TEIL D
ERGEBNISSE

In diesem Teil werden die im iterativen Entwicklungsprozess erzielten Ergebnisse sowohl hinsichtlich der Ausgestaltung der über die Schnittstelle bereitgestellten Datenschicht wie der konkreten Gestaltung von graphischer Oberfläche und Funktionalität dargestellt. Bereits vorab ist festzuhalten, dass die angestrebten Ziele bezüglich einer autonomen Datenversorgung aus allgemeinen Metadaten, Zeitreihenmetadaten und punktverorteten Zeitreihen des Potsdam-Instituts für Klimafolgenforschung durchweg erreicht werden konnten. Sämtliche im Kontext dieser Arbeit adressierten, zuvor nur isoliert zugänglichen und heterogenen Datenbanken des Instituts sind heute über die entstandene Schnittstelle in komfortabler und flexibler Weise zugänglich. Durch den von der Scientific Data Management Group durchgeführten Aufbau von integrierten Datenbanken zur Bereitstellung von Zeitreihenmetadaten und punktverorteten Zeitreihen können diese Daten den Anwendern nun zudem in homogener Weise zur Verfügung gestellt werden. Die Darstellung gliedert sich wie folgt:

In Kap. 18 wird zunächst die Ausgestaltung der über die Schnittstelle zugänglichen Datenschicht skizziert. Ausgehend von einer Beschreibung der iterativen Erweiterung der in die Datenschicht integrierten Datenräume werden die Entitäten dargestellt, die für den Zugriff auf allgemeine Metadaten, auf Zeitreihenmetadaten sowie auf Zeitreihen verwendet werden.

Die verbleibenden vier Kapitel dieses Teils beschreiben die in Wechselwirkung mit den Anwendern erzielten Ergebnisse bei der Gestaltung von graphischer Oberfläche und Funktionalität der Schnittstelle. Kap. 19 erläutert die gewählte Fensterstruktur sowie das Hauptfenster der Schnittstelle; nachfolgend werden die einzelnen Filtermodule vorgestellt, die für die Formulierung von Selektionskriterien anhand von Teilbedingungen entwickelt wurden (Kap. 20). Die zur Präsentation der jeweiligen Anfrageergebnisse aus allgemeinen Metadaten und Zeitreihenmetadaten sowie zur Interaktion mit diesen entwickelten Auswertungsmodule werden in Kap. 21 dargestellt; abschließend werden die Nutzerschnittstellen zum Zugriff auf Zeitreihen beschrieben (Kap. 22).

18 Ausgestaltung der zugänglichen Datenschicht

Kap. 18 beschreibt die Ausgestaltung der über die Schnittstelle zugänglichen Datenschicht. Die Darstellung gliedert sich in die iterative Erweiterung der Datenschicht (Kap. 18.1) sowie spezifische Aspekte bezüglich allgemeiner Metadaten (Kap. 18.2), Zeitreihenmetadaten (Kap. 18.3) und Zeitreihen (Kap. 18.4).

18.1 Iterative Erweiterung

18.1.1 Betriebsphase I – Metadaten

Die Entwicklung der Kernfunktionalität der Schnittstelle (erster Zyklus, vgl. Kap. 11.2.5) wurde 1999 abgeschlossen. Die - prototypische - Betriebsphase I (zweiter Zyklus) startete in der zweiten Hälfte des Jahres 1999[271] und eröffnete den Zugriff auf allgemeine Metadaten sowie auf Zeitreihenmetadaten (vgl. Abb. 18.1).

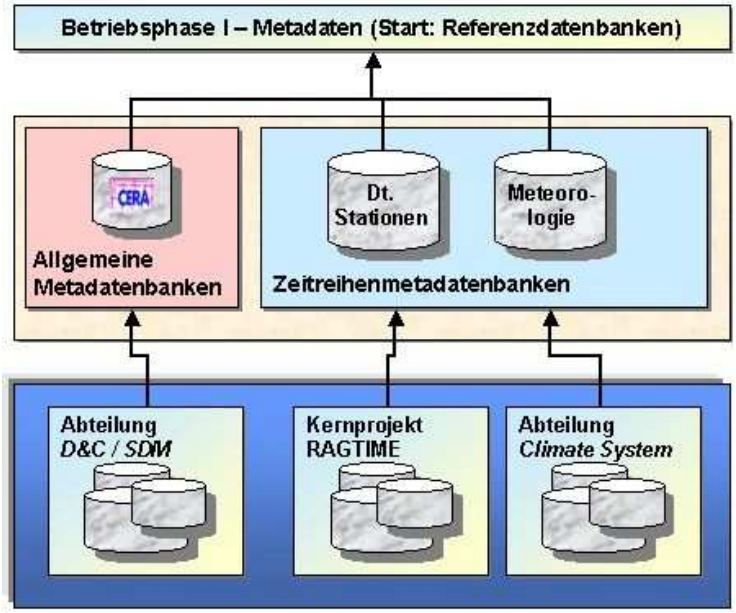

Abb. 18.1 - Zu Beginn von Betriebsphase I eingebundene Metadatenbanken.

Zu Beginn wurden die Referenzdatenbanken bereitgestellt, anhand deren die Entwicklung der Schnittstelle bis zu diesem Zeitpunkt durchgeführt wurde. Für die Wissenschaftler des Instituts bestanden damit zunächst Möglichkeiten zum Zugriff auf folgende Informationsquellen:

- verschiedene thematische Ausschnitte der allgemeinen Metadatenbank CERA-2, entwickelt von der Scientific Data Management Group;
- eine Datenbank mit Zeitreihenmetadaten zur Dokumentation deutscher Stationen, entwickelt vom PIK-Kernprojekt RAGTIME; sowie
- eine Datenbank mit Zeitreihenmetadaten zur Dokumentation internationaler Meteorologie-Stationen, entwickelt von der Abteilung Climate System.

Wie erwartet, entstand in der Folge schnell der Wunsch, die an den Referenzdatenbanken demonstrierten Möglichkeiten auch für weitere Datenräume mit Zeitreihenmetadaten des Institutes zu nutzen. Die Adaptierbarkeit der Schnittstelle unterstützte die Einbindung entsprechender weiterer Datenbanken durch Änderungen ihrer Konfiguration, so dass diese Erweiterungen der Datenschicht jeweils unaufwendig durchgeführt werden konnten. Zusätzlich wuchs die Datenschicht durch ein stetes Hinzukommen neuer Zeitreihenmetadaten

[271] Unter dem Namen AFRI (*Advanced Flexible Retrieval Interface*).

in den bereits angegliederten Datenräumen an; diese neuen Daten standen den Anwendern jeweils automatisch über die Schnittstelle zur Verfügung. Zu Ende von Betriebsphase I war den Datennutzern damit eine große Bandbreite unterschiedlicher Metadaten des Institutes zugänglich (vgl. Abb. 18.2):

- verschiedene thematische Ausschnitte der allgemeinen Metadatenbank CERA-2, entwickelt von der Scientific Data Management Group;
- eine Datenbank mit Zeitreihenmetadaten zur Dokumentation deutscher Stationen, entwickelt vom PIK-Kernprojekt RAGTIME;
- eine Datenbank mit Zeitreihenmetadaten zur Dokumentation internationaler Meteorologie-Stationen, entwickelt von der Abteilung Climate System;
- eine Datenbank mit Zeitreihenmetadaten zur Dokumentation nationaler hydrologischer Stationen, entwickelt von der Abteilung Climate System;
- eine Datenbank mit Zeitreihenmetadaten zur Dokumentation nationaler phänologischer Stationen, entwickelt von der Abteilung Global Change and Natural Systems;
- eine Datenbank mit Zeitreihenmetadaten zur Dokumentation von Stationen des Deutschen Wetterdienstes (DWD), entwickelt von der Scientific Data Management Group; sowie
- eine Datenbank mit Zeitreihenmetadaten zur Dokumentation von Stationen zur Messung von Wasserqualitäts-Parametern, entwickelt vom PIK-Kernprojekt RAGTIME.

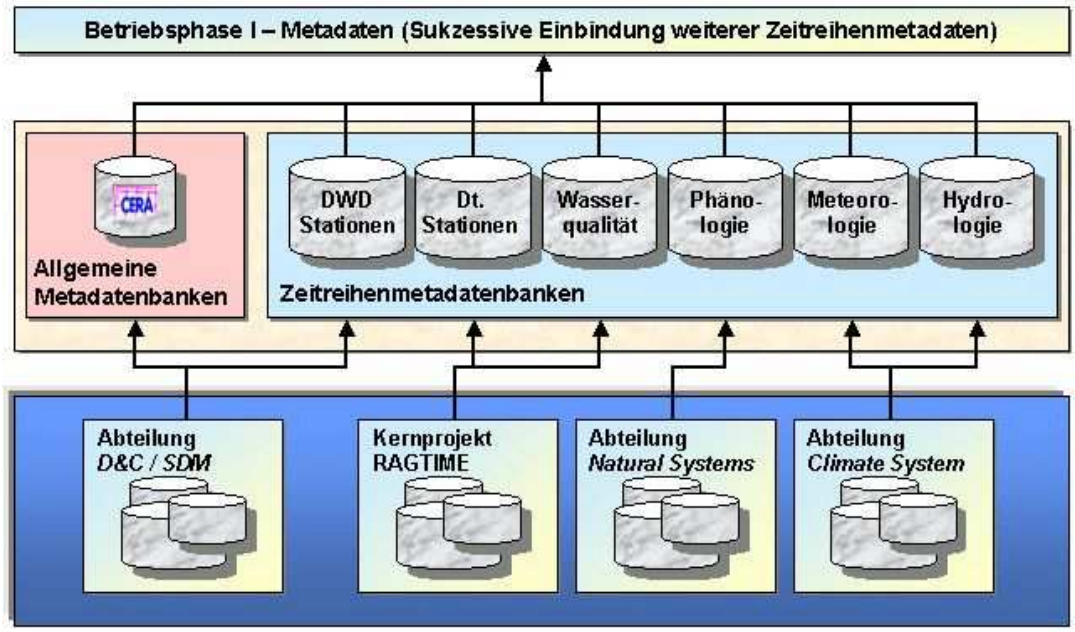

Abb. 18.2 - Erweiterung der zugänglichen Datengrundlage durch sukzessive Einbindung weiterer Zeitreihenmetadatenbanken.

