



7. Inetbibtagung 2003 in Frankfurt

Workshop

OAI

Dr. Bruno Klotz-Berendes
Hochschulbibliothek Münster
klotz-berendes@fh-muenster.de



Gliederung des Workshops

- Einführung in OAI
- Grundlagen des Protokolls
- Data - und Serviceprovider
- Realisierung auf Verbundebene (HBZ)
- Vermarktung des Dokumentenservers



Acknowledgements

- Einige wenige Folien sind von mir, die meisten haben ich aus anderen Vorträgen übernommen:
 - Heinrich Stammerjohanns
 - Uwe Müller
 - Andy Powell
 - Herbert Van de Sompel
 - Carl Lagoze
 - Hussein Suleman
 - Michael Nelson
 - Simeon Warner
 - (and others probably!)



Die Entstehung der Initiative

- Die Wurzeln der OAI sind in den Entwicklungen der Eprint Archive zu suchen.
arXiv, CogPrints, NACA (NASA), NCSTRL
- CrossSuche über viele Archive auf der Basis des Z39.50 Protokolls
 - Cross Search: Dienste laufen verteilt auf verteilten Daten (z.B. Metasuchmaschine, Z39.50)
- Harvesting von Metadaten - Einspielen der Daten von mehreren Archiven in einen zentrale Server mit einem Suchinterface
- OAI konzentriert sich auf Harvesting

Anforderungen an das Harvesting

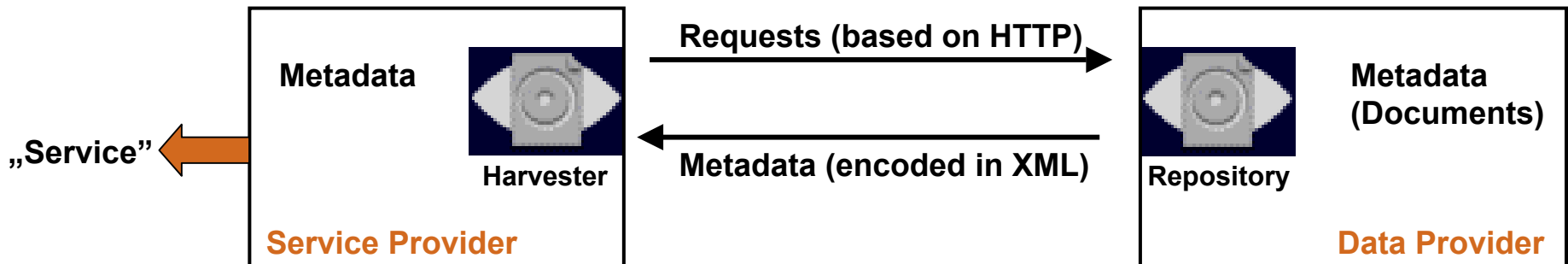
- Damit das Einsammeln der Metadaten funktioniert, müssen Absprachen in den folgenden Bereichen erfolgen:
- Transportprotokoll – HTTP vs. FTP vs. ...
- Metadatenformat – DC vs. MARC vs. ...
- Qualitätskriterien – vertrauenswürdigen Archiv
- Urheberrecht und Verwertungsrechte – Wer darf was mit dem Objekt machen?

The Open Archives Initiative (OAI)

➤ Zentrale Ideen

- weltweite Konsolidierung wissenschaftlicher Archive
- freier Zugang zu den Archiven (mindestens: Metadaten)
- konsistente Schnittstellen für Archive und Service Provider
- “low barrier protocol” / einfache Implementation
- auf existenten Standards basierend (e.g. HTTP, XML, DC)

➤ Grundsätzliche Funktionsweise





Data- und Serviceprovider

- Data Provider
 - Publiziert Inhalte der Community
 - Bietet Metadaten zu den Objekten des Archivs an - Schnittstelle
- Serviceprovider
 - Sammelt die Metadaten von den Data Providern ein
 - Ein Suchinterface für alle Archive, von denen Metadaten eingesammelt wurden.
- Anmerkung:
 - Der Data Provider kann immer noch ein Endnutzersuchinterface anbieten.

Organisationsstruktur von OAI

Steering Committee - besteht aus 12 Vertretern aus verschiedenen wissenschaftlichen Institutionen
politische Weiterentwicklung, richtungsweisende Diskussion und Promotion

- **Executive Committee** - C. Lagoze u.
H. Van de Sompel

Koordination der Aktivitäten

- **Technical Committee** - Evaluierung und Weiterentwicklung der OAI - Architektur, basierend auf Erfahrungen der Anwender

Santa Fe
convention

OAI-PMH
v.1.0/1.1

OAI-PMH
v.2.0

nature

experimental

experimental

stable

verbs

Dienst

OAI-PMH

OAI-PMH

requests

HTTP GET/POST

HTTP GET/POST

HTTP GET/POST

responses

XML

XML

XML

transport

HTTP

HTTP

HTTP

metadata

OAMS

unqualified
Dublin Core

unqualified
Dublin Core

about

eprints

document
like objects

resources

model

metadata
harvesting

metadata
harvesting

metadata
harvesting



Der Name OAI

Open

Archives

Initiative

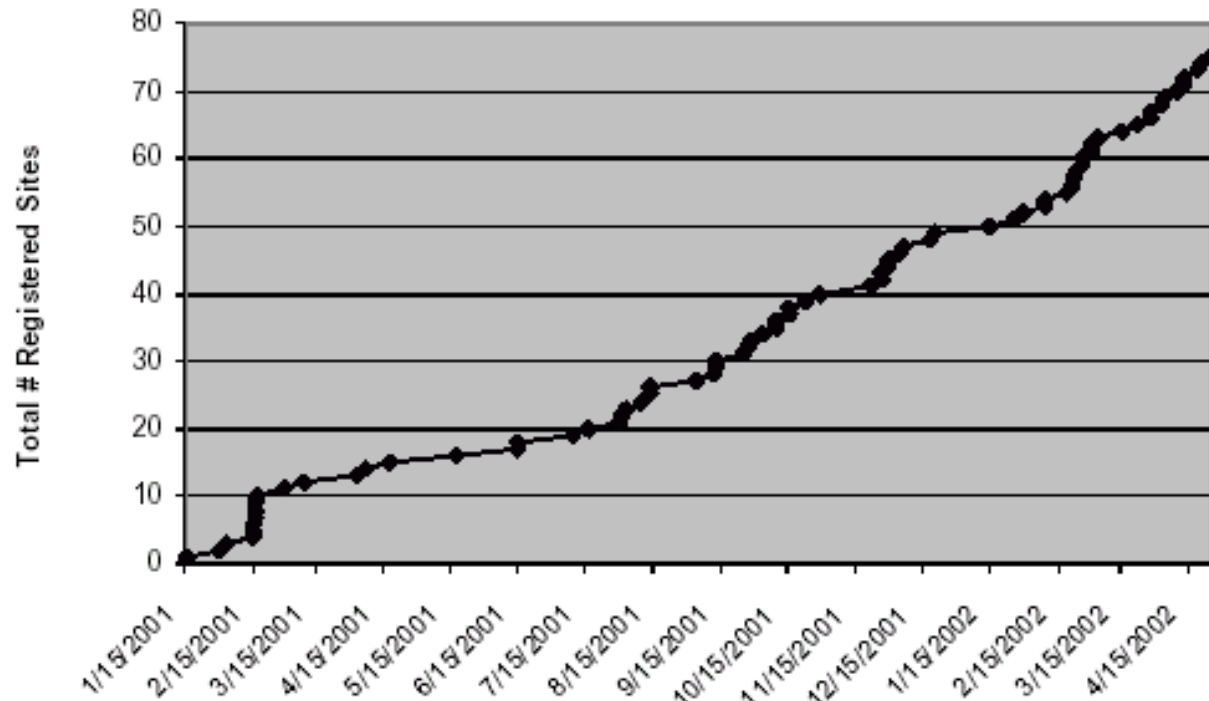
Die Protokollspezifikation ist für jeden zugänglich, und auf die Metadaten darf über die Schnittstelle für OAI-PMH zugegriffen werden. Ein Zugriffsrechtenmanagement für die Objekte ist möglich.

