

# Distributed Archive Networks in the Open Archives Initiative

Heinrich Stamerjohanns

[stamer@uni-oldenburg.de](mailto:stamer@uni-oldenburg.de)

Eberhard R. Hilf

[hilf@scineo.de](mailto:hilf@scineo.de)

Thomas Severiens

[severien@uni-oldenburg.de](mailto:severien@uni-oldenburg.de)

Institute for Science Networking, Oldenburg

Susanne Dobratz

[dobratz@rz.hu-berlin.de](mailto:dobratz@rz.hu-berlin.de)

Uwe Müller

[u.mueller@rz.hu-berlin.de](mailto:u.mueller@rz.hu-berlin.de)

Computing Centre, Humboldt-University Berlin

# Implementing OAI at German Universities

- DINI: (<http://www.dini.org>)
  - German Initiative for Networked Information
  - carries out guidance for implementations all over Germany
  - develop a strategy to cover German universities (libraries with document servers)
- Aim:
  - Serving a distributed archive network
  - Setting up a contact point for OAI in Germany



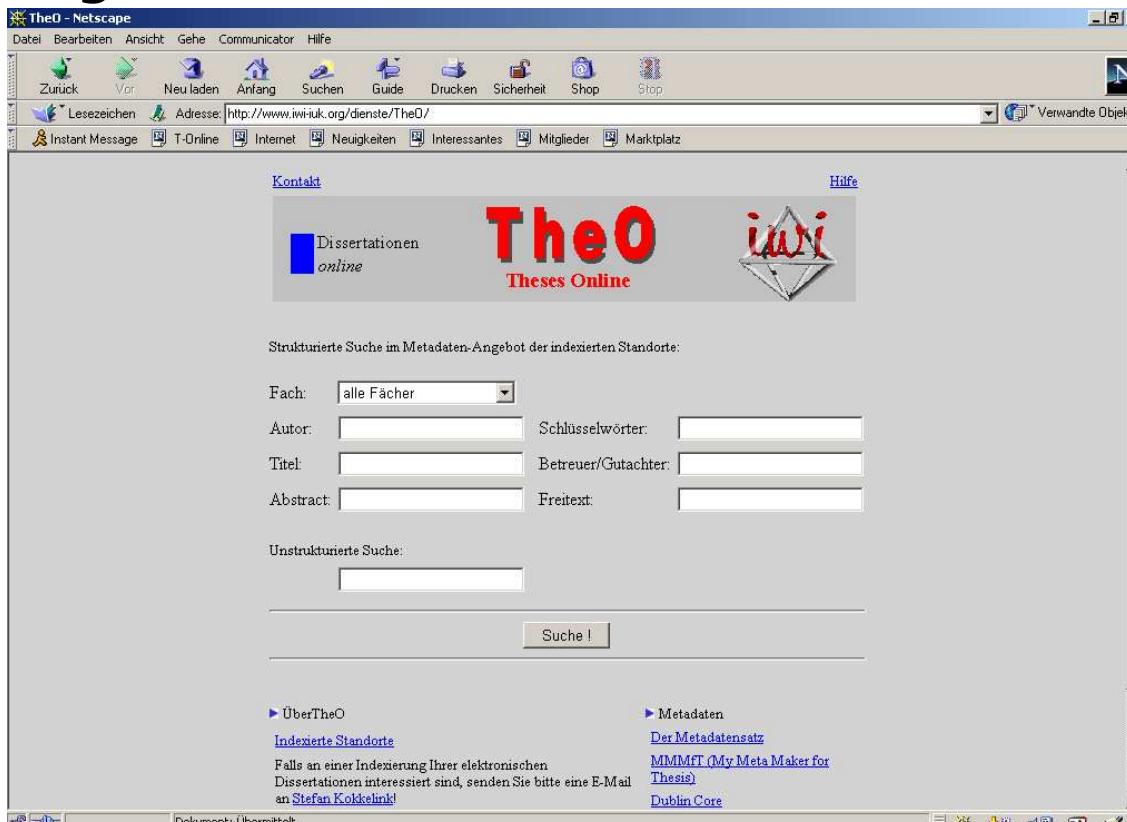
# Collections Represented

- PhysDoc:
  - Distributed document Database for Physics worldwide
  - using HARVEST as Retrieval mechanism
- Universities document servers
  - Humboldt-University as one example
  - small number of documents (up to 500)
  - document formats: PDF
  - DissOnline.de Initiative
  - Part of NDLTD