Nachdem auf diese Weise die überwiegende Mehrheit der Zeitreihenmetadaten des Institutes über die Schnittstelle zugänglich gemacht werden konnte, bestand großes Interesse daran, die erreichte Funktionalität auszuweiten und sowohl eine integrierte Auswertung bisher getrennt zugänglicher Zeitreihenmetadaten aus unterschiedlichen wissenschaftlichen Bereichen sowie insbesondere einen direkten Durchgriff von Zeitreihenmetadaten auf die im Institut bereitgestellten punktverorteten Zeitreihen zu ermöglichen.

18.1.2 Systemexterne Integrationsprozesse

Im dritten Zyklus wurden basierend auf den beim bisherigen Betrieb der Schnittstelle gewonnenen Erkenntnissen von der Scientific Data Management Group wesentliche Prozesse der Homogenisierung und Integration sowohl von Zeitreihenmetadaten wie von punktverorteten Zeitreihen durchgeführt, die die Basis für den in Betriebsphase II bereitge-

stellten umfassenden Funktionsumfang bilden. Dabei lassen sich zwei wichtige Schritte gegeneinander abgrenzen – die Zusammenfassung der am Institut verfügbaren punktverorteten Zeitreihen mit Data Warehouse-Methoden sowie der Aufbau von integrierten Zeitreihenmetadatenbanken zur Homogenisierung dieser Datenressourcen.

- **Integrierte Zeitreihenbereitstellung mit Data Warehouse-Methoden**

Um die integrierte Nutzung der am Institut in verschiedenen heterogenen Datenbanken vorgehaltenen punktverorteten Zeitreihen zu ermöglichen, wurde von der Scientific Data Management Group mit Data Warehouse-Methoden eine Datenbank zur integrierten Bereitstellung der punktverorteten Zeitreihen des Institutes aufgebaut (im folgenden kurz als *Data Warehouse* bezeichnet). Die einzelnen, getrennt entwickelten und betriebenen Datenbanken bleiben dabei bestehen, so dass für diese die jeweils vorhandenen Applikationen unverändert weiterverwendet werden können. Die Zeitreihen werden in einem dreistufigen ETL-Prozess (für *Extraction-Transformation-Loading,* vgl. Kap. 2.2.4) in das Data Warehouse überführt und in eine einheitliche Repräsentation umgewandelt; dabei werden die erforderlichen Konsistenzprüfungen durchgeführt sowie bestehende Heterogenitäten beseitigt. Sämtliche so integrierten Zeitreihen werden über das Data Warehouse in einheitlicher Form bereitgestellt; zusätzlich werden aus den Basisdaten weitere Werte durch zeitliche Aggregationen der Messwerte gewonnen und für einen schnellen Zugriff ebenfalls vorgehalten.

- **Aufbau integrierter Zeitreihenmetadatenbanken**

Sowohl zur Unterstützung einer erweiterten Erschließbarkeit der im Institut verfügbaren Zeitreihenmetadaten wie im Hinblick auf einen flexiblen Zugriff auf die unterliegenden Zeitreihen wurden von der Scientific Data Management Group zusätzlich zuvor getrennt bereitgestellte Zeitreihenmetadaten zusammengeführt. Die so entstandenen Zeitreihenmetadatenbanken vereinen nun in homogenisierter Form Metadaten zur Dokumentation von Stationen unterschiedlicher Wissenschaftsgebiete und erlauben somit eine integrierte Nutzung und Auswertung dieser Informationen.

18.1.3 Betriebsphase II – Metadaten und Zeitreihen

Betriebsphase II der Schnittstelle (vierter und letzter vorgesehener Entwicklungszyklus) begann - unter dem neuen Namen xDat[272] - im Juni 2001. Die Metadaten des Institutes werden dem Anwender nun in drei Datenbankgruppen zugänglich gemacht, die die eingebundenen Datenräume in allgemeine Metadaten, Zeitreihenmetadaten sowie Zeitreihenmetadaten mit Möglichkeit zum direkten Zeitreihenzugriff aufteilen (vgl. auch Abb. 18.3):

- **Allgemeine Metadatenbanken**

Die Gruppe der *allgemeinen Metadatenbanken* (*General Meta Databases*) dient zur Aufnahme derjenigen Datenräume, die allgemeine Metadaten bereitstellen. Hierunter fallen (Stand Oktober 2003)

- *PIK CERA-2.* Die anhand der Erfahrungen aus Betriebsphase I aus CERA-2 abgeleitete allgemeine Metadatenbank des Potsdam-Instituts für Klimafolgenforschung (vgl. Kap. 18.2);
- (geplant) *DKRZ CERA-2.* In naher Zukunft soll zusätzlich die CERA-2 Metadatenbank des DKRZ / MPIfM (Deutsches Klimarechenzentrum / Max Planck-Institut für Meteorologie) in Hamburg eingebunden werden.

- **Zeitreihenmetadatenbanken**

Der Gruppe der *Zeitreihenmetadatenbanken* (*Time Series Meta Databases*) sind Daten-

[272] xDat steht für *eXtensible Database Access Tool*.

räume zugeordnet, die allgemeine Zeitreihenmetadaten enthalten. Zum integrierten Vergleich sind hier sowohl Stationen dokumentiert, deren Messwerte bei externen Datengebern anzufordern sind, wie solche Stationen, deren Messwerte bereits im PIK vorgehalten werden. Gegenwärtig sind in diese Gruppe drei Datenräume eingebunden (Stand Oktober 2003):

- *PIK German Measurement Network Meta Database.* Die deutsche Messnetz-Metadatenbank des PIK zur Dokumentation deutscher Stationen aus den Bereichen Meteorologie, Hydrologie, Wasserqualität sowie Phänologie.
- *PIK Global Measurement Network Meta Database.* Die globale Messnetz-Metadatenbank des PIK zur Dokumentation internationaler Stationen aus den Bereichen Meteorologie und Hydrologie.
- *DWD Measurement Network Meta Database.* Die DWD-Messnetz-Metadatenbank des PIK zur Dokumentation von Stationen des Deutschen Wetterdienstes (DWD) aus den Bereichen Meteorologie und Phänologie.

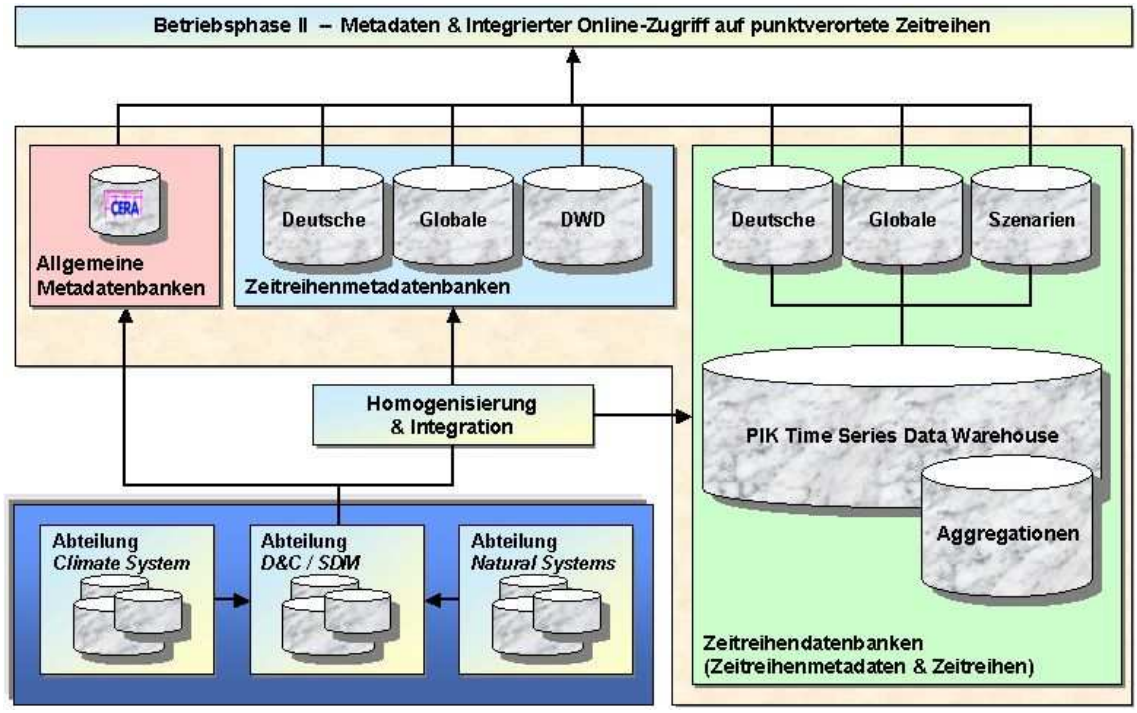

Abb. 18.3 - Zugängliche Datenräume in Betriebsphase II.

