24 Akzeptanz

Der erreichte Stand wird von den Wissenschaftlern des Instituts in informellen Rückmeldungen durchgehend positiv bewertet. Dass tatsächlich die angestrebte übergreifende Akzeptanz sowohl durch Datenbereitsteller wie Datennutzer erreicht wurde, lässt sich anhand der bis heute über die Schnittstelle bereitgestellten, zuvor isolierten Datenräume sowie der Zahl der Zugriffe auf diese objektivieren. Kap. 24.1 beschreibt die über die Schnittstelle zugänglichen Daten, Kap. 24.2 Zahl und Art der Zugriffe und Kap. 24.3 die institutsexterne Sichtbarkeit.

24.1 Zugängliche Daten

Die allgemeine Institutsmetadatenbank PIK CERA-2 enthält gegenwärtig[338] 358 Metadatensätze zur Dokumentation unterschiedlicher Datenressourcen (zur thematischen Verteilung der Metadatensätze vgl. Abb. 24.1a). Die über die Schnittstelle erschließbare Datenschicht umfasst zudem eine sehr umfangreiche Sammlung nationaler und internationaler Zeitreihenmetadaten sowie Zeitreihen hoher Qualität aus verschiedenen Wissenschaftsdisziplinen, die laufend erweitert wird (vgl. Abb. 24.1b bis 24.1d).