# Why OAI?

- DissOnline.de Retrieval interface TheO
  - <http://www.iwi-iuk.org/dienste/TheO/>
  - using Dublin Core Set for Theses and Diss.



# Why OAI?

- Interface to NDLTD and others
- Using Dublin Core Metadata Set for Theses and Dissertations
- Platform for integration
  - within subject specific gateways as PhysDoc, MathDiss,
  - Within local services (German library consortia)

# Context & Motivation

- OAI – different strategies emerge
  - incorporate existing repositories by implementing OAI compliant wrapper
  - fits for larger centers, institutions with active computer/library centers
  - unrealistic for large number of small places or even individual authors

# Strategy for OAI Compliance

- Most dissertation archives use
  - Harvest for Retrieval / HTML-Titlepages / HTML coded Dublin Core or
  - **Databases** for metadata storage
    - Sybase (Humboldt-University)
    - Oracle
    - mysql ...
- Implementing OAI protocol
  - scripts: perl / **php4** or 3
  - supporting **Dublin Core**

# OAI Compliance

- Humboldt University of Berlin, GERMANY, Document Server
  - OAI 1.0 compliant repository
  - since Feb. 16th 2001
  - <http://dochost.rz.hu-berlin.de/OAI-script>
    - ?verb=identify/ListRecords/...
  - original implementation (Uwe Müller) took a few hours
  - some problems with XML encoding
    - (UNICODE/UTF-8) not (ISO-8859-1) of German „Umlaute“ ä ö ü ß and XML-characters < >
    - -> php4 library for conversion used