Sammlung digitaler Objekte (nicht Archive). Häufig ist auch der Begriff "Repository" in Gebrauch.

OAI is happening at break-neck speed...



Zunahme der Datenprovider

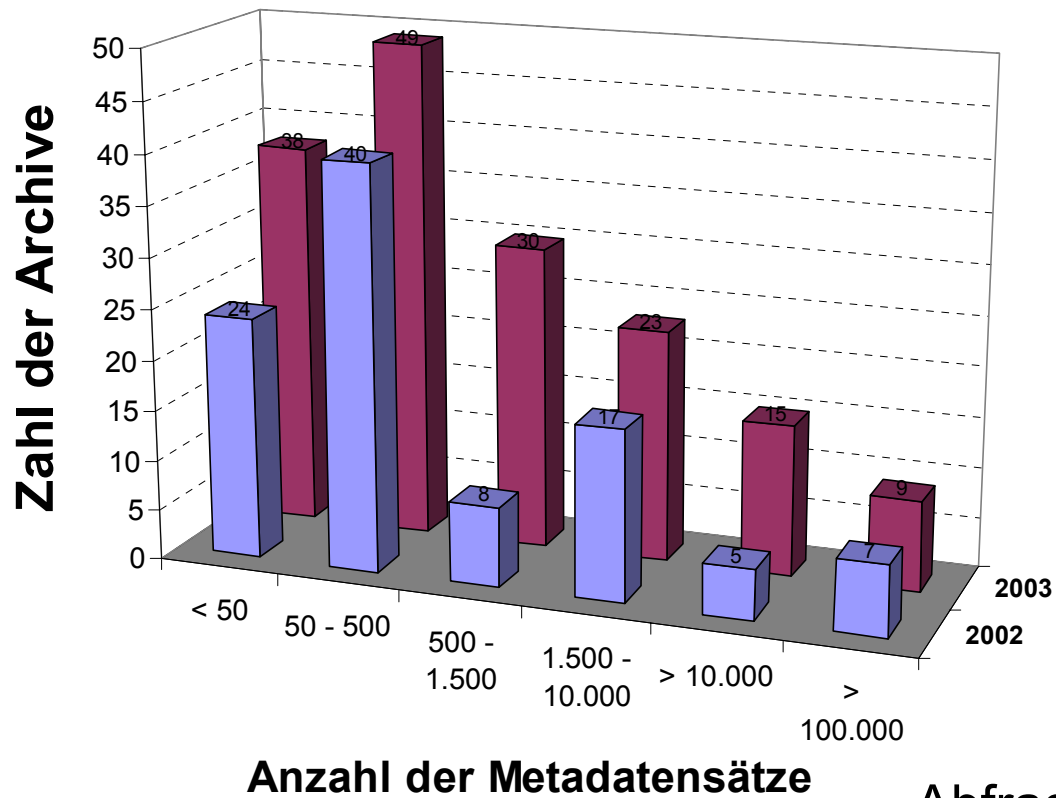


Quelle: H. Van de Sompel, C. Lagoze, Notes from the Interoperability Front:
A Progress Report on the Open Archives Initiative, ECDL 2002, Rom



OAI - Archive

Verteilung der OAI-Archive



Abfrage bei
<http://arc.cs.odu.edu/>

OAI-PMH Version 2.0 - 06/2002

- Ziel: dauerhafter Austausch der Metadaten zwischen Data Providern und Service Providern
- einfache Implementation
- Metadaten Harvesting model: Data Provider / Serviceprovider
- Metadaten der digitalen Objekte (resources)
- Unabhängiges Protokoll
- HTTP basiert
- XML basiert
- mindestens Dublin Core, ohne Qualifier
- Stabil - Produktionsbasis



Vorteile von OAI

- Einfaches Protokoll basierend auf HTTP and XML erlaubt eine zügige Weiterentwicklung
- Unabhängig von der eingesetzten Dokumentenserversoftware
- Verschiedene Serviceprovider können Metadaten von verschiedenen Data Provider einsammeln
- Aggregierende Data Provider dienen als Sammelstelle für kleine Data Provider
- Serviceprovider können ihr Suchinterface mit weiteren Suchen über Z 39.50 ausstatten.



Gliederung des Workshops

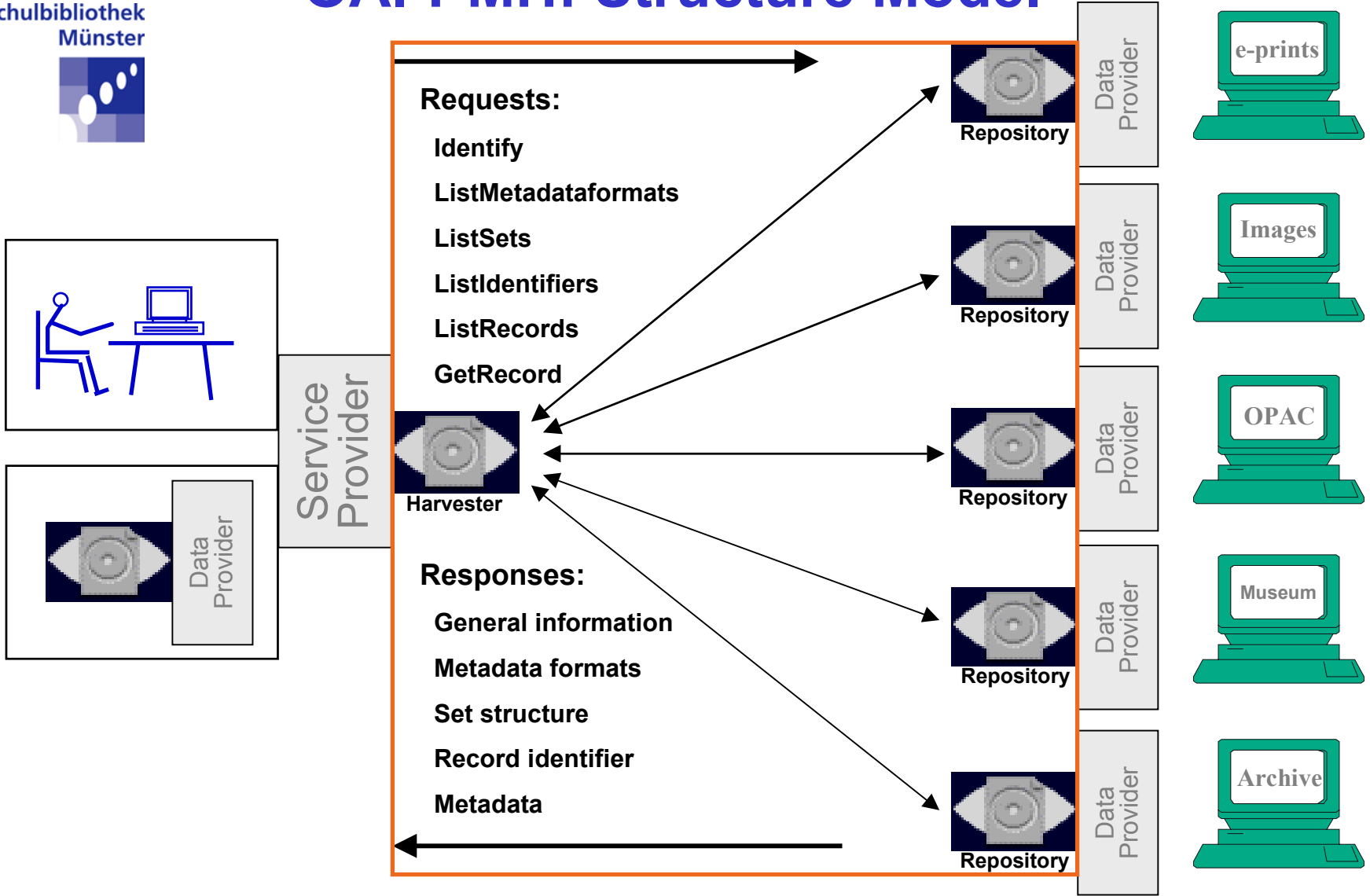
- Einführung in OAI
- **Grundlagen des Protokolls**
- Data - und Serviceprovider
- Realisierung auf Verbundebene (HBZ)
- Vermarktung des Dokumentenservers