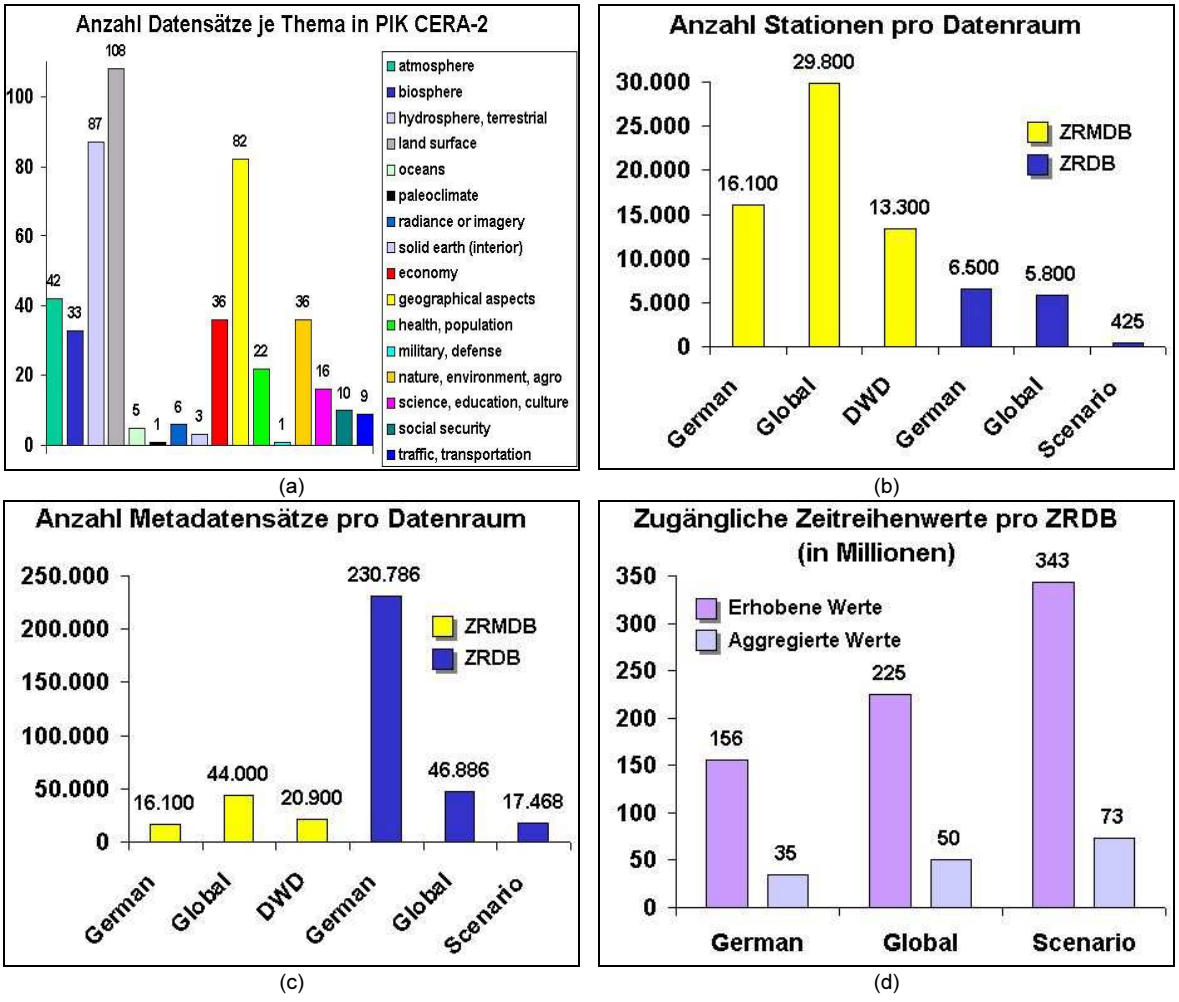


Abb. 24.1 - Über die Schnittstelle zugängliches Datenvolumen (Stand Okt. 2003): (a) Anzahl der Metadatensätze nach Themengebieten in PIK CERA-2; (b) Anzahl der in den einzelnen Zeitreihenmetadatenbanken (ZRMDB) und Zeitreihendatenbanken (ZRDB) dokumentierten Stationen; (c) Anzahl der dort jeweils enthaltenen Metadatensätze; (d) Anzahl der anhand der einzelnen Zeitreihendatenbanken jeweils aus dem Data Warehouse abrufbaren Zeitreihenwerte.

In den über die Schnittstelle adressierbaren Zeitreihenmetadatenbanken und Zeitreihendaten-

[338] Die Angaben in Kap. 24.1 beziehen sich auf den Stand vom Oktober 2003.

banken - mit teilweisen Überschneidungen in den einzelnen Datenräumen - sind insgesamt rund 376.140 Metadatensätze zu 71.925 nationalen und internationalen Stationen[339] enthalten (vgl. Abb. 24.1b und 24.1c). Das zur integrierten Bereitstellung der punktverorteten Zeitreihen verwendete Data Warehouse wird beständig erweitert und stellt annähernd eine Milliarde (rund 882 Millionen)[340] einzelner Zeitreihenwerte bereit, die sich aus erhobenen sowie aus diesen abgeleiteten, zeitlich aggregierten Werten zusammensetzen (vgl. Abb. 24.1d).

Nachfolgend wird kurz der Inhalt der eingebundenen Datenräume aus den Gruppen der Zeitreihenmetadatenbanken sowie der Zeitreihendatenbanken skizziert (vgl. auch Tab. 24.1, die diese Angaben überblicksartig zusammenfasst). In den drei *Zeitreihenmetadatenbanken* sind Metadaten zu insgesamt 59.200 Stationen enthalten, die sich wie folgt verteilen:

PIK German Measurement Network Meta Database
- Über die deutsche Messnetz-Metadatenbank des PIK sind anhand von 16.100 Einträgen Informationen über ebenso viele deutsche Stationen aus den Bereichen Meteorologie, Hydrologie, Wasserqualität und Phänologie abrufbar. Die maximale zeitliche Abdeckung der dokumentierten Stationen umfasst die Jahre von 1801 bis 2001.

PIK Global Measurement Network Meta Database
- Über die globale Messnetz-Metadatenbank des PIK können 44.000 Einträge über insgesamt 29.800 Stationen aus allen Kontinenten zu den Bereichen Meteorologie und Hydrologie ausgewertet werden. Die maximale zeitliche Abdeckung der in dieser Datenbank dokumentierten Stationen reicht von 1679 bis 2001.

DWD Measurement Network Meta Database
- Aus der DWD-Messnetz-Metadatenbank des PIK können 20.900 Einträge zur Dokumentation von insgesamt 13.300 Stationen des Deutschen Wetterdienstes aus den Bereichen Meteorologie und Phänologie abgerufen werden. Die maximale zeitliche Abdeckung der dokumentierten Stationen umfasst die Jahre von 1870 bis 2001.