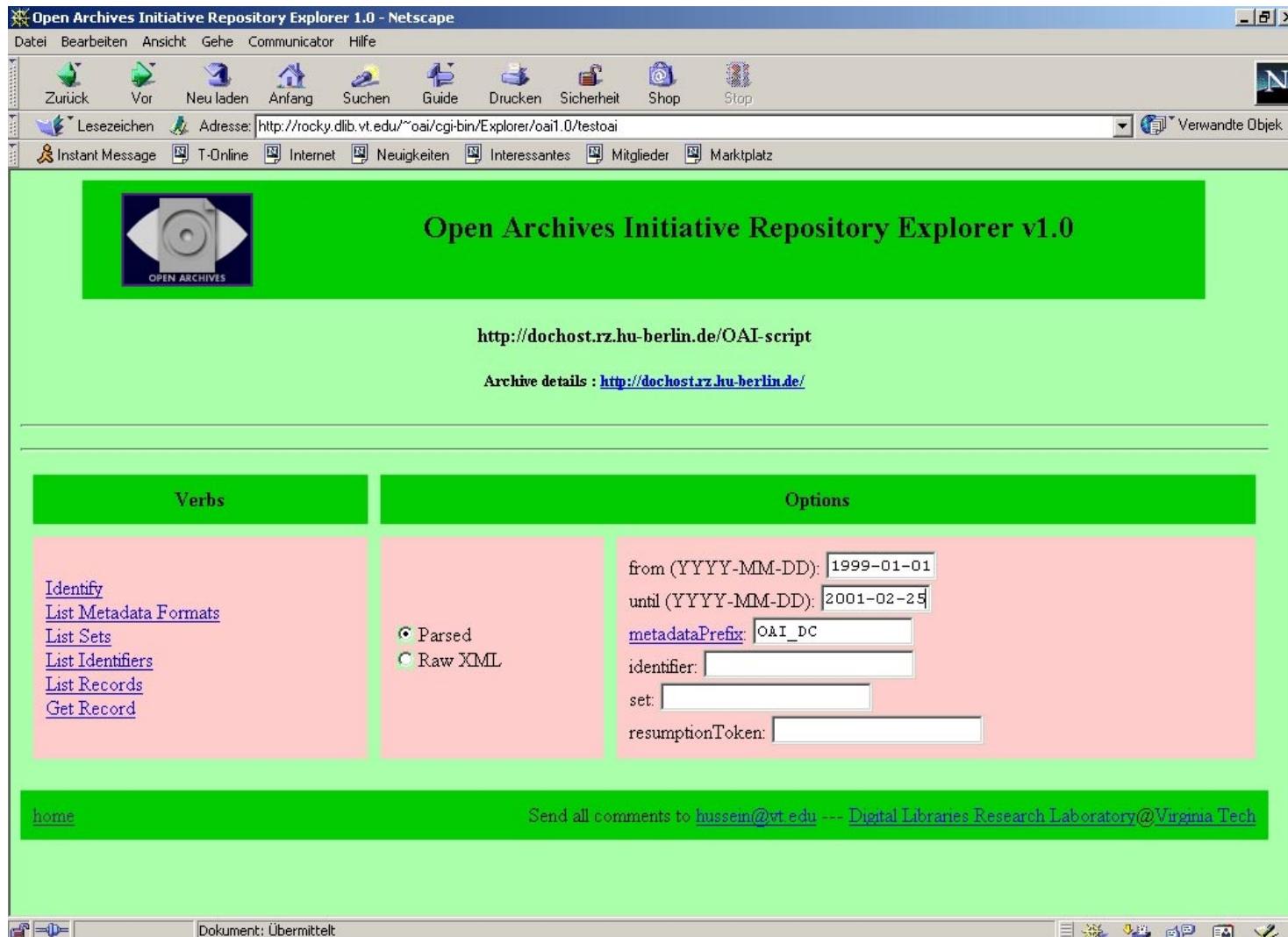
# OAI Compliance of HUBerlin

- Header
  - Unique Identifier: HUBerlin
  - Datestamp: response date
- protocol requests implemented:
  - GetRecord
  - Identify
  - ListIdentifiers
  - ListMetadataFormats: OAI\_DC
  - ListRecords
  - ListSets
- Resumption token (100 items)