- **Zeitreihendatenbanken**

Der Gruppe der *Zeitreihendatenbanken* (*Time Series Databases*) sind hingegen Datenräume zugeordnet, die Zeitreihenmetadaten zur Dokumentation derjenigen Stationen enthalten, deren Zeitreihen über die Schnittstelle interaktiv anhand der Metadaten aus dem Data Warehouse selektiert und online bereitgestellt werden können. Der Begriff *Zeitreihendatenbank* steht damit sowohl für Zeitreihenmetadaten wie zugehörige Zeitreihen; er wurde in Abstimmung mit der Scientific Data Management Group gewählt, um den Datennutzern eine intuitive Abgrenzung gegen reine Zeitreihenmetadatenbanken zu ermöglichen. Gegenwärtig sind in diese Datenbankgruppe ebenfalls drei Datenräume eingebunden (Stand Oktober 2003):

- *PIK German Time Series Database.* Die deutsche Zeitreihendatenbank des PIK zum Zugriff auf Zeitreihen von Stationen aus den Bereichen Meteorologie, Hydrologie, Wasserqualität und Phänologie.

- *PIK Global Time Series Database.* Die globale Zeitreihendatenbank des PIK zum Zugriff auf Zeitreihen nationaler und internationaler Stationen aus den Bereichen Meteorologie und Hydrologie.
- *PIK Scenario Time Series Database.* Die Szenarien-Zeitreihendatenbank des PIK zum Zugriff auf am Institut konstruierte Szenarien-Zeitreihen für die Jahre 2001 bis 2055 für deutsche Stationen aus dem Bereich Meteorologie sowie auf Referenzzeitreihen für diese Stationen für die Jahre von 1951 bis 2000.

Der zuletzt hinzugefügte neue Datenraum ist die PIK Scenario Time Series Database, die in Betriebsphase II im laufenden Betrieb (Feb. 2002) eingebunden wurde. Neu ins Institut geholte Zeitreihenmetadaten werden seither über die bereits eingebundenen Zeitreihenmetadatenbanken und Zeitreihendatenbanken, punktverortete Zeitreihen über das Data Warehouse verfügbar gemacht. Entsprechend ist in Betriebsphase II ein beständiges Anwachsen der über die Schnittstelle verfügbaren Daten primär durch *Erweiterungen der bereits eingebundenen Datenräume* zu beobachten.

18.2 Allgemeine Metadaten

18.2.1 Von CERA-2 zu PIK CERA-2

Es ist offensichtlich, dass eine so komplexe Struktur, wie sie das CERA-2-Datenbankmodell darstellt, entsprechend aufbereitet werden muss, um Anwendern ein möglichst effizientes und intuitives Auffinden der darin enthaltenen Metadaten zu ermöglichen. Um gemeinsam mit Datennutzern und dem CERA-Administrator des Institutes zu einer geeigneten Lösung zu gelangen, wurde ein zweistufiges Vorgehen gewählt. In Betriebsphase I wurde CERA-2 den Anwendern anhand verschiedener Teilausschnitte präsentiert, die aus einer Übersichtstabelle mit einer vom CERA-Administrator vorgeschlagenen Auswahl von Attributen, einzelnen Blöcken des CERA-Kerns sowie weiteren am Institut verwendeten CERA-Modulen bestanden. Dabei konnte jeder dieser Teilausschnitte als Ausgangsdatenraum für eine Anfrage ausgewählt werden; aus den Ergebnissen einer Anfrage konnten anschließend Metadatensätze selektiert werden, um für diese die zugehörigen Einträge aus einem der anderen CERA-Ausschnitte abzurufen. Die Ergebnisse für jeweils zwei Teilausschnitte wurden gleichzeitig dargestellt; der Prozess der Auswahl von Einträgen und des Abrufens der entsprechenden Informationen aus einem weiteren Ausschnitt konnte beliebig ausgeführt werden, so dass ein flexibles Navigieren zwischen Übersichtstabelle sowie Blöcken und Modulen von CERA-2 möglich war.

Dabei wurde deutlich, dass eine Navigation über gleichrangig präsentierten Ausschnitten der Metadatenbank insbesondere für Anwender, die mit der Bedeutung der einzelnen Datenbankausschnitte nicht vertraut waren, nicht intuitiv genug war. So gestaltete es sich schwierig, aus den Namen der angebotenen Teilausschnitte geeignete Rückschlüsse über die in ihnen enthaltenen Informationen abzuleiten; zudem erwarteten Anwender zumeist eine möglichst direkte Darstellung *aller* gesuchten Informationen nach Formulierung einer Anfrage und kamen oft gar nicht auf die Idee, ein Anfrageergebnis durch die Abfrage zugehöriger Informationen aus weiteren Ausschnitten zu verfeinern.

Aufbauend auf den beim Betrieb gewonnenen Erfahrungen wurde von der Scientific Data Management Group zu Beginn von Betriebsphase II aus CERA-2 die allgemeine Institutsmetadatenbank PIK CERA-2 abgeleitet. Basierend auf den in dieser ersten Betriebsphase aus Rückmeldungen der Anwender als besonders relevant identifizierten Informationen wurde eine Untermenge von Attributen aus CERA-2 ausgewählt. Hierauf aufbauend wurde das Datenbankmodell für die allgemeine Institutsmetadatenbank PIK CERA-2 entwickelt, die - kompatibel mit CERA-2, jedoch weniger komplex - seit April 2003 über ein von der

Scientific Data Management Group entwickeltes Eingabewerkzeug (vgl. Kap. 25.1) von den Wissenschaftlern des Instituts autonom befüllt werden kann.

Bezeichnung[273]	**Bedeutung**	**Selektion durch...**
Hierarchischer Thesaurus (Hierarchical Thesaurus)	Zuordnung jedes Eintrages zu thematischen Gebieten und Untergebieten anhand entsprechender Schlüsselbegriffe	Auswahl von Schlüsselbegriffen aus der Hierarchie
Identifikator (Entry ID)	Eine Kennzahl zur eindeutigen Identifikation jedes Metadateneintrages	Auswahl von Identifikatoren
Titel (Entry Title)	Eine kurze und eindeutige Beschreibung jedes Metadateneintrages	Auswahl von Titeln
Horizontale räumliche Abdeckung (North, South, East, West)	Die räumliche Abdeckung der dokumentierten Daten anhand einer Boundingbox (Angabe von minimaler sowie maximaler geographischer Breite und Länge)	Auswahl einer Boundingbox (vordefiniert oder frei definierbar)
Räumliche Auflösung[274] (Spatial Resolution)	Die kleinste räumliche Auflösung von Objekten, die von den dokumentierten Daten repräsentiert werden	Auswahl von räumlichen Auflösungen
Zeitliche Abdeckung (Begin Year, End Year)	Die zeitliche Abdeckung der dokumentierten Daten anhand des ersten sowie letzten Jahres der von diesen abgedeckten Zeitspanne	Auswahl einer zeitlichen Abdeckung
Zeitliche Auflösung[275] (Temporal Resolution)	Die zeitliche Auflösung der dokumentierten Daten	Auswahl von zeitlichen Auflösungen
Zugeordnetes Projekt (PIK Project)	Die Zuordnung jedes dokumentierten Datensatzes zu einem konkreten Projekt des Institutes	Auswahl von Projekten
Kontaktperson (PIK Contact)	Die jeweiligen Ansprechpartner im Institut für die dokumentierten Daten	Auswahl von Kontaktpersonen.
Verwendbarkeit (Access)	Die Klassifikation der dokumentierten Daten anhand der jeweils geltenden Restriktionen[276] für ihre Verwendung	Auswahl von Formen der Verwendbarkeit
Datengebende Institution (Originator)	Die Einrichtung, von der die dokumentierten Daten ursprünglich bereitgestellt wurden	Auswahl von datengebenden Institutionen
Speichermedium (Media)	Die Klassifikation der dokumentierten Daten anhand des zu ihrer physikalischen Speicherung verwendeten Mediums[277]	Auswahl von Speichermedien
Datenformat (Format)	Die Klassifikation der dokumentierten Daten anhand des jeweils verwendeten Datenformates[278]	Auswahl von Datenformaten
physikalische Einheit[279] (Unit)	Die Zuordnung einer physikalischen Einheit zu den dokumentierten Daten	Auswahl von physikalischen Einheiten

Tab. 18.1 - Selektionskriterien zur Auswahl von allgemeinen Metadaten aus PIK CERA-2.