OAI: Grundsätzliche Annahmen

- zwei unterschiedliche Gruppen von 'Teilnehmern'
- *Data Provider* (Open Archives, Repositories)
 - freier Zugriff auf Metadaten
 - nicht notwendigerweise: freier Zugriff auf Volltexte / vollständige Objekte
 - einfach zu implementieren, niedrige Barrieren
- *Service Provider*
 - nutzen OAI interface der *Data Provider*
 - sammeln und speichern Metadaten asynchron (keine synchrone verteilte Anfrage)
 - können einige Untermengen von *Data Provider einsammeln* (Mengen (set) Hierarchie, Datestamp)
 - können die Metadaten anreichern
 - bieten Mehrwertdienste, die auf den Metadaten basieren

OAI-PMH: Structure Model



OAI-PMH: Protokoll-Überblick

- Protokoll basiert auf HTTP
- Anfrageargumente als GET- oder POST-Parameter
- sechs Anfragetypen
- z.B. `http://archive.org?`
`verb=ListRecords&from=2002-11-01`
- Antworten werden in XML kodiert
- unterstützt jedes Metadatenformat (mindestens: Dublin Core als kleinster gemeinsamer Nenner)
- logische Mengenhierarchie, die von den Data Providern definiert werden
- Zeitmarken (Datestamps, letzte Änderung der Metadaten)
- Fehlermeldungen (auch in XML)
- Flusskontrolle

Metadatenformate

- OAI-PMH unterstützt das Anbieten / Übertragen mehrerer unterschiedlicher Metadatenformate durch ein Repository
- Repositories müssen mindestens Dublin Core unterstützen
- aber: Es können beliebige Metadatenformate definiert und mit dem OAI-PMH übertragen werden
- Metadaten müssen mit XML Namespace Spezifikation übereinstimmen
- Z.B. MARC (Bibliotheken), IMS (Lehre), ETDMS (Diplomarbeiten/Dissertationen), RFC1807 (Bibliographien)
- Beispiel für ein anderes Metadatenformat
 - Open Language Archives Community [OLAC]
 - <http://www.language-archives.org/OLAC/metadata.html>



Metadatenformate

- Beispiele mit dem Repository Explorer
 - OLAC:
<http://www.language-archives.org/cgi-bin/olaca3.pl?verb=ListMetadataFormats>

- Sehr interessante Seite
 - <http://gita.grainger.uiuc.edu/registry/searchform.asp>

Aussagen der Protokollspezifikation zu Set

- **A *set* is an optional construct for grouping items for the purpose of **selective harvesting**.** Repositories **may** organize items into sets. Set organization **may** be flat, i.e. a simple list, or hierarchical. Multiple hierarchies with distinct, independent top-level nodes are allowed. Hierarchical organization of sets is expressed in the syntax of the `setSpec` parameter as described below. When a repository defines a set organization it **must** include set membership information in the **headers** of items returned in response to the `ListIdentifiers`, `ListRecords` and `GetRecord` requests.



Protokolldetails: Sets

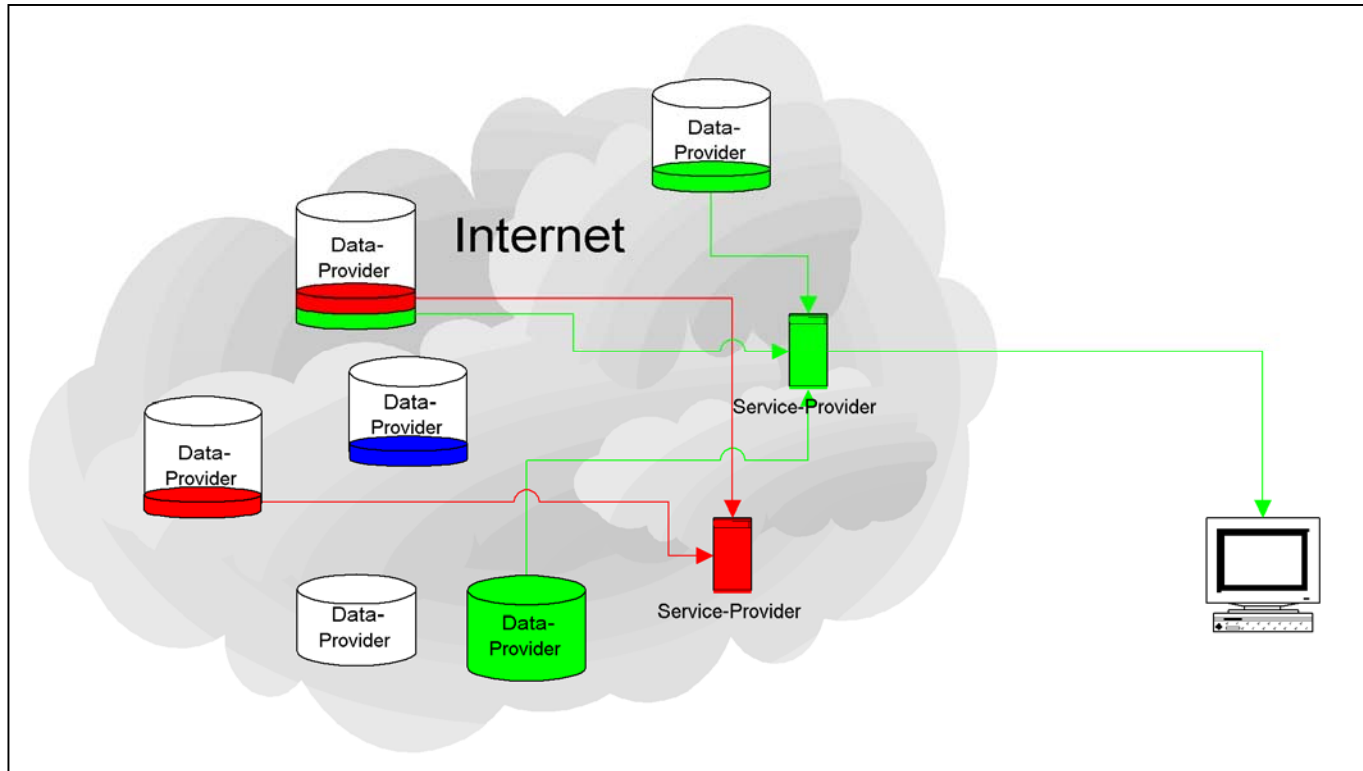
- Das Protokoll unterstützt einen Mechanismus, um Teilkollektionen (sub-collections) einzusammeln
- Keine wohldefinierte Semantik – diese hängt völlig von den lokalen Data Providern ab
- Semantik kann jedoch durch Absprache zwischen Data und Service Providern erreicht werden
- optional – Archive müssen keine Sets definieren
- Anwendungen:
Subject Gateways, Suchmaschine für Dissertationen, ...