# OAI \_Identify

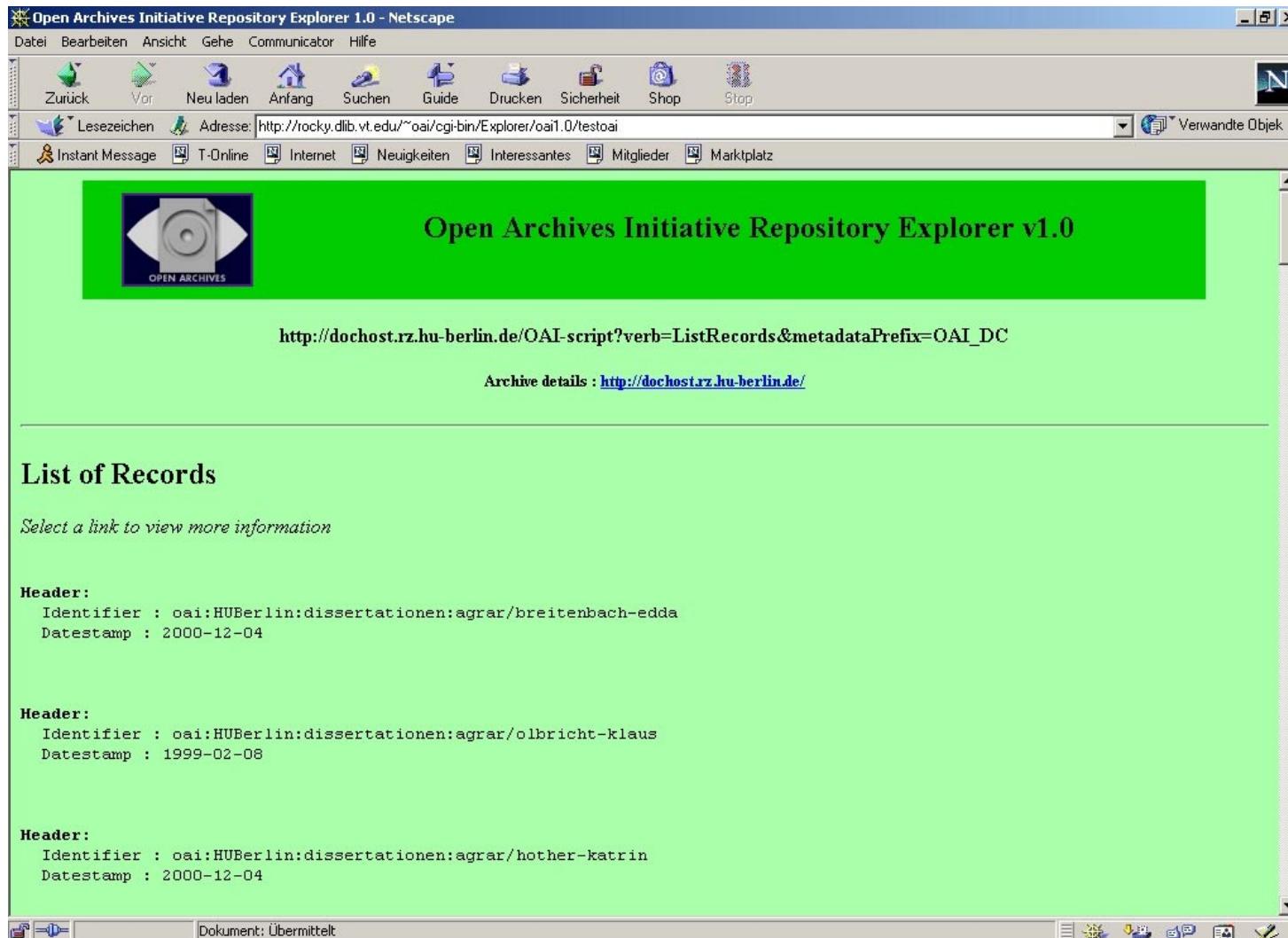
```
<?xml version="1.0" encoding="UTF-8" ?>
- <Identify xmlns="http://www.openarchives.org/OAI/1.0/OAI_Identifier"
  xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/1.0/OAI_Identifier
  http://www.openarchives.org/OAI/1.0/OAI_Identifier.xsd">
  <responseDate>2001-02-25T13:45:15+01:00</responseDate>
  <requestURL>http%3A%2F%2Fdochost.rz.hu-berlin.de%2FOAI-script%3Fverb%
    3DIdentifier</requestURL>
  <repositoryName>Humboldt University of Berlin, GERMANY, Document Server</repositoryName>
  <baseURL>http://dochost.rz.hu-berlin.de/OAI-script</baseURL>
  <protocolVersion>1.0</protocolVersion>
  <adminEmail>mailto:oai@rz.hu-berlin.de</adminEmail>
- <description>
  - <oai-identifier xmlns="http://www.openarchives.org/OAI/oai-identifier"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/oai-identifier
    http://www.openarchives.org/OAI/oai-identifier.xsd">
    <scheme>oai</scheme>
    <repositoryIdentifier>HUBerlin</repositoryIdentifier>
    <delimiter>:</delimiter>
    <sampleIdentifier>oai:HUBerlin:dissertationen:kemps-christoph-2000-06-
      18</sampleIdentifier>
  </oai-identifier>
</description>
</Identify>
```