18.2.2 Entitäten zur Selektion und Präsentation

Nachfolgend werden kurz die Entitäten von PIK CERA-2 beschrieben, die im Laufe der iterativen Entwicklung der Schnittstelle als relevant für einen effizienten Zugang zu den

[273] In Klammern die den Anwendern über die Schnittstelle präsentierten Bezeichnungen.
[274] Wird nur verwendet, wenn die dokumentierten Daten in einer spezifischen räumlichen Auflösung vorliegen.
[275] Wird nur verwendet, wenn die dokumentierten Daten einen Zeitbezug besitzen.
[276] Mögliche Klassifikationen sind bspw.: frei verwendbar, frei verwendbar innerhalb des zuständigen Projektes, institutsweit verwendbar, extern verwendbar innerhalb von Forschungskooperationen.
[277] Hier wird beispielsweise nach Datenhaltung in Datenbanken, Tape Library, Dateisystemen, CD-ROMs etc. unterschieden.
[278] Hierunter fallen bspw. Datenbankmodelle, Datenformate von Geoinformationssystemen wie ARC, selbstbeschreibende Dateiformate wie netCDF u.v.a.m.
[279] Wird nur verwendet, wenn für die dokumentierten Daten eindeutig eine entsprechende physikalische Einheit festgelegt werden kann.

allgemeinen Metadaten des Institutes identifiziert wurden. Zum Zugriff auf PIK CERA-2 wurde ein Satz aus rund 20 von den Anwendern als besonders relevant erachteten Attributen als zentraler Ausschnitt festgelegt. Dieser Ausschnitt bildet nun für die Anwender die zentrale Einsprungstelle, von der aus sie zu ihrem Anfrageergebnis gelangen. Tab. 18.1 gibt einen Überblick über die Entitäten von PIK CERA-2, die dem Anwender von der Schnittstelle als Selektionskriterien zur Auswahl allgemeiner Metadaten bereitgestellt werden. Sämtliche Selektionskriterien können *beliebig miteinander kombiniert* werden; zusätzlich können Datensätze anhand frei definierbarer, *kriterienübergreifender Suchbegriffe* selektiert werden[280]. In diesem Fall werden alle Metadateneinträge ausgewählt, in denen der eingegebene Suchbegriff in mindestens einer der textuellen Werteausprägungen eines Attributes enthalten ist.

Bezeichnung[281]	**Bedeutung**	**Bereitstellung**
Beschreibung (Entry Summary)	Eine ausführlichere textuelle Beschreibung der dokumentierten Daten	automatisch
Bezeichnung der horizontalen räumlichen Abdeckung (Location)	Eine zusätzliche textuelle Beschreibung der horizontalen räumlichen Abdeckung der dokumentierten Daten (Namen von Orten, von Regionen o.ä.)	automatisch
Vertikale räumliche Abdeckung (Min Alt, Max Alt)	Die vertikale räumliche Abdeckung der dokumentierten Daten anhand minimaler sowie maximaler Tiefe / Höhe in Metern	automatisch
Speicherbedarf (Size)	Der von den dokumentierten Daten belegte Speicherplatz in Byte[282]	automatisch
Zugriffsdetails (variable Bezeichnung)[283]	Detailinformationen für den Zugriff auf die Daten. Die Art dieser Informationen variiert in Abhängigkeit vom jeweiligen Speichermedium[284]	automatisch
Details über Kontaktpersonen	Detailinformationen für eine individuelle Kontaktaufnahme mit dem jeweiligen Ansprechpartner für die dokumentierten Daten[285]	interaktiv
Details über datengebende Institutionen	Detailinformationen für eine individuelle Kontaktaufnahme mit der Einrichtung, von der die dokumentierten Daten ursprünglich bereitgestellt wurden[286]	interaktiv
Referenzierungen	Detailinformationen über Publikationen, die im Zusammenhang mit den dokumentierten Daten stehen[287]	interaktiv

Tab. 18.2 - Überblick über die zusätzlich für jeden selektierten Metadatensatz aus PIK CERA-2 automatisch bereitgestellten bzw. interaktiv abrufbaren Informationen.

Für die Auswertung selektierter Metadatensätze werden dem Anwender die jeweiligen Werteausprägungen sämtlicher in Tab. 18.1 aufgeführter Selektionskriterien bereitge-

[280] Vgl. das hierfür entwickelte Filtermodul GlobalSearchFilter (Kap. 20.6).
[281] In Klammern die den Anwendern über die Schnittstelle präsentierten Bezeichnungen.
[282] Diese Informationen werden zur besseren Interpretierbarkeit von der Schnittstelle dynamisch in entsprechende Kilo-, Mega- oder Gigabyte-Werte transformiert (vgl. Kap. 21.1.4).
[283] Zugriffsdetails werden über zwei Attribute von PIK CERA-2 bereitgestellt, für die von der Schnittstelle zur schnellen Orientierung des Anwenders dynamisch je nach Semantik entsprechende Bezeichner generiert werden (vgl. Abb. 21.2).
[284] Bspw. werden bei Datenbanken Datenbankidentifikatoren und -Adressen beschrieben, bei Daten, die in Dateisystemen gehalten werden, hingegen Dateinamen und Verzeichnispfade, bei CDs wiederum ihr Standort (bspw. Institutsbibliothek) sowie ihre Signatur.
[285] Hierunter fallen bspw. Anrede, Titel, Telephonnummer, postalische und E-Mail-Adresse etc.
[286] Hierzu zählen die genaue Bezeichnung der datengebenden Institution, ihre Postanschrift, gegebenenfalls die URL einer Homepage, Name, Telephonnummer und E-Mail-Adresse von Ansprechpartnern etc.
[287] Liegen Informationen über solche Publikationen vor, enthalten die entsprechenden Metadateneinträge in PIK CERA-2 diesbezügliche Angaben wie Titel und Autor(en) der Publikationen, Herausgeber, Verlag, Datum der Veröffentlichung etc.

stellt[288]. Zusätzlich werden weitere Informationen zur Verfügung gestellt, die über die in die Selektionskriterien einbezogenen Attribute hinausgehen. Hierbei ist zwischen zwei Arten der Bereitstellung zu unterscheiden. Zum einen werden für diverse weitere Entitäten automatisch für jeden selektierten Metadateneintrag die jeweiligen Werteausprägungen angezeigt; ferner besteht für den Anwender die Option, bei Bedarf interaktiv spezifische Details für jeden selektierten Metadateneintrag aus PIK CERA-2 abzurufen (vgl. Tab. 18.2).

18.3 Zeitreihenmetadaten

18.3.1 Hierarchische räumliche Klassifikation

Bei einer raumbezogenen Selektion von punktverorteten Datensätzen - also bspw. von Zeitreihenmetadaten zur Beschreibung von Stationen - sind zwei Fälle zu unterscheiden, die jeweils andere Vorgehensweisen erfordern.

- **Selektion durch Koordinatenabgleich**

Eine Selektion anhand einer nutzerdefinierten Boundingbox erfolgt sinnvoll durch Abgleich der durch diese definierten Grenzen mit den geographischen Koordinaten der Datensätze. Um derartige Selektionen zu ermöglichen, genügt eine Georeferenzierung jedes Datensatzes über ein Paar geographischer Koordinaten; eine Vorabzuordnung zu einzelnen Boundingboxen durch entsprechende Kodierungen in den Datenbanken ist weder erforderlich noch aufgrund der potentiell unendlichen Zahl möglicher nutzerdefinierbarer Boundingboxen überhaupt durchführbar.

- **Selektion durch räumliche Klassifikatoren**

Um hingegen eine effiziente Selektion anhand vordefinierter hierarchischer administrativer oder naturräumlicher geographischer Einheiten - Kontinente, Staaten, Bundesländer etc. oder Flusseinzugsgebiete - zu ermöglichen, ist es sinnvoll, entsprechende Zuordnungen jedes Datensatzes vorab vorzunehmen und in den entsprechenden Datenbanken zu kodieren[289]. Um die über den interaktiven digitalen Atlas IDA (vgl. Kap. 17) unterstützte graphisch-interaktive Selektion vordefinierter Einheiten aus unterschiedlichen Hierarchien effizient in raumbezogene Abfragen abbilden zu können, wurden vor Beginn der Betriebsphase I von der Scientific Data Management Group anhand der Referenzdatenbanken *hierarchische räumliche Klassifikatoren* entwickelt bzw. adaptiert, die die gleichzeitige Zuordnung von Datensätzen zu administrativen sowie naturräumlichen Hierarchien erlauben und leicht erweitert werden können.

Diese Klassifikatoren wurden in die von IDA verarbeiteten Kartendaten integriert und können von diesem für jede nutzerdefinierte Selektion vordefinierter Einheiten bereitgestellt und von einem entsprechenden Filtermodul[290] in eine raumbezogene Selektion umgesetzt werden. Die in die Schnittstelle eingebundenen Datenräume mit Zeitreihenmetadaten wurden während der Betriebsphasen I und II sukzessive um entsprechende Klassifikatoren erweitert. Nachfolgend werden zunächst die Klassifikatoren für Zuordnungen zu administrativen Einheiten sowie zu Flusseinzugsgebieten vorgestellt; im Anschluss wird die Abbildung einer nutzerdefinierten Raumauswahl zur Selektion entsprechend klassifizierter Datensätze beschrieben.