Die drei *Zeitreihendatenbanken* enthalten Metadaten zu insgesamt 12.725 Stationen, für die zusätzlich die jeweils zugehörigen Zeitreihen aus dem Data Warehouse direkt über die Schnittstelle selektiert und bereitgestellt werden können:

PIK German Time Series Database
- Über die deutsche Zeitreihendatenbank des Institutes stehen 230.786 Einträge zur Dokumentation von insgesamt 6.500 deutschen Stationen mit einer gesamten zeitlichen Abdeckung von 1781 bis 2000 zur Verfügung. Es sind insgesamt 313 unterschiedliche Variablen aus den Bereichen Meteorologie, Hydrologie, Wasserqualität und Phänologie in insgesamt 5 unterschiedlichen zeitlichen Auflösungen dokumentiert (stündliche, tägliche, monatliche, jährliche und unregelmäßige Erhebung). Anhand dieser Zeitreihenmetadaten können rund 156 Millionen Zeitreihenwerte in Basisauflösung sowie weitere 35 Millionen zeitlich aggregierte Werte aus dem Data Warehouse abgerufen werden.

PIK Global Time Series Database
- Die globale Zeitreihendatenbank des Institutes beinhaltet 46.886 Metadatensätze über 5.800 internationale Stationen mit einer gesamten zeitlichen Abdeckung von 1775 bis 2002. Es sind insgesamt 13 unterschiedliche Variablen aus den Bereichen Meteorologie und Hydrologie in insgesamt 2 unterschiedlichen zeitlichen Auflösungen (tägliche und

[339] Die im Vergleich zur Anzahl der Stationen teilweise deutlich höhere Zahl der Metadatensätze resultiert aus der Struktur der unterliegenden Datenbanken, die bspw. pro zu dokumentierender Kombination aus Station, Variable und zeitlicher Auflösung bzw. aus Station und Subtyp jeweils einen Datensatz bereitstellen.
[340] Die Anzahl der Einzelwerte stieg bis Dezember 2003 bereits auf rund 905 Millionen.

monatliche Erhebung) dokumentiert; anhand der Zeitreihenmetadaten kann auf 225 Millionen einzelne Zeitreihenwerte in Basisauflösung sowie 50 Millionen zeitlich aggregierte Werte im Data Warehouse zugegriffen werden.

PIK Scenario Time Series Database

Über die Szenarien-Zeitreihendatenbank des Institutes sind anhand von 17.468 Einträgen Informationen über 425 Stationen im Flusseinzugsgebiet der Elbe sowie in Süddeutschland verfügbar, für die sowohl Zeitreihen für Referenzdaten wie für am Institut konstruierte Szenarien bis zum Jahr 2055 zur Verfügung gestellt werden. Dabei werden für jede Station 11 Variablen aus dem Bereich Meteorologie in jeweils 5 unterschiedlichen Varianten[341] dokumentiert. Davon entfallen zwei Varianten auf Referenzdaten (homogenisierte sowie interpolierte Messwerte für die Jahre von 1951 bis 2000) und die weiteren auf drei konstruierte Szenarien für die Jahre von 2001 bis 2055 (das als wahrscheinlichstes eingestufte Szenario sowie Szenarien mit normaler bzw. zunehmender Niederschlagsentwicklung). Für jede dieser 54 Variablen können jeweils vollständig vorliegende Zeitreihen in täglicher Auflösung aus dem Data Warehouse abgerufen werden; insgesamt kann so auf insgesamt 343 Millionen einzelne Zeitreihenwerte in Basisauflösung und weitere 73 Millionen zeitlich aggregierte Werte zugegriffen werden.