# OAI\_ListRecords / Resumption Token



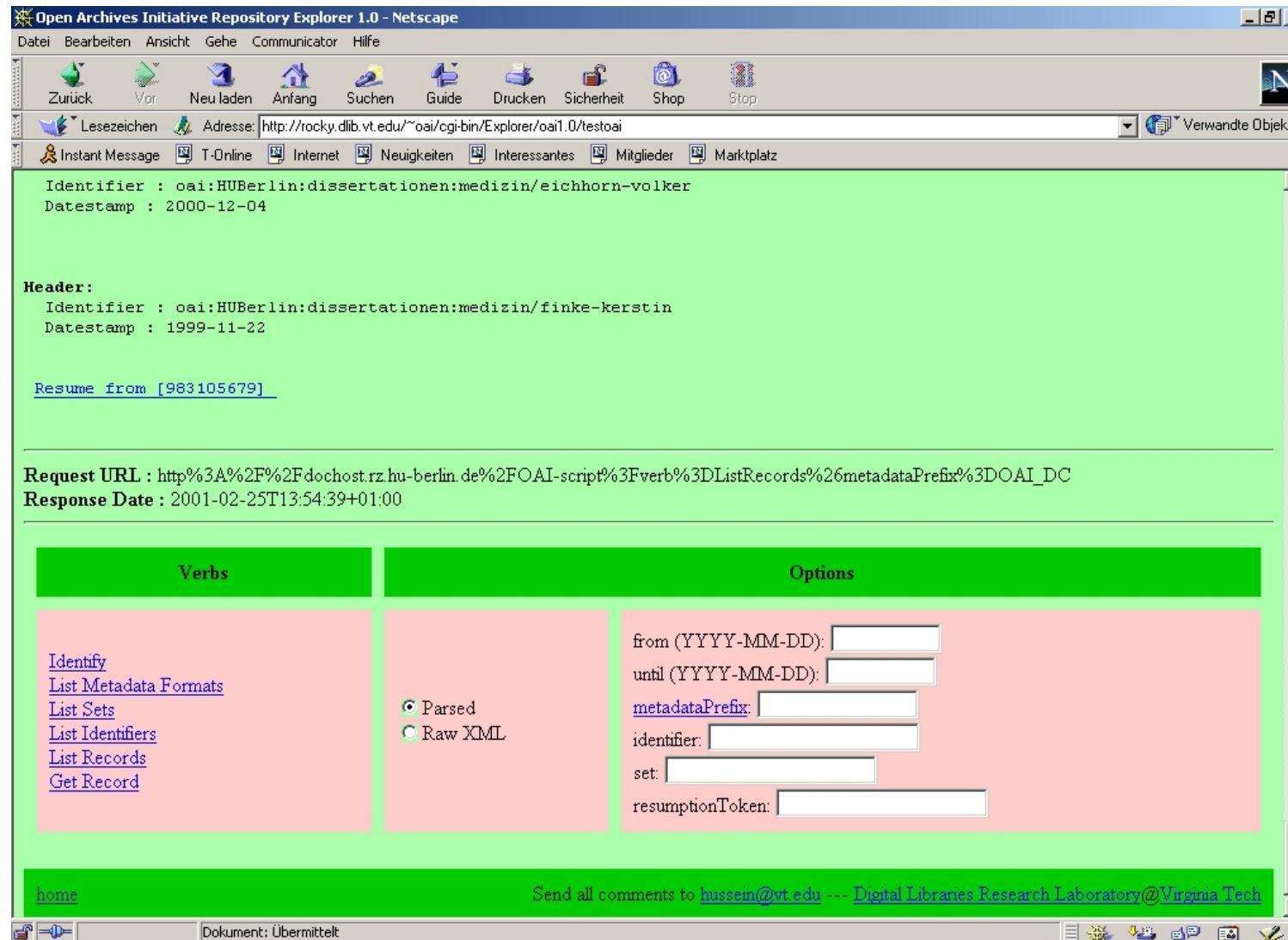
Using Virginia Tech Repository Explorer at [http://purl.org/net/oai\\_explorer](http://purl.org/net/oai_explorer)  
February 26th 2001 Heinrich Stamerjohanns & Susanne Dobratz 11