- **Administrative Klassifikation**

Der Schlüssel für die administrative Klassifikation setzt sich aus verschiedenen Bestand-

[288] Die hierfür entwickelten Auswertungsmodule werden in Kap. 21.1 beschrieben.
[289] Ein Abgleich der geographischen Koordinaten von Datensätzen mit den Umrissen geographischer Einheiten für jede Selektion ist rechenaufwendiger als die Auswertung vordefinierter Zuordnungen.
[290] Das realisierte Filtermodul zur graphisch-interaktiven Raumauswahl über IDA ist der SpatialFilter (vgl. Kap. 20.3).

teilen zusammen. Er besteht aus einem festen Präfix (Element PX_A) zur Abgrenzung gegen andere Hierarchien, Elementen des FIPS-Code[291] zur Unterscheidung von Kontinenten (Element KO) und Staaten (Element ST) sowie - für deutsche Stationen - weiteren Teilschlüsseln zur Unterscheidung von Bundesländern (Element BL) und Landkreisen (Element LK). Der Schlüssel wird nach folgender Regel gebildet:

$$PX_A + KO + ST [+ BL] [+ LK]$$

Tab. 18.3 gibt einen Überblick über die Bedeutung der einzelnen Schlüsselbestandteile und führt Beispiele für die Schlüsselbildung an. So setzt sich der Schlüssel ADSABR für den Staat Brasilien aus den Elementen AD (Präfix administrative Klassifikation) + SA (Kontinent Südamerika) + BR (Staat Brasilien) zusammen. Entsprechend wird der Schlüssel für den Landkreis Fulda (ADEUGM05631) gebildet aus den Elementen AD (Präfix administrative Klassifikation) + EU (Kontinent Europa) + GM (Staat Deutschland) + 05 (Bundesland Hessen) + 631 (Landkreis Fulda).

Schlüsselbestandteile			Beispiele für die Schlüsselbildung		
Element	**Bedeutung**	**Kodierung**	**Element**	**Bedeutung**	**Vollständiger Schlüssel**
PX_A	Präfix Administrative Klassifikation	AD (fix)			
KO	Kontinent (FIPS)	2 Zeichen	EU	Europa	ADEU
			SA	Südamerika	ADSA
			AF	Afrika	ADAF
ST	Staat (FIPS)	2 Zeichen	GM	Deutschland	ADEUGM
			BR	Brasilien	ADSABR
			CM	Kamerun	ADAFCM
BL [292]	Bundesland	2 Ziffern	05	Hessen	ADEUGM05
			11	Brandenburg	ADEUGM11
LK [293]	Landkreis	3 Ziffern	631	Fulda	ADEUGM05631
			054	Potsdam-Stadt	ADEUGM11054

Tab. 18.3 - Hierarchische administrative Klassifikation.

Der hierarchische Aufbau der Schlüssel spiegelt dabei die hierarchische Struktur administrativer Einheiten wider und enthält alle Informationen, um einen so referenzierten Datensatz der vordefinierten Einheit mit dem höchsten Detailgrad und zugleich allen übergeordneten Einheiten zuordnen zu können – so kann beispielsweise für einen Datensatz, der dem Landkreis Fulda zugeordnet wurde, zugleich geschlossen werden, dass dieser ferner dem Bundesland Hessen, dem Staat Deutschland sowie dem Kontinent Europa zugeordnet ist.

- **Klassifikation nach Flusseinzugsgebieten**

Ein ähnliches Prinzip der Beschreibung liegt auch dem Schlüssel für die Zuordnung zu Flusseinzugsgebieten zugrunde. Er setzt sich aus einem festen Präfix (Element PX_F) zur Abgrenzung gegen andere Hierarchien sowie Teilschlüsseln zur Kodierung von Kontinenten bzw. Ozeanbecken[294] (Element KO), Haupteinzugsgebieten bzw. Hauptbecken (Element H) sowie Teileinzugsgebieten bzw. Teilbecken[295] erster und zweiter Ordnung (Elemente T_1 und T_2) zusammen. Der Schlüssel wird nach folgender Regel gebildet:

[291] Federal Information Processing Standards (http://www.itl.nist.gov/fipspubs/index.htm).
[292] Wird nur für deutsche Stationen verwendet.
[293] Wird nur für deutsche Stationen verwendet.
[294] Die Zuordnung zu Ozeanbecken erfolgt, um auch solche Stationen in die Hierarchie einordnen zu können, die auf Inseln liegen, deren Umrisse aufgrund geringer Größe bei der Generalisierung der Vektordaten für IDA (vgl. Kap. 17.3.5) nicht berücksichtigt werden.
[295] Ozeanbecken werden gegenwärtig nur in Hauptbecken unterteilt.

$$PX_F + KO + H [+ T_1] [+ T_2]$$

Wie beim Schlüssel zur administrativen Klassifikation wird eine Zuordnung durch Kombination von Werten der einzelnen Schlüsselelemente durchgeführt. Tab. 18.4 beschreibt die Bedeutung der einzelnen Bestandteile und gibt Beispiele für die Schlüsselbildung. So setzt sich der der Schlüssel für das Flusseinzugsgebiet der Nuthe (BAEU00584), eines Nebenflusses der Havel, aus dem Elementen BA (Präfix Flusseinzugsgebiets-Hierarchie) + EU (Kontinent Europa) + 005 (Haupteinzugsgebiet Elbe) + 8 (Teileinzugsgebiet Level 1 Havel) + 4 (Teileinzugsgebiet Level 2 Nuthe) zusammen. Auch hier enthält der Schlüssel alle Informationen, um einen entsprechend klassifizierten Datensatz beispielsweise dem Subsystem Nuthe und zugleich den übergeordneten Flusseinzugssystemen Havel und Elbe zuordnen zu können.

Schlüsselbestandteile			**Beispiele für die Schlüsselbildung**		
Element	**Bedeutung**	**Kodierung**	**Element**	**Bedeutung**	**Vollständiger Schlüssel**
PX_F	Präfix Flusseinzugsgebiets-Hierarchie	BA (fix)			
KO	Kontinent bzw. Ozeanbecken	2 Zeichen	EU	Europa	BAEU
			SA	Südamerika	BASA
			PA	Pazifik	BAPA
H	Haupteinzugsgebiet bzw. Hauptbecken	3 Ziffern	005	Elbe	BAEU005
			016	Amazonas	BASA016
			003	Südpazifik	BAPA003
T_1 [296]	Teileinzugsgebiet bzw. Teilbecken Level 1	1 Ziffer	4	Mulde (Dtld.)	BAEU0054
			8	Havel	BAEU0058
T_2 [297]	Teileinzugsgebiet bzw. Teilbecken Level 2	1 Ziffer	4	Nuthe	BAEU00584
			9	Dosse	BAEU00589

Tab. 18.4 - Hierarchische Klassifikation nach Flusseinzugsgebieten.

- **Selektion durch Musterabgleich**

In den einzelnen Datenbanken des PIK werden Zeitreihenmetadaten anhand der beschriebenen Klassifikatoren jeweils bestimmten administrativen Einheiten und Flusseinzugsgebieten zugeordnet. Jeder Klassifikator wird dabei als Zeichenkette in einem entsprechenden Attribut abgelegt. Eine entsprechende Selektion dieser Daten kann durch Musterabgleich erfolgen und dabei von der hierarchischen Struktur der Klassifikatoren profitieren. Dabei kann die Tatsache ausgenutzt werden, dass sich die Zuordnung über eine entsprechende Zeichenkette von links nach rechts verfeinert.

Selektion von Stationen durch Gebietsauswahl (selektiert = ✓)					
Stationen und vorab zugeordnete Schlüssel		**Selektierte Gebiete und generierte Vergleichsmuster**			
		Europa	Deutschland	Brandenburg	Potsdam-Stadt
Name	**Schlüssel**	ADEU	ADEUGM	ADEUGM11	ADEUGM11054
Neapel	ADEUIT	✓			
Neuruppin	ADEUGM09041	✓	✓		
Brandenburg	ADEUGM11068	✓	✓	✓	
Potsdam	ADEUGM11054	✓	✓	✓	✓

Tab. 18.5 - Beispiele für raumbezogene Selektion durch Musterabgleich.

Werden beispielsweise diejenigen Zeichenketten selektiert, die mit der Zeichenfolge ADEU

[296] Eine Zuordnung zu Teileinzugsgebieten erster Ordnung findet zur Zeit nur für deutsche Stationen statt; *Teilbecken* erster Ordnung finden gegenwärtig keine Verwendung.
[297] Eine Zuordnung zu Teileinzugsgebieten erster Ordnung findet zur Zeit nur für deutsche Stationen statt; *Teilbecken* erster Ordnung finden gegenwärtig keine Verwendung.

beginnen, so trifft dieses Muster auf die administrativen Schlüssel aller Datensätze zu, die Europa oder einer beliebigen Subeinheit von Europa - Ländern, Landkreisen usw. - zugeordnet sind. Werden hingegen solche Zeichenketten selektiert, die mit der Zeichenfolge ADEUGM beginnen, führt dies entsprechend zur Auswahl derjenigen Datensätze, die Deutschland oder dessen Subeinheiten zugeordnet sind. Da dieser Musterabgleich - stets von links beginnend - auf unterschiedlichen Stufen der Hierarchien stattfinden kann, steht so ein flexibler Mechanismus zur Verfügung, um für ein beliebiges vordefiniertes Gebiet die diesem zugeordneten Datensätze zu identifizieren. Tab. 18.5 veranschaulicht dieses Prinzip anhand unterschiedlicher Abfragemuster und den jeweils durch diese selektierten Datensätzen.

