingbar sind. Es ist von wesentlicher Bedeutung, dass sich *potentiell jeder* Wissenschaftler innerhalb und außerhalb des PIK zu jeder Zeit und an jedem Ort autonom und effizient entsprechend seiner individuellen Anforderungen anhand der im PIK vorgehaltenen Zeitreihenmetadaten über die so dokumentierten Stationen informieren kann. Zentrale Voraussetzung hierfür ist jeweils die nutzerdefinierbare Selektion von Zeitreihenmetadaten anhand von Anforderungskriterien, die die von diesen beschriebenen Stationen erfüllen müssen; selektierte Zeitreihenmetadaten sind zudem so zu präsentieren, dass dem einzelnen Wissenschaftler ihre effiziente Auswertung gemäss seiner jeweiligen Anforderungen ermöglicht wird.

7.3 Punktverortete Zeitreihen

7.3.1 Charakterisierung

Es wurde bereits ausgeführt, dass für die überwiegende Zahl der hier betrachteten Daten sowohl ein Bezug zu geographischen Räumen wie ein Bezug zur Zeit gegeben ist (vgl. Kap. 6.4.2). Zentrale Datenressourcen der Erdsystemanalyse liegen dabei in Form von punktverorteten Zeitreihen vor.

- **Punktverortete Zeitreihen**

Als *Zeitreihe* wird hier eine Sammlung zeitlich aufeinanderfolgender geordneter Werte bezeichnet, die für jeweils einen definierten Parameter und für einen definierten geographischen Bezug vorliegen. Zeitreihen können sowohl durch Messungen, durch Aufbereitung empirischer Daten oder durch Berechnungen entstehen. Um von den Aspekten der Datengewinnung zu abstrahieren, soll, wenn eine diesbezügliche Unterscheidung nicht erforderlich ist, nachfolgend einheitlich von *erhobenen*[218] Zeitreihen die Rede sein. Als *punktverortete Zeitreihen* werden hier solche Zeitreihen bezeichnet, deren geographischer Bezug jeweils durch einen Koordinatenpaar beschrieben ist. Im hier betrachteten Kontext sind punktverortete Zeitreihen dabei stets im Zusammenhang mit jeweils fest verorteten Stationen zu sehen. Die Identifikation punktverorteter Zeitreihen erfordert daher in jedem Fall die Einbeziehung von Zeitreihenmetadaten, die eine geeignete Charakterisierung dieser Stationen erlauben (vgl. Kap. 7.2). Da im Rahmen dieser Arbeit ausschließlich punktverortete Zeitreihen[219] adressiert werden, werden diese im Folgenden auch kurz als *Zeitreihen* bezeichnet.

Jede Zeitreihe wird neben Angaben zu ihrem geographischen Bezug anhand mehrerer weiterer Beschreibungskriterien charakterisiert, aus denen abgeleitet werden kann, welche Parameter erhoben wurden, in welchen zeitlichen Abständen die Erhebung stattgefunden hat und über welchen Zeitraum sie durchgeführt wurde. Jede Zeitreihe ist daher durch entsprechende Angaben über die erhobene Variable, die zeitliche Auflösung, die zeitliche Abdeckung sowie die Vollständigkeit charakterisiert:

- **Variablen und zeitliche Auflösungen**

Der jeweilige Parameter, der für eine individuellen Zeitreihe erhoben wurde, wird hier als *Variable* bezeichnet. Jede Zeitreihe enthält dabei nur Werte, die für genau eine Variable erhoben wurden; liegen zeitlich aufeinanderfolgende Werte für unterschiedliche Variablen und den selben geographischen Bezug vor, handelt es sich entsprechend um unterschiedliche Zeitreihen. Entsprechend der Vielfalt der im PIK zusammengeführten Wissenschaftsdisziplinen liegen im Institut Zeitreihen zu einer erheblichen Anzahl unterschiedlicher Variablen vor. Die Frequenz, mit der die Werte einer Zeitreihe erhoben werden, wird hier als die *zeitliche Auflösung* der Zeitreihe bezeichnet. Die Werteerhebung kann dabei in unterschiedlichen regelmäßigen oder auch unregelmäßigen Abständen stattfinden. Jede hier

[218] Vgl. die Herangehensweise von [Schumann, Müller 2000, 29], die im Kontext der Visualisierung von Daten mit dem Begriff des Beobachtungsraumes ebenfalls von der Art der Datenerzeugung abstrahieren (vgl. Kap. 4.4.1).