Protokolldetails: Sets (2)

- Die DINI-Arbeitsgruppe zu OAI hat eine Empfehlung zur Verwendung von Sets erarbeitet.
<http://www.dini.de/documents/2003-10-08-OAI-Empfehlungen.pdf>
- Publikationstypen (thesis, article, ...)
- Dokumenttypen (text, audio, image, ...)
- inhaltliche Sets, zunächst nach DNB (Medizin, Biologie, ...) jetzt DDC.
- Im HBZ-Verbund sollen Sets verwendet werden.

Benutzung von Sets



Von Bernd Diekmann



Protokolldetails: Flusskontrolle

- Flusskontrolle auf zwei Protokollebenen
 - HTTP (503, retry-after)
 - OAI-PMH, Resumption-Token
- HTTP “retry-after” Mechanismus kann eingesetzt werden, um Anfragen eines Clients zurückzustellen
- Resumption Tokens werden benutzt, um nur Teilantworten zurückzugeben
- Der Client bekommt einen Token, den er für eine neue Anfrage am Server benutzen kann, um weitere Antworten zu bekommen



Data Provider: Flusskontrolle (2)

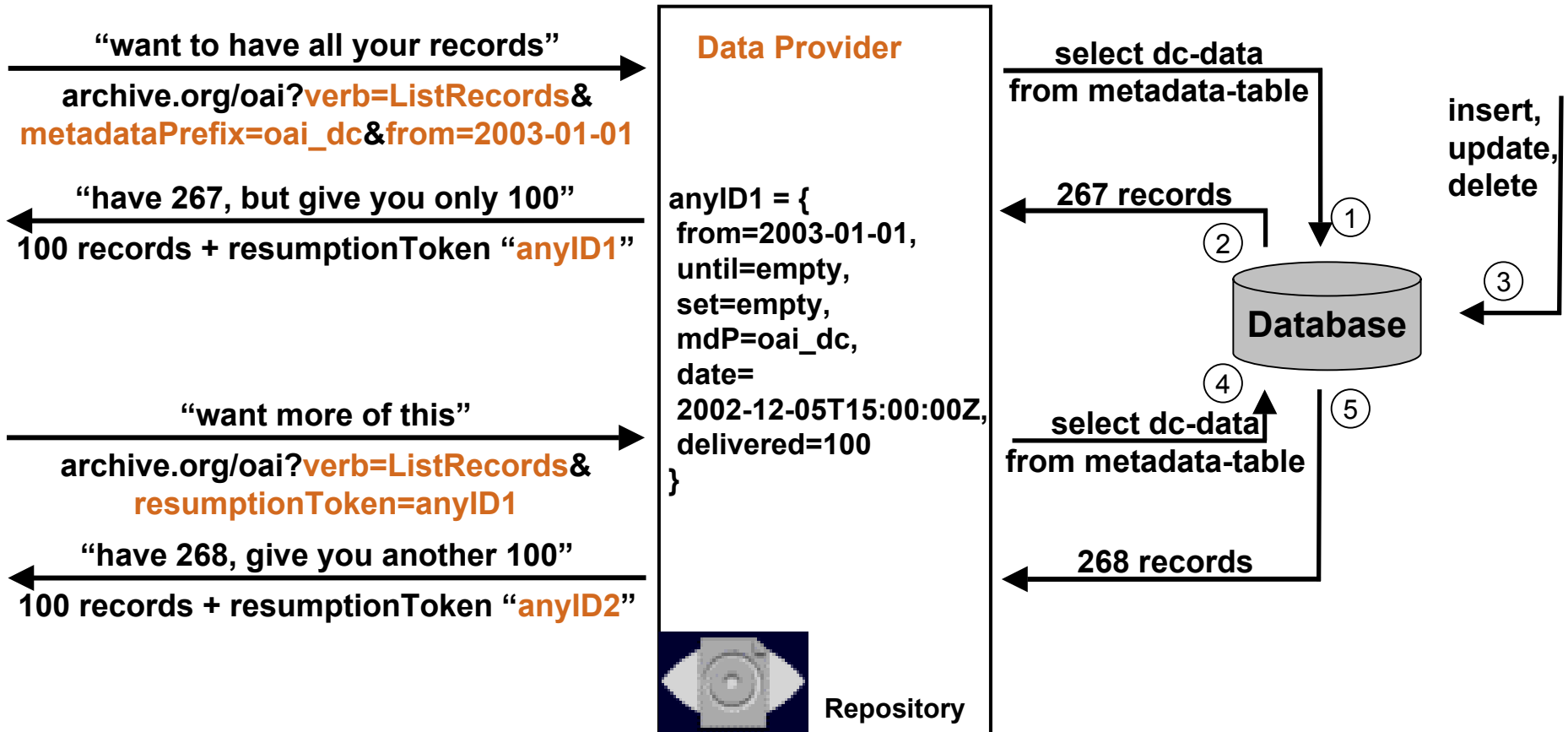
Beispiel





Data Provider: Flusskontrolle (3)

Beispiel (2)



Ausführliche Einführung zu OAI

- Ein online Tutorium ist im Open Archive Forum unter folgender URL zu finden:
 - <http://www.oaforum.org/tutorial>
- Dieses online Tutorium basiert auf mehreren Workshops, die im Rahmen von Tagungen stattgefunden haben.



Gliederung des Vortrages

- Einführung in OAI
- Grundlagen des Protokolls
- **Data - und Serviceprovider**
- Realisierung auf Verbundebene (HBZ)
- Vermarktung des Dokumentenservers



Grundsätzliches: Erste Fragen

Data Provider

- Welche Daten möchte ich anbieten?
- Welchen Service Providern biete ich diese Daten an?

Service Provider

- Welchen Dienst möchte ich anbieten?
- Von welchen Data Providern werde ich die Daten einsammeln?
- Welche Metadatenformate soll ich unterstützen?
- Wie müssen Metadaten weiterverarbeitet werden?

Data Provider & Service Provider

- Auf welche Aspekte muss man sich einigen?
- Metadatenformate ...