# OAI\_ListRecords / Resumption Token



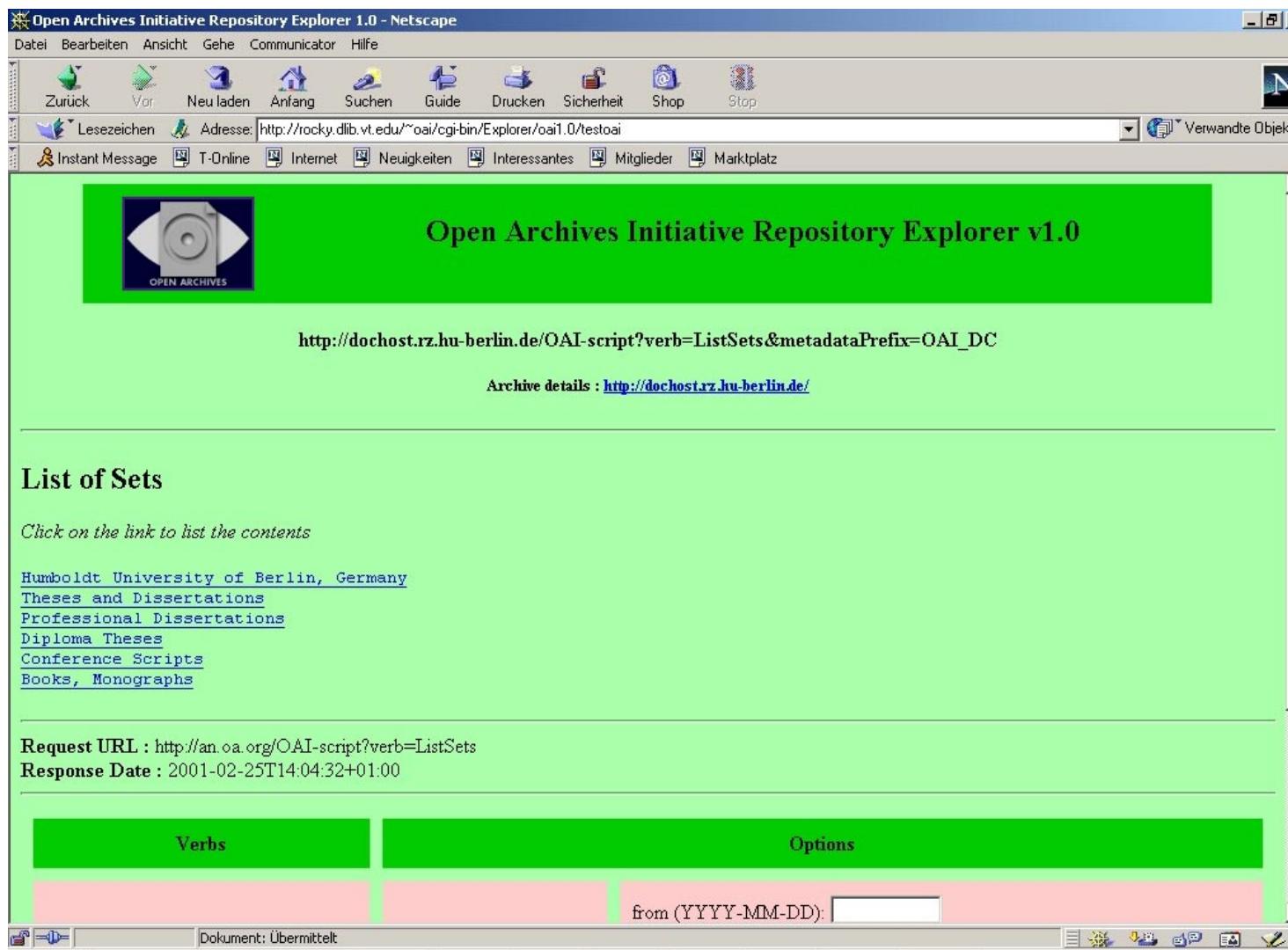
Using Virginia Tech Repository Explorer at [http://purl.org/net/oai\\_explorer](http://purl.org/net/oai_explorer)  
February 26th 2001 Heinrich Stamerjohanns & Susanne Dobratz 12