[219] Neben punktverorteten Zeitreihen lassen sich im Kontext der Erdsystemanalyse zwei weitere Gruppen von Zeitreihen abgrenzen. *Zeitreihen auf Polygonstrukturen* sind solche Zeitreihen, deren geographischer Bezug jeweils durch ein oder mehrere Polygone beschrieben ist; typische Vertreter dieser Gruppe sind wirtschaftliche und soziologische Zeitreihen, die für einzelne Kontinente, Staaten oder andere vordefinierte geographische Einheiten erhoben werden. In Abgrenzung hierzu fallen in die Gruppe der *Zeitreihen auf regelmäßigen Gittern* solche Zeitreihen, deren geographischer Bezug jeweils durch eine eindeutige Position auf einem regelmäßigen Gitter beschrieben ist. Hierbei handelt es sich um berechnete Daten, die basierend auf gemessenen (punktverorteten) Zeitreihen erzeugt werden, oder um Ausgabedaten von Modellen. Vgl. den abschließenden Ausblick auf die geplante Ausweitung der Datenerschließung und die hierfür entwickelten Prototypen (Kap. 26.3).

betrachtete Zeitreihe besitzt dabei genau *eine* zeitliche Auflösung; liegen zeitlich aufeinanderfolgende Werte für die selbe Variable und den selben geographischen Bezug in unterschiedlichen zeitlichen Auflösungen vor, handelt es sich entsprechend um unterschiedliche Zeitreihen.

- **Zeitliche Abdeckungen und Vollständigkeit**

Als *zeitliche Abdeckung* einer Zeitreihe wird hier der maximale Zeitraum bezeichnet, für den Werte erhoben wurden; die zeitliche Abdeckung einer Zeitreihe ist entsprechend definiert durch den frühesten und den spätesten Zeitpunkt, für den für diese Reihe Werte vorliegen. Die *Vollständigkeit* einer Zeitreihe ergibt sich aus dem Verhältnis der Anzahl möglicher Werte - definiert durch zeitliche Abdeckung sowie zeitliche Auflösung der Zeitreihe - sowie der Anzahl der tatsächlich erhobenen Werte[220].

7.3.2 Bedarf

Zu Beginn dieser Arbeit bestand akuter und übergreifender Bedarf nach einem autonomen und effizienten Zugang zu den am PIK vorgehaltenen punktverorteten Zeitreihen, die als einzigartige und kostbare Datenressource für vielfältige Forschungsaufgaben einzustufen sind. Es ist von wesentlicher Bedeutung, dass *potentiell jeder* Wissenschaftler innerhalb und außerhalb des PIK zu jeder Zeit und an jedem Ort autonom und effizient entsprechend seiner individuellen Anforderungen eine Auswahl aus den im PIK vorgehaltenen punktverorteten Zeitreihen treffen und auf diese zur wissenschaftlichen Nutzung *direkt* zugreifen kann. Zentrale Voraussetzung hierfür ist jeweils ihre Selektion anhand der im Institut vorgehaltenen Zeitreihenmetadaten. Selektierte punktverortete Zeitreihen sind dabei den Wissenschaftlern so zugänglich zu machen, dass ihre Weiterverwendung bspw. für Aufgaben der Modellierung oder Simulation in möglichst geeigneter Weise sichergestellt ist.

[220] Die zeitliche Abdeckung einer Zeitreihe lässt entsprechend noch keine Rückschlüsse auf ihre Vollständigkeit zu: So besitzt etwa eine Zeitreihe mit täglicher Auflösung, die Werte für jeden Tag zwischen dem 1.1.1800 und dem 31.12.1999 enthält, die selbe zeitliche Abdeckung wie eine Zeitreihe mit täglicher Auflösung, die nur aus jeweils einem Wert für den 1.1.1800 und dem 31.12.1999 besteht.