Abbilden von Metadaten (Mapping)

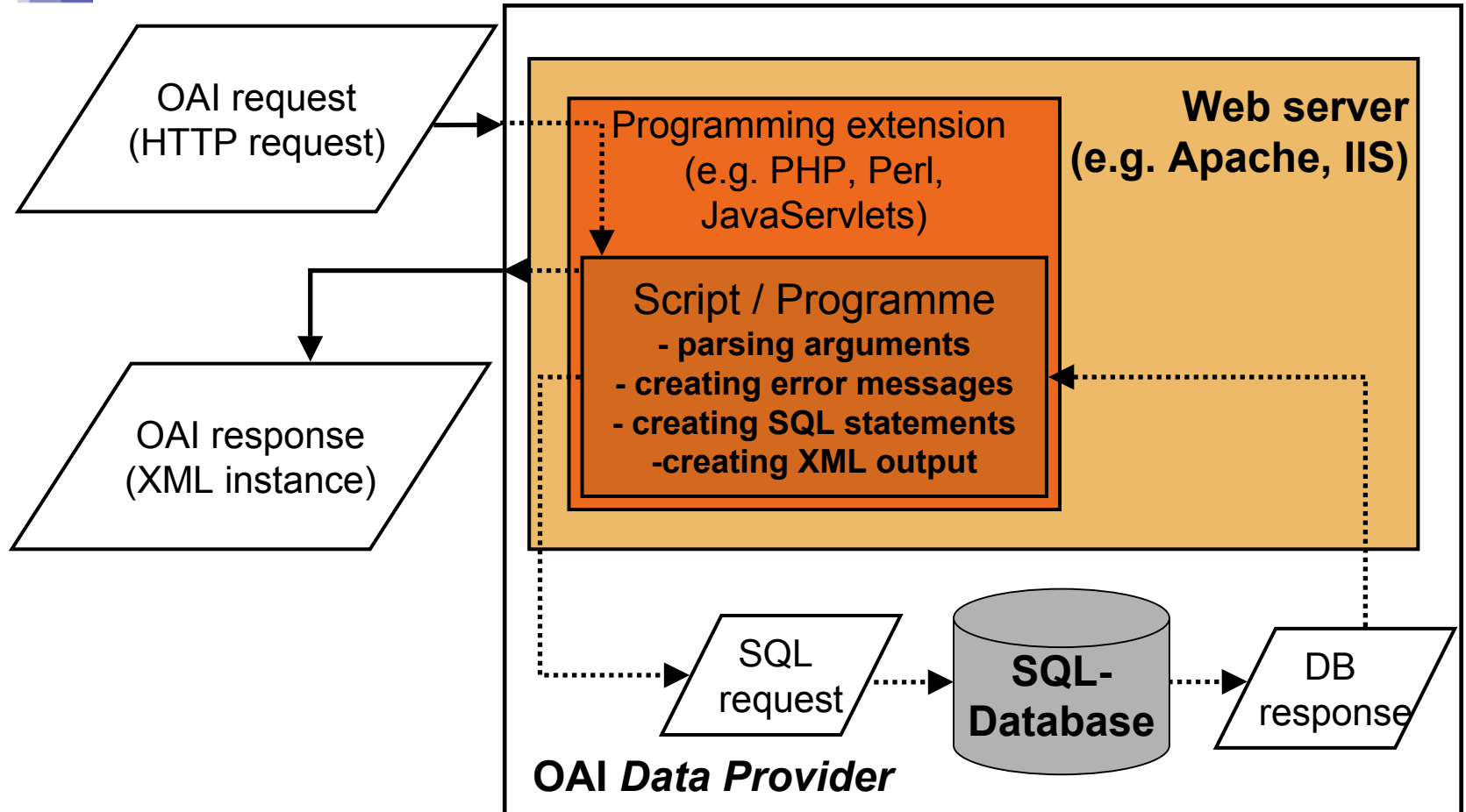
- Data Provider muss seine Metadaten auf das Format, welches er durch das OAI Interface anbietet, abbilden (map).
- Unqualified Dublin Core ist als kleinster gemeinsamer Nenner notwendig
 - Üblicherweise wird ein Link in dem <identifier> Tag zur Resource oder zumindest zu einer lesbaren Webseite angeboten
- Ursprungsformate werden empfohlen
- Metadatenformate der eigenen Community werden empfohlen



Organisation

- Besondere Fächer/Themengebiete/Communities: Andere Metadatenbeschreibungen sind vielleicht notwendig
 - beschreibt Ressourcen in einer besonderen Weise
 - Definition eines eigenen XML-Schemas (welches für Validierung öffentlich zugänglich sein sollte)
- definiere eine Set-Hierarchie
 - um die eigenen Metadaten für “selective harvesting” aufzuteilen
 - Einigung zwischen Data Provider und Service Provider
- zusammengefasste Data Provider
 - wenn ein Service Provider einsammelt, sollte er nicht auch die “sub data providers” befragen (Dubletten)
- Subject Gateways
 - sind bei Einigung auf bestimmte Sets einfach möglich

Data Provider: Architecture





Data Provider: Test und Registrierung

- Erzeugen und Verwenden eigener OAI-PMH-Anfragen – Überprüfen der Ergebnisse
- Benutzen des Repository Explorer (VT University)
 - <http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai/>
 - Eingabe der Argumente über ein HTML-Formular
 - Antworten werden automatisch validiert
 - Möglichkeit des 'Browsing' zu anderen Anfragen
 - automatisches Programm zum Testen der Korrektheit
- offizielle Registrierungsseite
 - <http://www.openarchives.org/data/registerasprovider.html>
 - Angeben der Basis-URL
 - genauer Korrektheitstest (inkl. Fehlermeldungen usw.)
 - Information über inkorrektes Verhalten des Data Providers
 - im Falle des korrekten Verhaltens: Hinzufügen zur offiziellen Liste registrierter Data Provider
 - regelmäßige Überprüfungen

Test des neuen Data Providers

- Nutzen Sie die Hilfen zur Implementation!
 - zu finden unter:
<http://www.openarchives.org/tools/tools.html>

- Testen Sie ihr System, bevor Sie es anmelden mit Hilfe des Repository Explorers
 - Testbeispiele



Serviceprovidertypen

- Man kann inzwischen zwei Typen von Serviceprovidern unterscheiden:
- Reine Suchfunktionalität
Beispiele ARC, Scirus und HU
- Erweiterte Funktionen
Beispiele: Citebase Search, Torii und MyOAI



Service Provider: Beispiele

- Repository Explorer:
 - <http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai/>
- Suchmaschinen / Subject Gateways
 - Cross Archive Searching Service: <http://arc.cs.odu.edu/>
 - DINI: <http://edoc.hu-berlin.de/oaisearch/>
 - Physnet: <http://physnet.uni-oldenburg.de/oai/query.php>
 - NCSTRL: <http://www.ncstrl.org>
- Mehrwertdienste
 - ProPrint: <http://www.proprint-service.de>
 - Citation Indexing: <http://icite.sissa.it:8888>
 - MyOAI: <http://www.myoai.org/>

Service Provider: Beispiele (2)

- Suchbeispiel ARC

- MyOAI
 - SDI-Suchen
 - E-Mailbenachrichtigung
 - Annotation bei gespeicherten Literaturverweisen

Service Provider: Struktur (1)

Archiv-Management

- Auswahl der Archive von den gesammelt werden soll
- manuelle Eingabe oder
- automatische Hinzufügung/Löschung von Archiven mittels der offiziellen Registry

Anfrage-Komponente

- erzeugt HTTP Anfragen und sendet sie an OAI-Archive (Data Provider)
- verlangt Metadaten mittels OAI-PMH
- möglicherweise selective harvesting (**set-Parameter**)



Service Provider: Struktur (2)

Scheduler

- sorgt für regelmäßige Abfragen von den Archiven
- einfachster Fall: manueller Start
- sonst: z.B. cron job, relationale Datenbank ...