	Zeitreihenmetadatenbanken			**Zeitreihendatenbanken**		
	German	*Global*	*DWD*	*German*	*Global*	*Scenario*
Metadatensätze	16.100	44.000	20.900	230.786	46.886	17.468
Stationen	16.100	29.800	13.300	6.500	5.800	425
Meteorologie	✓	✓	✓	✓	✓	✓
Hydrologie	✓	✓		✓	✓	
Wasserqualität	✓			✓		
Phänologie	✓		✓	✓		
Räumliche Abdeckung	Deutschland	Welt	Deutschland	Deutschland	Welt	Deutschland
Max. zeitliche Abdeckung	1801 bis 2001	1679 bis 2001	1870 bis 2001	1781 bis 2000	1775 bis 2002	1951 bis 2055
Variablen				313	13	54
Zeitliche Auflösungen				5	2	1
Zugängliche Zeitreihenwerte[342]				156 Mio. + 35 Mio.	225 Mio. + 50 Mio.	343 Mio. + 73 Mio.

Tab. 24.1 - Überblick über den Inhalt der einzelnen Datenräume aus den Gruppen der Zeitreihenmetadatenbanken und Zeitreihendatenbanken (Stand Okt. 2003).

24.2 Zahl und Art der Zugriffe

Seit Beginn von Betriebsphase II konnten - bei rund 120 wissenschaftlichen Mitarbeitern des Institutes - bis Oktober 2003 mehr als 10.000 Anwenderzugriffe über die Schnittstelle verzeichnet werden. Aus dem Data Warehouse wurden in diesem Zeitraum von Wissenschaftlern des PIK rund 330 Millionen Zeitreihenwerte extrahiert, also rund ein Drittel des enthaltenen Gesamtvolumens.

[341] Eine Ausnahme bildet die Variable Precipitation (Niederschlag), für die nur homogenisierte Werte bereitgestellt werden, so dass diese nur in 4 Varianten (Referenzwerte und 3 Szenarien) vorliegt.
[342] Basiswerte sowie aus diesen durch Aggregation abgeleitete zeitbezogene sekundäre statistische Werte.

Insgesamt lassen sich für die einzelnen Datenräume folgende Gewichtungen dokumentieren. Zunächst (vgl. Abb. 24.2a) zeigt eine Aufschlüsselung der Zugriffe nach den drei Datenbankgruppen mit 70 Prozent der Gesamtzugriffe eine deutliche Präferenz der Anwender für die Gruppe der Zeitreihendatenbanken, gefolgt von den Gruppen der allgemeinen Metadaten (24 Prozent) und der Zeitreihenmetadatenbanken (6 Prozent).