Eine Generierung geeigneter Abfragemuster durch ein Filtermodul für raumbezogene Teilbedingungen setzt voraus, dass dieses für beliebige nutzerselektierbare vordefinierte Gebiete auf die jeweiligen Klassifikatoren zurückgreifen kann. Um dies zu ermöglichen, enthalten die von IDA verwendeten Kartendaten für jede geographische Einheit jeweils genau den Klassifikator, der als Muster für eine entsprechende raumbezogene Selektion verwendet werden kann. Hat der Anwender in IDA ein vordefiniertes Gebiet in einer beliebigen Kartenhierarchie ausgewählt, stellt die Komponente den entsprechenden Klassifikator zur Verfügung. Das für die graphisch-interaktive Raumauswahl konzipierte Filtermodul (vgl. Kap. 20.3, SpatialFilter) erhält so die erforderlichen Schlüsselwerte und kann auf diese Weise eine in IDA getroffene Gebietsauswahl jeweils in eine entsprechende Teilbedingung für die Datenbankanfrage umsetzen.

18.3.2 Entitäten zur Selektion und Präsentation

In diesem Kapitel werden die Entitäten beschrieben, die im Laufe der iterativen Entwicklung der Schnittstelle als relevant für einen effizienten Zugang zu den in Zeitreihenmetadatenbanken und Zeitreihendatenbanken des Institutes dokumentierten Stationen identifiziert wurden. Dabei ist zu beachten, dass nicht alle eingebundenen Datenräume den gleichen Umfang an Attributen zur Dokumentation von Stationen bereitstellen, so dass die beschriebenen Entitäten in den einzelnen Datenräumen in unterschiedlicher Zusammensetzung auftreten können. Die einzelnen Selektionskriterien können folgendermaßen aufgeteilt werden:

Klassifikationskriterien
- Klassifikationskriterien beschreiben individuelle Stationen anhand von Eigenschaften bezüglich Beobachtungsaufgabe und Datenverfügbarkeit. Hierunter fallen die übergreifend verwendete Klassifizierung anhand von *Stationstypen* sowie weitere, jeweils *spezifische Klassifikationskriterien* für Zeitreihenmetadatenbanken und Zeitreihendatenbanken.

Weitere Kriterien
- Weitere Selektionskriterien umfassen die *zeitliche Abdeckung*, *Georeferenzierung*, die *direkte Referenzierung* individueller Stationen sowie die Selektion von Stationen aus Zeitreihendatenbanken anhand *statistischer Eigenschaften* der von ihnen erhobenen Messwerte.

Die verschiedenen Selektionskriterien für Zeitreihenmetadatenbanken und Zeitreihendatenbanken werden nachfolgend vorgestellt und daran anschließend überblicksartig zusammengefasst (vgl. Tab. 18.12).

- **Übergreifende Klassifizierung - Stationstypen**

Jeder Station wird, unabhängig davon, ob sie in einer Zeitreihenmetadatenbank oder einer Zeitreihendatenbank dokumentiert ist, anhand ihrer jeweiligen Beobachtungsaufgabe ein bestimmter *Stationstyp* (Attribut Stat_Type) zugeordnet. Gegenwärtig werden vier unter-

schiedliche Stationstypen zur Klassifizierung der am PIK dokumentierten und über die Schnittstelle zugänglichen Stationen verwendet (vgl. Tab. 18.6).

Bezeichnung	Bedeutung
meteorology	Stationen zur Messung meteorologischer Größen
hydrology	Stationen zur Messung hydrologischer Größen
water quality	Stationen zur Messung des Anteils spezifischer Stoffe im Wasser
phenology	Stationen zur Messung des Auftretens wiederkehrender natürlicher Phänomene, bspw. der zeitlichen Abfolge und des Eintretens pflanzlicher Entwicklungsphasen

Tab. 18.6 - Übersicht über die in den eingebundenen Zeitreihenmetadatenbanken und Zeitreihendatenbanken dokumentierten Stationstypen (Stand Oktober 2003).

Die in Zeitreihenmetadatenbanken dokumentierten Stationen können ferner anhand zweier weiterer Kriterien klassifiziert werden. Hierunter fallen eine - nicht obligatorische - zusätzliche Zuordnung zu Subtypen sowie Angaben über den Status der jeweiligen Datenverfügbarkeit.

- **Subklassifizierung in Zeitreihenmetadatenbanken I – Subtypen**

Eine genauere Klassifizierung individueller Stationen kann durch die Verwendung von *Subtypen* (Attribut Stat_Subtype) erfolgen, die einzelne Stationen zusätzlich Unterklassen zuordnen; so können bspw. bei phänologischen Stationen die Subtypen Agrarwirtschaft, Forstwirtschaft, Obstanbau, natürliche Vegetation etc. Verwendung finden, je nachdem, welchem Bereich die beobachteten Spezies zuzuordnen sind. Eine solche Unterteilung findet sich gegenwärtig in der DWD Measurement Network Meta Database, die die Stationstypen *meteorology und phenology sowie* neun bzw. fünf Subtypen zur Klassifizierung der in ihr dokumentierten Stationen verwendet. Die PIK German Measurement Network Meta Database und die PIK Global Measurement Network Meta Database verwenden hingegen keine weitere Untergliederung; in diesen Fällen beschränkt sich die typbezogene Klassifizierung der Stationen jeweils auf Stationstypen.

- **Subklassifizierungen in Zeitreihenmetadatenbanken II – Datenverfügbarkeit**

Der *Verfügbarkeitsstatus* (Attribut PIK_Availability) einer Station gibt an, in welcher Weise der Anwender auf die an einer Station erhobenen Daten zugreifen kann. Dabei sind vier Formen der Datenverfügbarkeit zu unterscheiden (vgl. Tab. 18.7).

Bezeichnung	Bedeutung
online	Zeitreihen dieser Station können mit der Schnittstelle online über die Zeitreihendatenbanken abgerufen werden
digital	Zeitreihen dieser Station liegen dem Institut zwar in digitaler Form vor, können jedoch (noch) nicht online über die Zeitreihendatenbanken abgerufen werden [298]
paper copies	Zeitreihen dieser Station liegen im Institut in (noch nicht digitalisierter) Papierform vor
no	Zeitreihen dieser Station sind am Institut nicht direkt verfügbar; der Anwender wendet sich in diesem Fall direkt an den entsprechenden Datengeber

Tab. 18.7 - Übersicht über die in den eingebundenen Zeitreihenmetadatenbanken dokumentierten Formen der Datenverfügbarkeit (Stand Oktober 2003).

Während diese Information in Zeitreihendatenbanken, für die ein direkter Online-Durchgriff auf die Messdaten bereitgestellt wird, entfallen kann, gibt sie in den Zeitreihenmetadatenbanken, die Stationen mit unterschiedlichen Verfügbarkeitskriterien dokumentieren, wichtige Hinweise auf die erforderlichen Schritte zum Datenzugriff.

[298] Dabei kann es sich um Zeitreihen handeln, die auf CDs oder im Dateisystem des Institutes gehalten werden, oder solche, die über eine PIK-Kontaktperson via Internet bei Datengebern zugänglich sind.

- **Ausprägungen der Klassifizierungskriterien in den Zeitreihenmetadatenbanken**

Die Klassifizierungskriterien Stationstyp, Subtyp und Datenverfügbarkeit finden sich in den drei gegenwärtig eingebundenen Zeitreihenmetadatenbanken (PIK German Measurement Network Meta Database, PIK Global Measurement Network Meta Database und DWD Measurement Network Meta Database) in jeweils unterschiedlichen Ausprägungen. Tab. 18.8 gibt einen Überblick über die jeweilige Zusammensetzung.

Zeitreihen-meta-datenbank	Stationstypen				Sub-typen	Datenverfügbarkeit			
	hydro-logy	*meteoro-logy*	*pheno-logy*	*water quality*		*online*	*digital*	*paper copies*	*no*
German	✓	✓	✓	✓		✓		✓	✓
Global	✓	✓				✓	✓		✓
DWD		✓	✓		✓[299]	✓			✓

Tab. 18.8 - Ausprägungen von Stationstypen, Subtypen sowie Formen der Datenverfügbarkeit in den eingebundenen Zeitreihenmetadatenbanken (Stand Oktober 2003).

Anders als in den Zeitreihenmetadatenbanken erfolgt in den Zeitreihendatenbanken die Subklassifizierung der dokumentierten Stationen in einer Weise, die es erlaubt, Stationen anhand der ihnen zugeordneten Zeitreihen zu beschreiben. Dies erfolgt über die beiden Entitäten Variable sowie zeitliche Auflösung, durch deren Kombination eine eindeutige Identifikation individueller Zeitreihen einer Station ermöglicht wird.