Flusskontrolle

- Resumption-Token: weitere Anfragen bei Rückgabe eines Resumption-Tokens
- HTTP-Fehler 503 (service not available) – Analyse der Antwort, um das Archiv nach “retry-after” Zeitraum erneut anzufragen

Service Provider: Struktur (3)

Update-Mechanismus

- fügt alte und neue Daten zusammen (oder ersetzt diese)
- einfachster Fall: lösche alle Einträge mit “alten” Metadaten, bevor diese von einem Archive eingesammelt werden
- besser: inkrementelle Aktualisierung (**from** parameter) – füge *neue* Metadaten ein und überschreibe *geänderte / gelöschte* Metadaten (anhand der eindeutigen Identifier)

XML-Parser

- analysiert die Antworten von den Archiven
- Validierung anhand des XML-Schemas
- extrahiert die Metadaten
- transformiert die Metadaten in eine interne Datenstruktur



Service Provider: Struktur (4)

Normalisierer und Mapper

- normalisiert die Darstellung (z.B. Datum, Autor, Sprachcode)
- transformiert die Daten in eine homogene Struktur (bei unterschiedlichen Metadaten-Formaten)

Datenbank

- Abbildung der XML-Struktur der Metadaten in eine relationale Datenbank (Mehrfach-Werte)
- oder: Nutzung einer XML-Datenbank



Service Provider: Struktur (5)

Dubletten-Checker

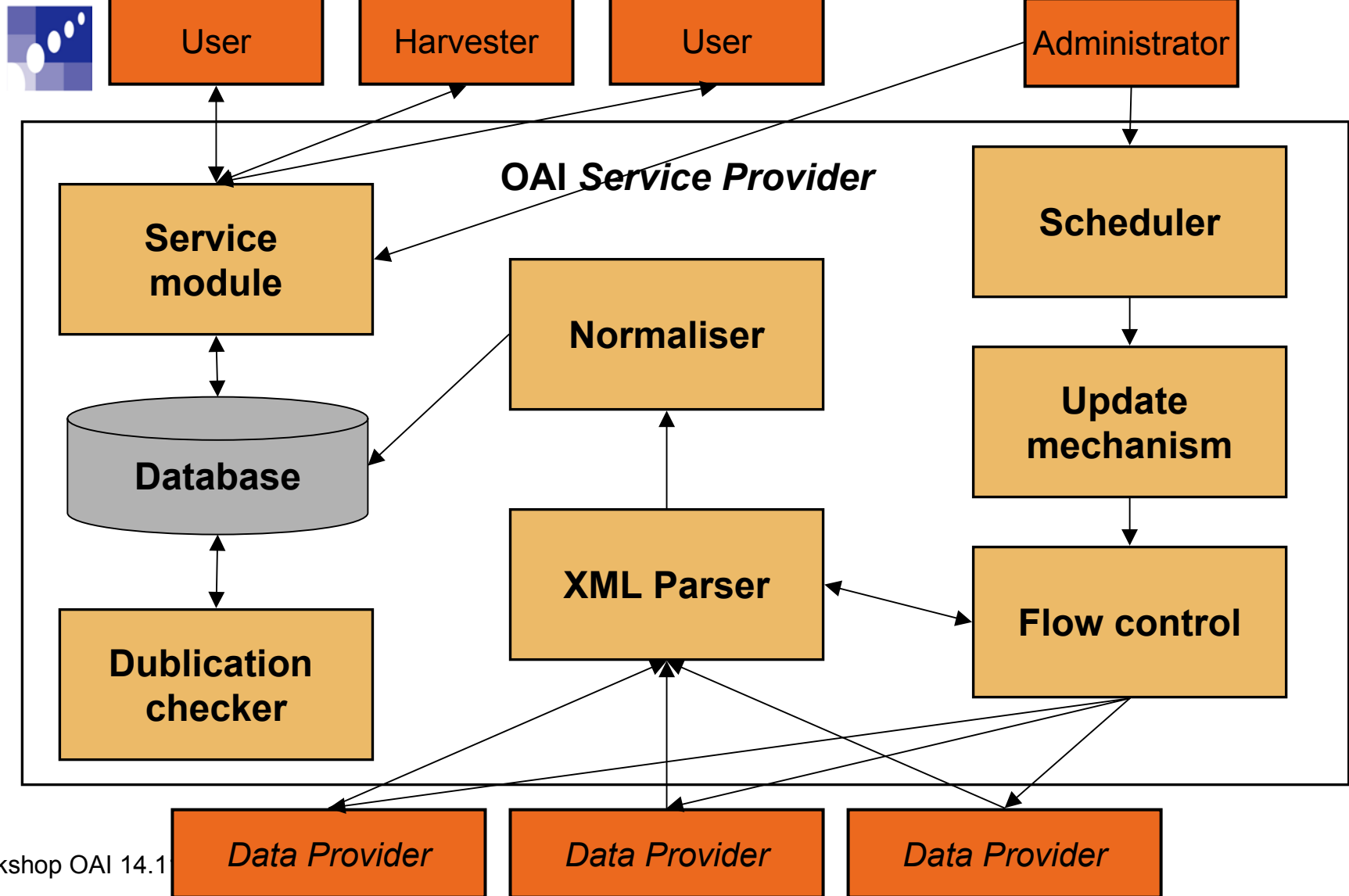
- führt identische Records von verschiedenen Data Providern zusammen
- z.B.: eindeutiger Identifier für ein Item (z.B. URN, ...)
- jedoch: oft nicht einfach zu handhaben und nicht fehlerfrei

Service-Modul / Dienst

- bietet den Dienst für die “Öffentlichkeit” an
- Basis: eingesammelte und gespeicherte Records der Archive
- benutzt ausschliesslich die lokale Datenbank für Suchen



Serviceprovider: Architektur



Serviceprovider: OAI Communities

- gemeinsame Metadatenformate
 - z.B. E-Print-Format, ETD-MS, VRA Core, IMS
- gemeinsame Semantik
 - kontrolliertes Vokabular für Felder
 - spezifische Felder für externe Links (DC: Identifier)
- Geschlossene OAI-Netzwerke
 - Transfer von Daten zwischen heterogenen Systemen innerhalb einer Organisation
 - globale Optimierungen möglich (Sets, Metadatenformate)
- OAI innerhalb von Digital Libraries (DL)
 - z.B. Browsing über Sets
 - Reviews, Annotations können unabhängige OAI-Data Provider sein
 - Open Digital Library Project: <http://oai.dlib.vt.edu/odl>



Gliederung des Vortrages

- Einführung in OAI
- Grundlagen des Protokolls
- Data - und Serviceprovider
- **Realisierung auf Verbundebene (HBZ)**
- Vermarktung des Dokumentenservers



Verbundlösung in NRW

Empfehlungen für die
Dokumentenserverbetreiber und dem
Hochschulbibliothekszenrum in Köln