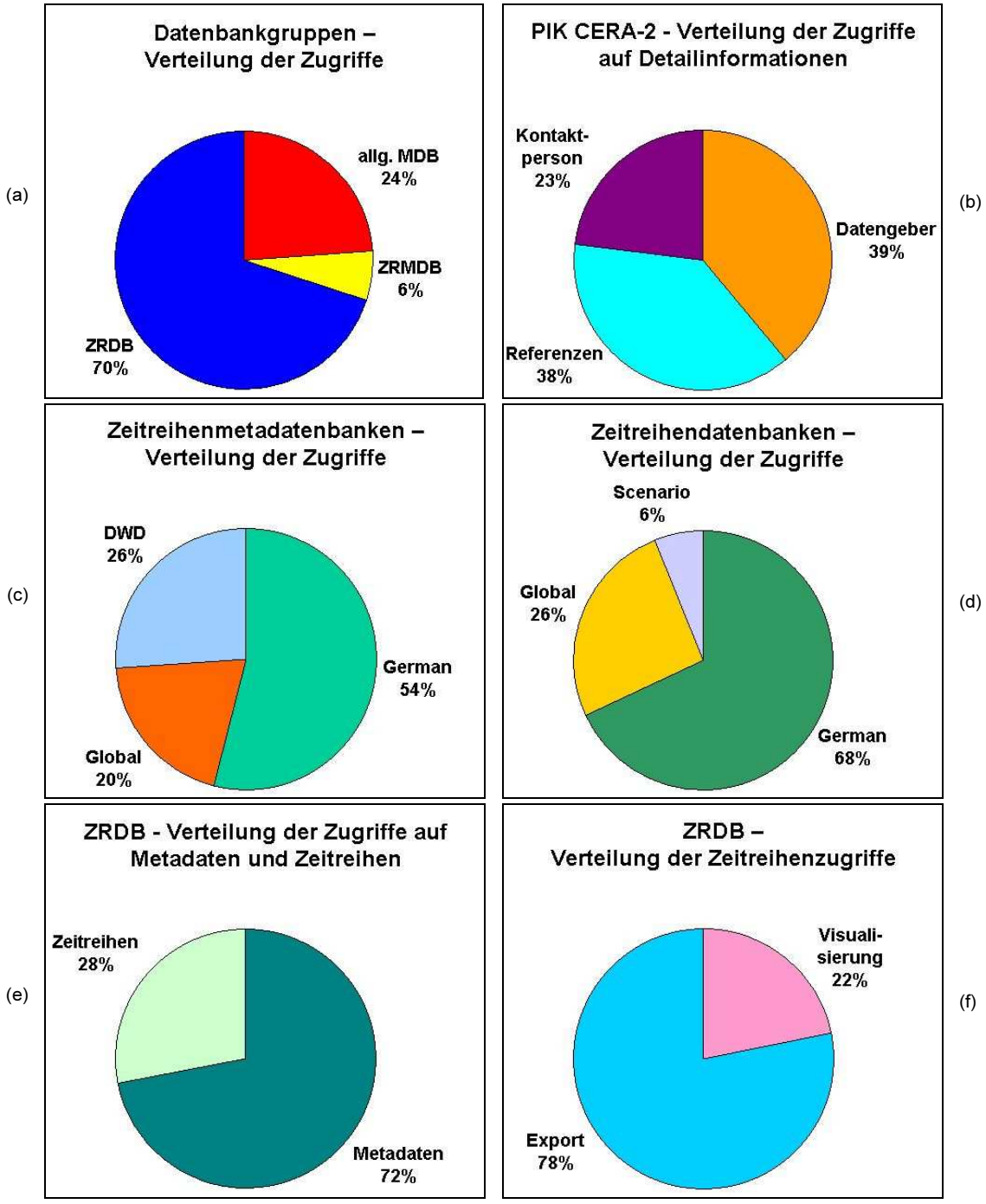
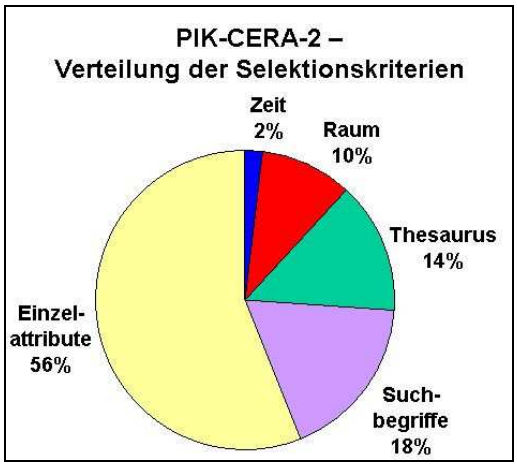


Abb. 24.2 - Verteilung der erfolgten Zugriffe: (a) Aufteilung auf die drei Datenbankgruppen; (b) Aufteilung der Detailzugriffe in PIK CERA-2; Aufteilung auf die einzelnen Datenräume in den Datenbankgruppen der Zeitreihenmetadatenbanken (c) sowie der Zeitreihendatenbanken (d); (e) Verhältnis der Zugriffe auf Metadaten und Zeitreihen in der Gruppe der Zeitreihendatenbanken; (f) Aufteilung der Zeitreihenzugriffe auf Visualisierung und Export zum Anwender.