- **Subklassifizierung in Zeitreihendatenbanken I – Variablen**

Die individuellen Parameter, die an einzelnen Stationen erhoben werden, werden als *Variablen* (Attribut Variable) bezeichnet. Da Stationstypen unterschiedliche Beobachtungsaufgaben von Stationen beschreiben, differieren die jeweiligen Variablen zunächst in Abhängigkeit von den jeweiligen Stationstypen. So erheben meteorologische Stationen typischerweise Werte wie Niederschlag, Luft- und Bodentemperatur, Luftdruck oder Bewölkungsdichte, während hydrologische Stationen Parameter wie Wasserstände und Durchflussmengen messen. Wasserqualitäts-Stationen erheben bspw. den Gehalt von Stoffen wie Ammonium, Nitriten, Nitraten oder Phosphaten im Wasser; phänologische Stationen dokumentieren bspw. für unterschiedliche Pflanzen eine Vielzahl von Phänomenen wie den jeweiligen Zeitpunkt von erster Blüte, Fruchtreife, Blattabfall etc. Dabei variiert die Zahl der einzelnen Variablen zunächst je nach Stationstyp sowie in den individuellen Zeitreihendatenbanken beträchtlich (vgl. Tab. 18.10). Zudem kann nicht davon ausgegangen werden, dass verschiedene Stationen des gleichen Typs jeweils auch den gleichen Satz von Variablen erheben. In der Praxis treten vielmehr häufig sehr unterschiedliche Kombinationen pro Station auf, deren Zusammensetzung überdies im Lauf der Zeit Veränderungen unterliegen kann (vgl. u. zeitliche Abdeckung).

- **Subklassifizierung in Zeitreihendatenbanken II – zeitliche Auflösung**

Messreihen werden jeweils in bestimmten Frequenzen erhoben, die als *zeitliche Auflösungen* (Attribut Temp_Resol) bezeichnet werden. Die zeitliche Auflösung kann in unterschiedlichen Ausprägungen auftreten; die eingebundenen Zeitreihendatenbanken dokumentieren gegenwärtig fünf Messfrequenzen (vgl. Tab. 18.9). Bei der großen Bandbreite der im Institut dokumentierten Stationen können dabei vielfältige Kombinationen aus Variablen und zeitlichen Auflösung sowohl innerhalb einer Zeitreihendatenbank wie für individuelle Stationen vorliegen; so ist es bspw. möglich, dass dieselbe Variable von einer Station in täglicher Auflösung, von einer weiteren Station in monatlicher Auflösung und von einer dritten Sta-

[299] Gegenwärtig werden 9 Subtypen zur Klassifizierung von Meteorologie-Stationen (*MIRIAM/AFMS2 system*, *aerological station*, *climate station*, *precipitation station*, *soil temp. station*, *sunshine station*, *synoptic station*, *temp./humidity station* und *wind station*) sowie 5 Subtypen zur Klassifizierung von Phänologie-Stationen (*agriculture*, *horticulture*, *nat. vegetation*, *silviculture* und *unknown*) verwendet.

tion sowohl in täglicher wie monatlicher Auflösung erhoben bzw. bereitgestellt wird.

Bezeichnung	Bedeutung
hourly	Die Messdaten werden im Stundentakt erhoben
daily	Die Messdaten werden einmal pro Tag erhoben
monthly	Die Messdaten werden einmal pro Monat erhoben
yearly	Die Messdaten werden einmal pro Jahr erhoben
nonregular	Die Messdaten werden in unregelmäßigen Zeitabständen erhoben

Tab. 18.9 - Übersicht über die in den eingebundenen Zeitreihendatenbanken dokumentierten zeitlichen Auflösungen (Stand Oktober 2003).

- **Ausprägungen der Klassifizierungskriterien in den Zeitreihendatenbanken**

Die Klassifizierungskriterien Stationstyp, Variable und zeitliche Auflösung finden sich in den drei gegenwärtig eingebundenen Zeitreihendatenbanken (PIK German Time Series Database, PIK Global Time Series Database und PIK Scenario Time Series Database) in unterschiedlichen Ausprägungen. Tab. 18.10 gibt einen Überblick über die jeweilige Zusammensetzung.

Zeitreihen-datenbank	Anzahl Variablen pro Stationstyp				Zeitliche Auflösungen				
	hydro-logy	*meteoro-logy*	*pheno-logy*	*water quality*	*hourly*	*daily*	*monthly*	*yearly*	*non-regular*
German	3	41	257	12	✓	✓	✓	✓	✓
Global	1	12				✓	✓		
Scenario		54 [300]				✓			

Tab. 18.10 - Anzahl unterschiedlicher Variablen pro Stationstyp sowie verfügbare zeitliche Auflösungen in den eingebundenen Zeitreihendatenbanken (Stand Oktober 2003).

- **Zeitliche Abdeckung**

Sowohl Zeitreihenmetadatenbanken wie Zeitreihendatenbanken dokumentieren für jede Station den ersten sowie letzten Zeitpunkt der Datenerhebung beim Datengeber (Attribute Begin_Date und End_Date). In Zeitreihenmetadatenbanken wird diese *zeitliche Abdeckung* entweder für eine Station als ganze oder - wenn eine zusätzliche Klassifizierung über Subtypen unterstützt wird - für jeden Subtyp einer Station beschrieben. Zeitreihendatenbanken dokumentieren hingegen die zeitliche Abdeckung[301] getrennt für jede an einer Station erhobene Variable, für die Werte vorliegen. Die zeitliche Abdeckung kann für einzelne Stationen jeweils unterschiedlich ausfallen. Zudem kann sie sich bei derselben Station für einzelne dort erhobene Variablen unterschiedlich gestalten, bspw. wenn dort die Anzahl der erhobenen Parameter im Lauf der Zeit erweitert oder - durch die Einstellung von Messungen ab einem bestimmten Zeitpunkt - verringert wurde.

- **Georeferenzierung**

Die Verortung der dokumentierten Stationen im *geographischen Raum* erfolgt in Zeitreihenmetadatenbanken wie Zeitreihendatenbanken jeweils durch mehrere einander ergänzende Entitäten:

Position ▸ Sämtliche Zeitreihenmetadatenbanken wie Zeitreihendatenbanken dokumentieren für jede Station jeweils zwei Werte (Attribute

[300] Die PIK Scenario Time Series Database dokumentiert für jede Station 11 unterschiedliche Variablen in mehreren Varianten: Als Referenzdaten werden für die Jahre von 1951 bis 2000 für 10 Variablen sowohl homogenisierte wie interpolierte Messwerte sowie für eine weitere Variable (Niederschlag) homogenisierte Messwerte bereitgestellt. Ferner werden drei konstruierte Szenarien für die Jahre von 2001 bis 2055 (das als wahrscheinlichstes eingestufte Szenario sowie Szenarien mit normaler bzw. zunehmender Niederschlagsentwicklung) bereitgestellt.

[301] Bei Zeitreihendatenbanken bildet die zeitliche Abdeckung neben der zeitlichen Auflösung damit ein zweites zeitbezogenes Selektionskriterium.

Stat_Lat und Stat_Lon), die ihre Position im geographischen Koordinatensystem definieren.

Hierarchische räumliche Klassifikation
- Zudem ordnen sämtliche Zeitreihenmetadatenbanken wie Zeitreihendatenbanken jede Station anhand eines entsprechenden hierarchischen räumlichen Klassifikators (vgl. Kap. 18.3.1) in eine globale Hierarchie administrativer Einheiten aus Kontinenten, Staaten und gegebenenfalls feineren Aufteilungen ein (Attribut Admin_ID). Zusätzlich kommen entsprechende Klassifikatoren zur Einordnung der dokumentierten Station in eine globale Hierarchie von Flusseinzugsgebieten zum Einsatz (Attribut Basin_ID).

Höhe über dem Meeresspiegel
- Sowohl Zeitreihenmetadatenbanken wie Zeitreihendatenbanken dokumentieren für jede Station ihre Höhe über dem Meeresspiegel in Metern (Attribut Stat_Alt).

Flussnamen
- In Zeitreihenmetadatenbanken wird für Hydrologie-Stationen und für Wasserqualitäts-Stationen zusätzlich der Name des Flusses (Attribut River) dokumentiert, an dem diese ihre Messungen durchführen.

- **Direkte Referenzierung – Stationsidentifikatoren und Stationsnamen**

Zeitreihenmetadatenbanken wie Zeitreihendatenbanken weisen jeder dokumentierten Station sowohl eine eindeutige Kennnummer als *Identifikator* (Attribut Stat_ID) als auch einen *Namen* (Attribut Stat_Name) zu. Diese Entitäten können dazu herangezogen werden, einzelne Stationen direkt zu selektieren.

- **Statistische Kriterien**

Die Zeitreihendatenbanken stellen zusätzlich Jahresstatistiken[302] mit vorberechneten Kennwerten bereit. Anhand dieser Daten können für jede Variable jeder dokumentierten Station verschiedene Kennwerte für den jeweiligen Verfügbarkeitszeitraum - oder nutzerdefinierbare Ausschnitte aus diesem - abgeleitet werden (vgl. Tab. 18.11).

Bezeichnung	Bedeutung
Values_Num	Die Anzahl der Messwerte für den ausgewählten Zeitraum
Values_Sum	Die Summe der Messwerte für den ausgewählten Zeitraum
Values_Min	Der kleinste Messwert für den ausgewählten Zeitraum
Values_Max	Der größte Messwert für den ausgewählten Zeitraum
Values_Avg	Der Durchschnitt der Messwerte für den ausgewählten Zeitraum
Values_Var	Die Varianz der Messwerte für den ausgewählten Zeitraum
Values_Cpl	Die prozentuale Vollständigkeit der Messwerte für den ausgewählten Zeitraum
Values_ Rel_Cover	Die relative Abdeckung (Verhältnis der Anzahl der in eine Berechnung einbezogenen Jahre zur Anzahl der Jahre, die durch den ausgewählten Zeitraum vorgegeben wird)

Tab. 18.11 - Ableitbare Kennwerte aus den Jahresstatistiken der Zeitreihendatenbanken.