**Expertengruppe aus NRW und
Rheinland -Pfalz**



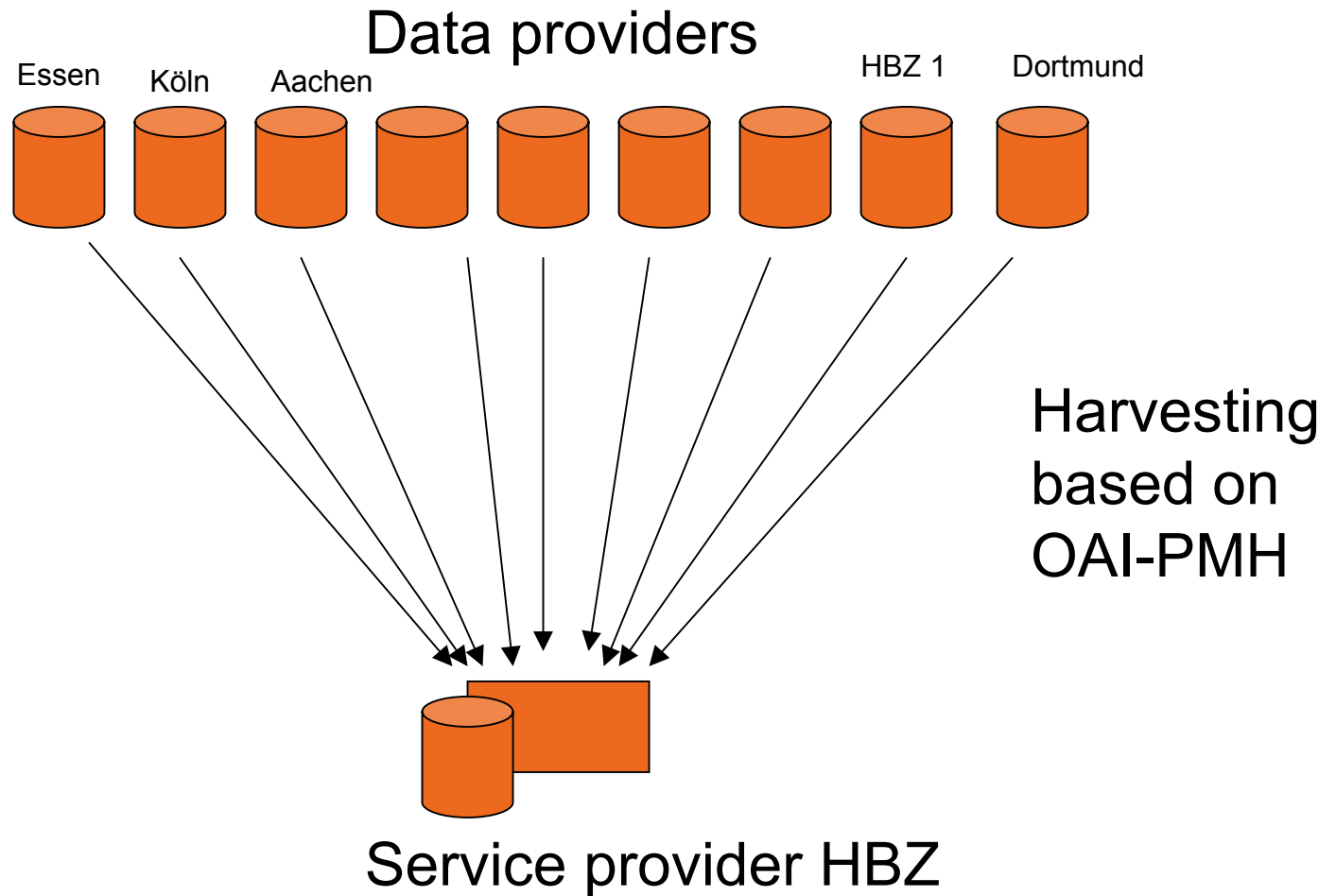
Ausgangslage

- Jede Bibliothek / Hochschulrechenzentrum betreibt einen eigenen Dokumentenserver
- Die Digitale Bibliothek NRW wird vom Hochschulbibliothekszenrum (HBZ) für alle Bibliotheken mit einer lokalen Sicht betrieben.
- Der erste Versuch einer gemeinsamen Dokumentenserversuche ist gescheitert.

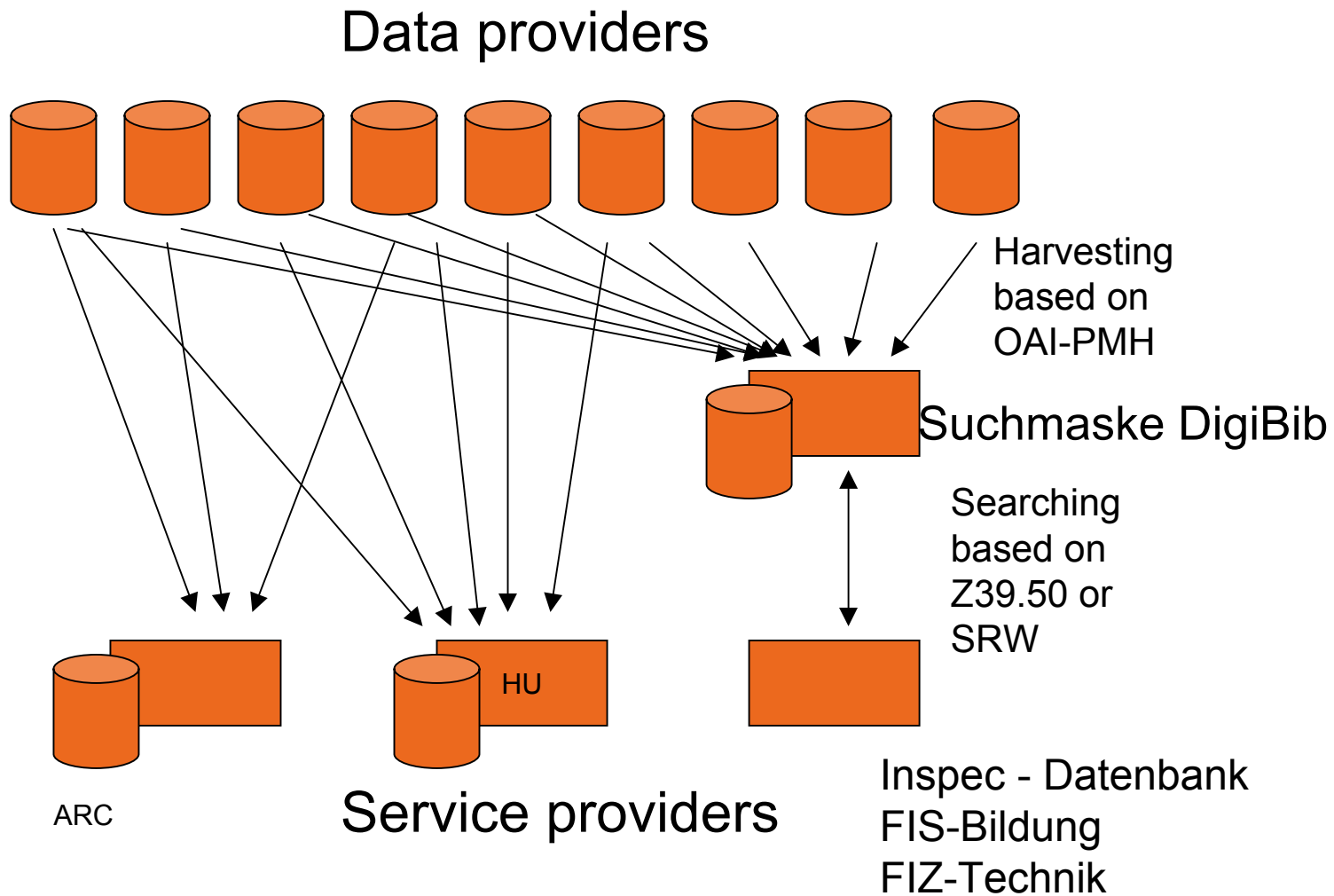
OAI ist die Lösung

- Die Expertengruppe hat sich auf die folgenden Punkte verständigt:
- OAI erlaubt jeder Institution den Betrieb einer eigenen Lösung für den Dokumentenserver
- Einige Dokumentenserver verfügen bereits über eine solche Schnittstelle
- Hilfestellung bei der Implementierung der Schnittstelle durch die OAI-Community
- Support des HBZs für OPUS