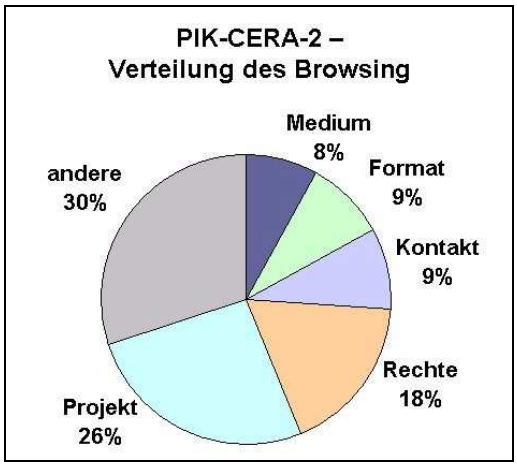
Das folgende Diagramm (Abb. 24.2b) zeigt die Verteilung der Zugriffe auf zusätzliche Detailinformationen für Ergebnisdatensätze aus der allgemeinen Metadatenbank PIK CERA-2 auf die hierfür angebotenen Entitäten Datengeber, Kontaktpersonen und Referenzen (vgl. Kap. 18.2.2).

Während in der Gruppe der allgemeinen Metadatenbanken im Beobachtungszeitraum alleine PIK CERA-2 zur Verfügung stand, setzten sich die beiden anderen Gruppen aus jeweils drei Datenbanken zusammen, so dass hier die Verteilung der Zugriffszahlen nach ihrer relativen Häufigkeit weiter aufgeschlüsselt werden kann. Die am häufigsten angesprochenen Datenräume sind demnach innerhalb der Gruppe der Zeitreihenmetadatenbanken die PIK German Measurement Network Meta Database mit 54 Prozent aller in dieser Gruppe erfolgten Zugriffe (vgl. Abb. 24.2c) sowie innerhalb der Gruppe der Zeitreihendatenbanken die PIK German Time Series Database mit 68 Prozent aller in dieser Gruppe erfolgten Zugriffe (vgl. Abb. 24.2d). Für die Gruppe der Zeitreihendatenbanken kann ferner das Verhältnis von Metadatenzugriffen sowie Zeitreihenzugriffen zueinander bestimmt werden (vgl. Abb. 24.2e). 72 Prozent aller Zugriffe in der Gruppe der Zeitreihendatenbanken bezogen sich auf Zeitreihenmetadaten in den diesbezüglichen Datenräumen; die verbleibenden 28 Prozent der Zugriffe bestanden in hierauf basierenden Extraktionen von Zeitreihenwerten aus dem Data Warehouse. Eine weitere Aufgliederung der Data Warehouse-Zugriffe (vgl. Abb. 24.2f) zeigt, dass 22 Prozent dieser Zugriffe zur interaktiven Visualisierung von Zeitreihen über die Schnittstelle genutzt wurden, während der überwiegende Anteil von 78 Prozent auf den Download von Zeitreihen auf den Rechner des Anwenders zur nachfolgenden Weiterverarbeitung entfiel.

(a)



(b)



(c)

(d)

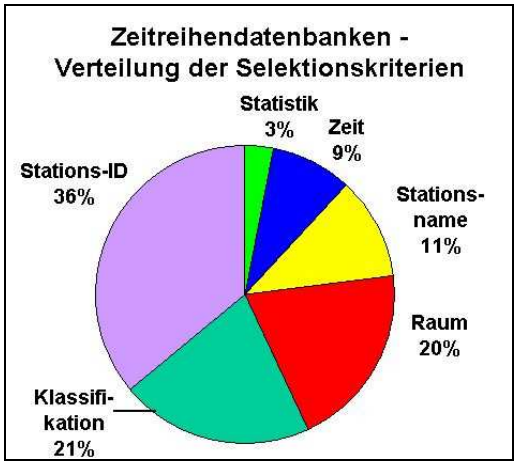


Abb. 24.3 - Nutzung von Filtermodulen in unterschiedlichen Datenräumen: Verteilung von Selektionskriterien (a) und Browsing (b) in PIK CERA-2; Verteilung der Selektionskriterien in den Zeitreihenmetadatenbanken (c) und den Zeitreihendatenbanken (d).

Die Diagramme in Abb. 24.3 veranschaulichen die Bandbreite, in der die verschiedenen Filtermodule der Schnittstelle von Anwendern für die Selektion von Datensätzen genutzt werden. Zunächst wird der jeweilige Anteil einzelner Selektionskriterien an sämtlichen auf der allgemeinen Metadatenbank PIK CERA-2 formulierten Bedingungen dargestellt (Abb. 24.3a). Zu beachten ist hierbei, dass sich der große Anteil der Selektionskriterien auf Einzelattributen von 56 Prozent aus

zehn einzeln adressierbaren Teilbedingungen auf jeweils unterschiedlichen Attributen zusammensetzt.