Durch nutzerdefinierbare Bedingungen auf diesen Kennwerten (vgl. das Filtermodul StatisticFilter, Kap. 20.9) wird eine Selektion von Stationen anhand von Eigenschaften der von ihnen erhobenen Zeitreihen ermöglicht.

[302] Zeitreihen sind intervallskalierte Daten (vgl. Kap. 1.3.1), auf denen verschiedene statistische Auswertungen durchgeführt werden können.

Bezeichnung[303]	Bedeutung	Selektion durch...	Verwendung in	
			ZRMDB	ZRDB
Stations-identifikator (Stat_ID)	Kennzahl zur eindeutigen Identifikation jeder dokumentierten Station	Auswahl von Identifikatoren	✓	✓
Stationsname (Stat_Name)	Bezeichner für jede dokumentierte Station	Auswahl von Stationsnamen	✓	✓
Stationstyp (Stat_Type)	Übergreifende Klassifizierung der dokumentierten Stationen anhand ihrer Beobachtungsaufgabe	Vorgabe von Stationstypen	✓	✓
Zeitliche Abdeckung (Begin_Date, End_Date)	Erster sowie letzter Zeitpunkt der Datenerhebung an den dokumentierten Stationen	Auswahl einer zeitlichen Abdeckung	✓	✓
Position (Stat_Lon, Stat_Lat)	Die Position der dokumentierten Stationen im geographischen Koordinatensystem	Auswahl einer Boundingbox[304]	✓	✓
Administrative Einheit (Admin_ID)	Einordnung der dokumentierten Stationen in eine administrative Hierarchie	Auswahl vordefinierter administrativer Einheiten	✓	✓
Flusseinzugsgebiet (Basin_ID)	Einordnung der dokumentierten Stationen in eine globale Hierarchie von Flusseinzugsgebieten	Auswahl vordefinierter Flusseinzugsgebiete	✓	✓
Höhe (Stat_Alt)	Die Höhe (in Metern) der dokumentierten Stationen über dem Meeresspiegel	Vorgabe unterer und / oder oberer Grenzen	✓	✓
Subtyp[305] (Stat_Subtype)	Feinere Klassifizierung der dokumentierten Stationen anhand von Unterklassen	Vorgabe von Subtypen	✓	
Status der Datenverfügbarkeit (PIK_Availability)	Klassifizierung der dokumentierten Stationen anhand der Möglichkeiten zum Zugriff auf die zugeordneten Zeitreihen	Vorgabe von Formen der Datenverfügbarkeit	✓	
Flussname[306] (River)	Bezeichnung des Flusses, an dem dokumentierte Stationen ihre Messungen durchführen	Auswahl von Flussnamen	✓	
Variable (Variable)	Die an den dokumentierten Stationen erhobenen Parameter	Vorgabe von Variablen		✓
Zeitliche Auflösung (Temp_Resol)	Die Erhebungsfrequenzen der Parameter an den dokumentierten Stationen	Vorgabe von zeitlichen Auflösungen		✓
Jahresstatistiken	Statistische Eigenschaften der von den dokumentierten Stationen erhobenen Zeitreihen	Auswahl von Kennwerten sowie Vorgabe von Bedingungen auf diesen		✓

Tab. 18.12 - Selektionskriterien für Zeitreihenmetadatenbanken (ZRMDB) und Zeitreihendatenbanken (ZRDB).

Tab. 18.12 gibt einen zusammenfassenden Überblick über die Selektionskriterien, die

[303] In Klammern die den Anwendern über die Schnittstelle präsentierten Bezeichnungen.
[304] Es können vordefinierte oder frei definierbare Boundingboxen gewählt werden. Die Auswahl vordefinierter Boundingboxen erfolgt durch Selektion vordefinierter geographischer Einheiten. Sie ist nur erforderlich, wenn der ausgewählte Datenraum für diese Einheiten keine entsprechenden hierarchischen räumlichen Klassifikatoren bereitstellt.
[305] Wird nur in der DWD Measurement Network Meta Database verwendet.
[306] Wird nur für Hydrologie-Stationen und Wasserqualitäts-Stationen verwendet.

übergreifend für Zeitreihenmetadatenbanken und Zeitreihendatenbanken anwendbar sind sowie über solche, die spezifisch für Zeitreihenmetadatenbanken bzw. für Zeitreihendatenbanken sind.

Sämtliche für einen konkreten Datenraum verfügbaren Selektionskriterien können jeweils beliebig miteinander kombiniert werden. Für die Präsentation von Anfrageergebnissen kann der Anwender vor Beginn einer Anfrage zusätzlich beliebige Kombinationen[307] der jeweiligen Datenbankattribute auswählen.

18.4 Zeitreihen

Die im Data Warehouse integriert vorgehaltenen punktverorteten Zeitreihen des Institutes können vom Anwender anhand der in den Zeitreihendatenbanken dokumentierten Metadaten für eine Visualisierung oder den Export auf den Rechner des Anwenders (Download) ausgewählt werden. In diesem Kapitel werden zunächst die Entitäten beschrieben, die zur Auswahl der Zeitreihen verwendet werden. Abschließend wird auf die unterschiedlichen zeitlichen Aggregationen eingegangen, in denen Zeitreihen für Visualisierung und Download bereitgestellt werden können.

18.4.1 Auswahl von Zeitreihen

Die Metadaten jeder in einer Zeitreihendatenbank dokumentierten Station erlauben eine eindeutige Identifikation der dieser jeweils zugeordneten Zeitreihen im Data Warehouse. Die Identifikation jeder individuellen Zeitreihe erfolgt dabei durch die Kombination von Werteausprägungen der Entitäten:

- Stationsidentifikator ‣ eindeutige Identifikation einer Station;
- Variable ‣ eindeutige Identifikation einer Messvariable;
- Zeitliche Auflösung ‣ eindeutige Identifikation einer zeitlichen Auflösung.

Die jeweils relevanten Ausschnitte der Zeitreihen werden durch die zusätzliche Angabe eines zeitlichen Ausschnitts definiert:

- Zeitfenster ‣ Festlegung des ersten sowie letzten Zeitpunktes, für den Werte von Zeitreihen zu extrahieren sind.

Dabei soll es dem Anwender überlassen bleiben, welche Zeitreihen von welchen Stationen er jeweils für eine Extraktion miteinander kombiniert. Er soll anhand der Auswahl einer Untermenge zuvor selektierter Zeitreihenmetadaten zunächst entsprechende Kombinationen der Werteausprägungen von Stationsidentifikatoren, Variablen und zeitlichen Auflösungen festlegen und anschließend einen Ausschnitt aus den so ausgewählten Zeitreihen durch die Eingabe eines Zeitfensters definieren können.

18.4.2 Auswahl der zeitlichen Aggregation

Durch die integrierte Bereitstellung sämtlicher relevanter punktverorteter Zeitreihen des Institutes über das Data Warehouse können Zeitreihen aus zuvor nur getrennt zugänglichen Datenräumen gemeinsam extrahiert werden. Zusätzlich wurde Wert darauf gelegt, dass auch Zeitreihen mit unterschiedlichen zeitlichen Auflösungen für Visualisierung oder Download miteinander kombiniert werden können. Die im Data Warehouse vorgehaltenen Zeitreihen werden zusätzlich zu der jeweiligen zeitlichen Auflösung, in der sie erhoben wurden, zeitlich aggregiert vorgehalten. Dabei finden entsprechend Aggregationen hin zu jeweils gröberen zeitlichen Auflösungen statt. So werden bspw. aus Tageswerten Monats-

[307] Eine Ausnahme bilden hierbei nur Attribute wie Stat_ID (Stationsidentifikator), die intern zur korrekten Zuordnung von Datensätzen zu einzelnen Stationen verwendet werden und daher nicht vom Anwender deselektiert werden dürfen.

sowie Jahreswerte berechnet; jede Aggregation führt jeweils zu einer Reihe von statistischen Kennwerten wie Summe, Extrema, Varianz etc.

Der Anwender kann entsprechend als Zielauflösung einer Zeitreihenextraktion entweder die zeitliche Auflösung der Ausgangsdaten oder eine der jeweils verfügbaren zeitlichen Aggregationen auswählen. Auf diese Weise können auch Zeitreihen mit unterschiedlicher zeitlicher Auflösung gemeinsam extrahiert und bereitgestellt werden; dabei werden diejenigen Zeitreihen, deren Ausgangsdaten eine feinere zeitliche Auflösung als die Zielauflösung besitzen, entsprechend auf diese abgebildet.

Zeitliche Auflösung der Ausgangsdaten	**Mögliche zeitliche Auflösungen für Visualisierung bzw. Export**		
	daily	*monthly*	*yearly*
daily	✓	✓	✓
monthly		✓	✓
yearly			✓
nonregular			✓

Tab. 18.13 - Unterstützte Abbildungen von Basis- auf Zielauflösungen (Stand Oktober 2003)[308].

Tab. 18.13 gibt einen Überblick über die unterschiedlichen zeitlichen Auflösungen der Ausgangsdaten im Data Warehouse sowie die für diese jeweils vom Anwender auswählbaren zeitlichen Aggregationen.

[308] Ein Zugriff auf Stundenwerte, die für einige Stationen der deutschen Zeitreihendatenbank des Institutes vorliegen, wird gegenwärtig noch nicht unterstützt.