# OAI\_ListRecords / Resumption Token

A screenshot of the Open Archives Initiative Repository Explorer 1.0 interface, displayed in a Netscape browser window. The title bar reads "Open Archives Initiative Repository Explorer 1.0 - Netscape". The menu bar includes "Datei", "Bearbeiten", "Ansicht", "Gehe", "Communicator", and "Hilfe". The toolbar contains icons for "Zurück", "Vor", "Neu laden", "Anfang", "Suchen", "Guide", "Drucken", "Sicherheit", "Shop", and "Stop". The address bar shows the URL "http://rocky.dlib.vt.edu/~oai/cgi-bin/Explorer/oai1.0/testoai". Below the address bar is a toolbar with links for "Lesezeichen", "Adresse", "Verwandte Objekt", "Instant Message", "T-Online", "Internet", "Neugkeiten", "Interessantes", "Mitglieder", and "Marktplatz". The main content area displays the following text:  
Identifier : oai:HUBerlin:dissertationen:medizin/eichhorn-volker  
Datestamp : 2000-12-04  
  
**Header:**  
Identifier : oai:HUBerlin:dissertationen:medizin/finke-kerstin  
Datestamp : 1999-11-22  
  
[Resume from \[983105679\]](#)  
  
Request URL : http%3A%2F%2Fdochostrz.hu-berlin.de%2FOAI-script%3Fverb%3DListRecords%26metadataPrefix%3DOAI\_DC  
Response Date : 2001-02-25T13:54:39+01:00  
  
The interface has two main sections: "Verbs" (left) and "Options" (right). The "Verbs" section contains links: Identify, List Metadata Formats, List Sets, List Identifiers, List Records, and Get Record. The "Options" section contains input fields for "from (YYYY-MM-DD)", "until (YYYY-MM-DD)", "metadataPrefix", "identifier", "set", and "resumptionToken". At the bottom, there are links for "home" and "Send all comments to [hussein@vt.edu](mailto:hussein@vt.edu) --- Digital Libraries Research Laboratory@Virginia Tech".