Das Diagramm in Abb. 24.3b zeigt, wie breit gefächert in diesem Kontext das über den SingleAttributeFilter bereitgestellte Browsing von Werteausprägungen (vgl. Kap. 20.2) zur Unterstützung bei der Anfrageerstellung eingesetzt wird. Die Abfrage gültiger Werteausprägungen vor einer eigentlichen Datenselektion aus PIK CERA-2 wird demnach von Anwendern am häufigsten für PIK-Projekte (26 Prozent) sowie Verwendungsrechte (18 Prozent) in Anspruch genommen. Auch die Darstellung des jeweiligen Anteils einzelner Selektionskriterien an sämtlichen formulierten Bedingungen auf den Zeitreihenmetadatenbanken (vgl. Abb. 24.3c) sowie den Zeitreihendatenbanken (vgl. Abb. 24.3d) bestätigt den Eindruck, dass alle angebotenen Filtermodule - in unterschiedlicher Gewichtung - zur Datenselektion herangezogen werden. Es sei an dieser Stelle darauf hingewiesen, dass diese Diagramme keine Aussagen über die konkrete Zusammensetzung einzelner Datenbankanfragen treffen, die ja aus beliebigen Kombinationen von Selektionskriterien gebildet werden können.

Abschließend werden die unterschiedlichen Formen dargestellt, die von Anwendern zur Definition von Raumbezügen zur Auswahl von Stationen in Zeitreihenmetadatenbanken und Zeitreihendatenbanken eingesetzt werden (vgl. Abb. 24.4). Zu beachten ist, dass das Selektionskriterium *Flussname* ausschließlich in Zeitreihenmetadatenbanken, die Hydrologie- oder Wasserqualitäts-Stationen dokumentieren - also nur in der deutschen sowie der globalen Messnetz-Metadatenbank des PIK - zur Verfügung steht (vgl. Tab. 24.1). Ferner ist zu berücksichtigen, dass die Auswahl von Flussnamen und Stationshöhe (jeweils über Instanzen des SingleAttributeFilter) zusätzlich zu einer über den SpatialFilter möglichen Raumauswahl erfolgen kann, die ihrerseits entweder in der Auswahl einer Boundingbox über das Gummiband oder der Auswahl vordefinierter Flusseinzugsgebiete oder administrativer Einheiten bestehen kann.

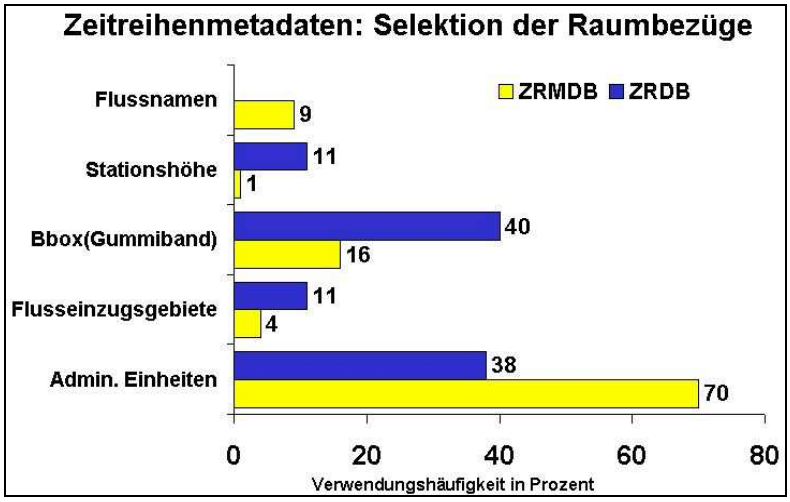


Abb. 24.4 - Nutzung verschiedener Kriterien zur Raumselektion in Zeitreihenmetadatenbanken und Zeitreihendatenbanken.