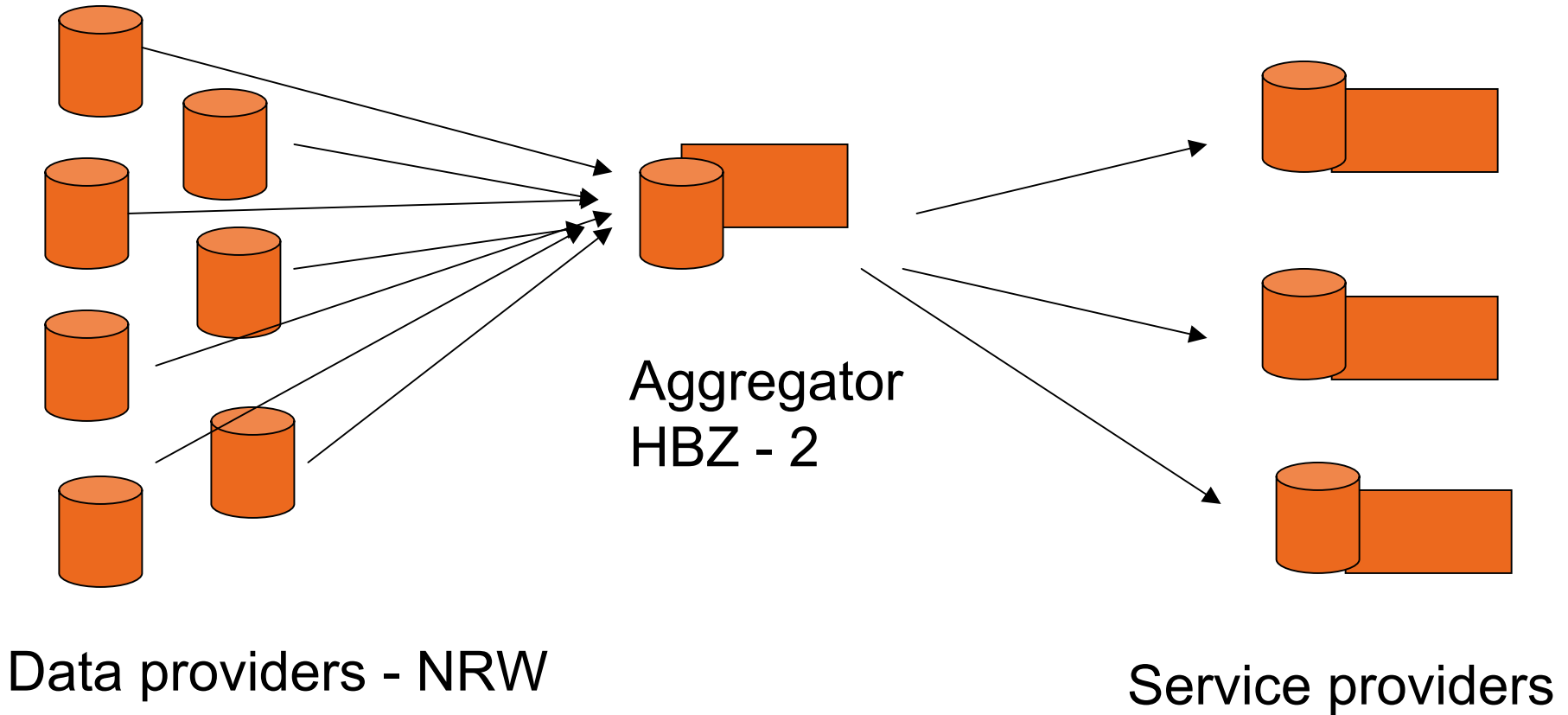
Data Provider und SP HBZ



Integration in die DigiBib



Aggregierender DP im HBZ



OAI ist die Lösung

- Aufbau eines aggregierenden Datenproviders beim HBZ
- Ausweitung der SP – Funktionen und Archive
- Nutzen wir unsere Verbundstrukturen für den Aufbau neuer Strukturen auf Dokumentenserverebene, um die wissenschaftlichen Dokumente unserer Hochschulen in der internationalen Wissenschaft bekannt zu machen.



Gliederung des Vortrages

- Einführung in OAI
- Grundlagen des Protokolls
- Data - und Serviceprovider
- Realisierung auf Verbundebene (HBZ)
- **Vermarktung des Dokumentenservers**



Inhalte einwerben

- Dissertationen sind nicht das Problem

- Wie erreiche ich die anderen Autoren der Hochschule?
 - Jeder Autor muss einen persönlichen Nutzen davon haben, dass er seine Dokumente auf den Dokumentenserver einspielt
 - Die Universität (Rektorat) muss einen Nutzen darin sehen, dass die Publikationen der Hochschule an einer zentralen Stelle bereitgestellt werden.



Inhalte einwerben (2)

- Zusätzliche Dienste entwickeln
 - Print on demand / on CD
 - Meldung VG-Wort
 - rechtliche Betreuung
 - Autorenbetreuung
 - Publikationslisten (auf CD) erstellen
 - Betreuung der Metadaten (Status der Veröffentlichung)
 - Einreichen der Veröffentlichung
 -

- Zertifizierter Dokumentenserver

Inhalte einwerben (3)

Rektorat

- Hochschulbibliographie
- Forschungsreports
- Erfolgsparemetermodell im Etat



Verbündete suchen

- Budapest Open Access Initiative
 - <http://www.soros.org/openaccess/g/commitment.shtml>
 - 3028 Einzelpersonen u. 255 Institutionen

- Public Library of Science (PLoS)
 - <http://www.plos.org/>

- Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities
Aufnahme in die Grundsätze zur Sicherung guter wissenschaftlicher Praxis der DFG?

Beispiel System + Marketing

- Beispiel DSpace
 - Building your DSpace system is only part of the challenge you face in implementing DSpace. You need to market DSpace to let faculty, library staff, administrators, and users know what DSpace can do for them.

- zu finden unter:
<http://dspace.org/implement/market.html>

Zusammenfassung

- Unabhängig von der Dokumentenserversoftware
- Kostengünstiger Metadatentransfer vom Datenprovider zum Serviceprovider
- basiert auf HTTP und XML – Web – friendly
- Flexible Anpassung der Metadaten
- Steuerung des Datenflusses zwischen DP u. SP (Tokenfunktion)
- Aggregierende DP
- Umfangreiche Implementationshilfen



Zusammenfassung (2)

- Mindestens DC simpel als Metadatenformat, aber offen für alle anderen Formate, die in XML encoded sind.
- OAI-PMH ist kein Endnutzersuchprotokoll
- Metadaten und Volltext sind üblicherweise frei zugänglich – Volltexte müssen es aber nicht sein.
 - OAI-PMH kann auch innerhalb geschlossener Gruppen benutzt werden.
- Zugriffskontrolle basiert auf dem zugrunde liegenden HTTP Protokoll



Wichtige Ressourcen

- OAI Web site:
 - <http://www.openarchives.org/>
- OAI-PMH specification:
 - <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- Implementation guidelines:
 - <http://www.openarchives.org/OAI/2.0/guidelines.htm>
- Discussion lists:
 - <http://www.openarchives.org/mailman/listinfo/oai-general>
 - <http://oaisrv.nsd.cornell.edu/mailman/listinfo/oai-implementers>
- Repository explorer:
 - <http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai>
- Tools: <http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai>



Danksagung

Meinen Mitstreitern in der DINI Arbeitsgruppe
Susanne Dobratz HU Berlin,
Uwe Müller HU Berlin,
Frank Scholze UB Stuttgart,
Bernd Diekmann BIS Oldenburg und
Heinrich Stamerjohanns, Universität Oldenburg

allen, bei denen ich mir Folien „entliehen“ habe,
und Ihnen danke ich für Ihre Aufmerksamkeit !