Auch hier zeigt sich, dass sämtliche angebotenen Formen der Raumselektion jeweils ihre Berechtigung besitzen, wobei diese von den Anwendern in den Gruppen der Zeitreihenmetadatenbanken und Zeitreihendatenbanken jeweils in unterschiedlichen Ausprägungen zur Datenselektion eingesetzt werden.

24.3 Institutsexterne Sichtbarkeit

Aus datenrechtlichen Gründen wird die Schnittstelle vorerst primär im Intranet des Institutes betrieben. Auf Vorträgen wurden Prototypen der Schnittstelle sowie des interaktiven digitalen Atlas IDA gemeinsam mit CERA-2 im Workshop Heterogene, aktive Umweltdatenbanken (1998) des

Arbeitskreises Umweltdatenbanken[343] auf der Insel Vilm [Wrobel 1999] [Toussaint et al. 1999] sowie auf der International Conference on Quality, Management and Availibility of Data for Hydrology and Water Resources Management 1999 in Koblenz [Toussaint, Wrobel 1999] vorgestellt. Verschiedene nationale und internationale Institutionen, die für ihre Forschungsarbeit ebenfalls das Internet für einen effizienten Zugriff auf komplexe Datenbestände nutzen wollen, haben das PIK und den Autor dieser Arbeit zwecks Erfahrungsaustausch und Beratung konsultiert; die Schnittstelle wurde in verschiedenen Entwicklungsstadien u.a. Mitarbeitern des GeoForschungsZentrums Potsdam (GFZ)[344], des Biosphärenreservats Schorfheide/Chorin[345] oder des Botanischen Gartens Potsdam[346] demonstriert sowie anhand von Kurzvorträgen Studenten der Humboldt Universität Berlin (Informatik) und der FU Berlin (Mathematik) vorgestellt. Zeitlich befristete Demonstrationsclients wurden dem Max-Planck-Institut für BioGeoChemie (MPI-BGC)[347] in Jena, der National Energy Authority (NEA)[348] in Reykjavik / Island sowie dem Tyndall-Centre[349] in Norwich / Großbritannien zur Verfügung gestellt.

Gegenwärtig wertet das Institut für Meteorologie[350] der FU Berlin PIK CERA-2 und xDat anhand eines DemonstrationsClient auf ihre Eignung für die Dokumentation und internetbasierte Erschließung von Metadaten aus mehreren Disziplinen aus; die Bereitstellung eines weiteren DemonstrationsClient für das Max-Planck-Institut für Meteorologie (MPIfM)[351] in Hamburg ist vorgesehen. Ferner ist in der Diskussion, dass xDat vom Tyndall-Centre in naher Zukunft im Rahmen einer Forschungskooperation sowohl für die Erschließung von dort administrierten Zeitreihenmetadaten und punktverorteten Zeitreihen wie für den Zugriff der dort zusammengeschlossenen Wissenschaftler auf die Datenbestände des PIK genutzt werden soll. Die entstandene Software sowie autonome Komponenten wie IDA sollen in Zukunft ferner institutsextern als *open source* Software über GNU GPL[352] zur Verfügung gestellt werden.

[343] Der Arbeitskreis Umweltdatenbanken ist ein Arbeitskreis der Fachgruppe 4.6.1 *Informatik im Umweltschutz* (http://www.iai.fzk.de/Fachgruppe/GI/welcome.html) der Gesellschaft für Informatik e.V. (http://www.gi-ev.de/).
[344] http://www.gfz-potsdam.de/
[345] http://www.schorfheide-chorin.de/
[346] http://www.bio.uni-potsdam.de/botgar/
[347] http://www.bgc-jena.mpg.de/
[348] http://www.os.is/english/
[349] Das Tyndall-Centre (http://www.tyndall.ac.uk/) verbindet dreizehn britische Forschungseinrichtungen wie Cambridge und Southhampton zu einem temporären virtuellen Institut und kann als Prototyp für ein europäisches Forschungsnetzwerk angesehen werden.
[350] http://www.met.fu-berlin.de/
[351] http://www.mpimet.mpg.de/
[352] GNU GPL steht für die *GNU General Public License* (http://www.gnu.org/copyleft/gpl.html).