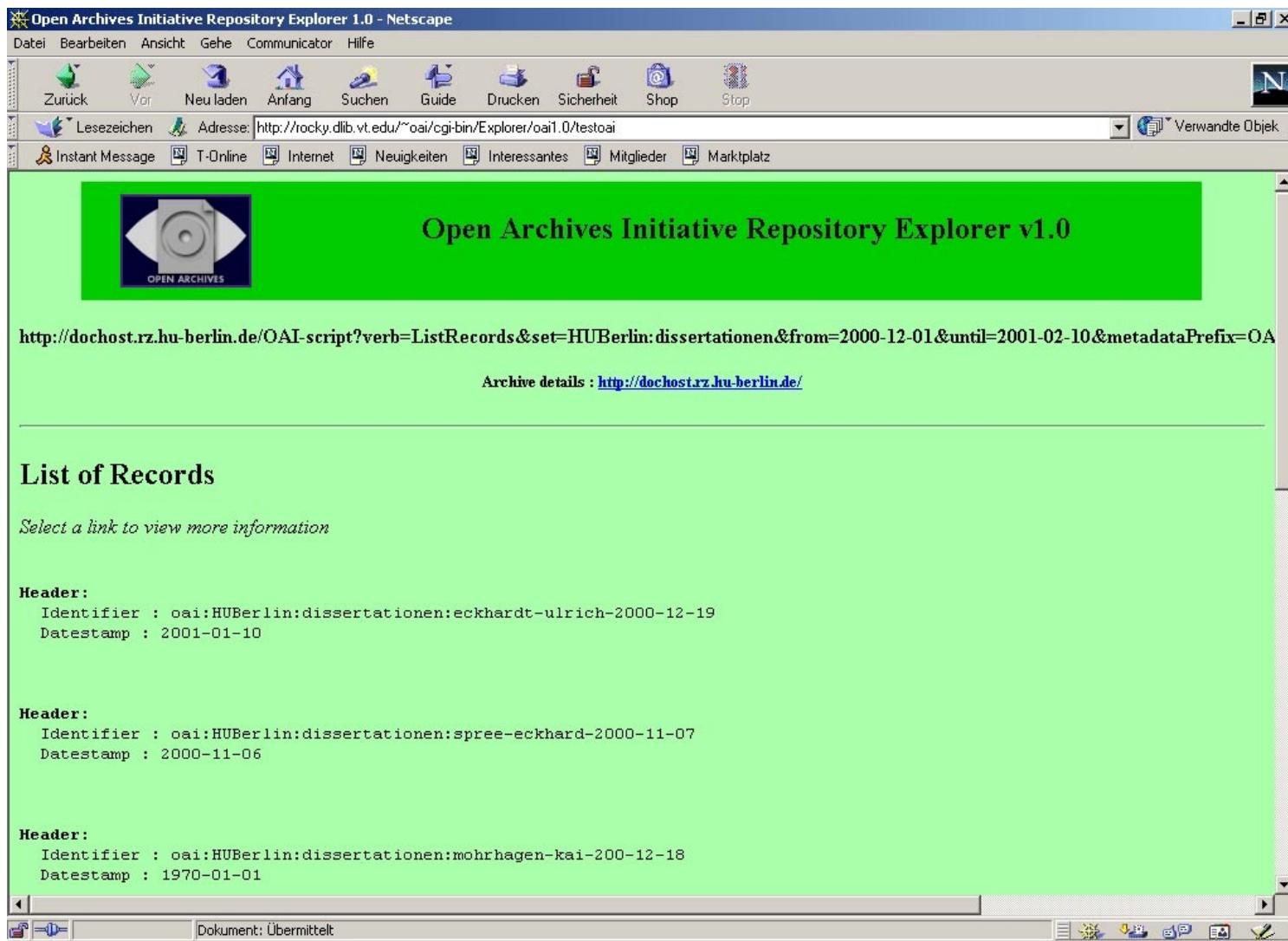
Using Virginia Tech Repository Explorer at [http://purl.org/net/oai\\_explorer](http://purl.org/net/oai_explorer)  
February 26th 2001 Heinrich Stamerjohanns & Susanne Dobratz 13

# OAI\_ListSets



Using Virginia Tech Repository Explorer at [http://purl.org/net/oai\\_explorer](http://purl.org/net/oai_explorer)  
February 26th 2001 Heinrich Stamerjohanns & Susanne Dobratz 14

# OAI\_ListRecords



Using Virginia Tech Repository Explorer at [http://purl.org/net/oai\\_explorer](http://purl.org/net/oai_explorer)  
February 26th 2001 Heinrich Stamerjohanns & Susanne Dobratz 15

# Experiences at HUBerlin

- Use Databases
  - agree to datamodel for Dublin Core for theses and dissertations
  - php4 from HUBerlin scripts can be used
- Next:
  - Integration into NDLTD and PhysDis
  - within DINI: usage e.g. for educational materials
- Protocol may be not sufficient for specific user domains

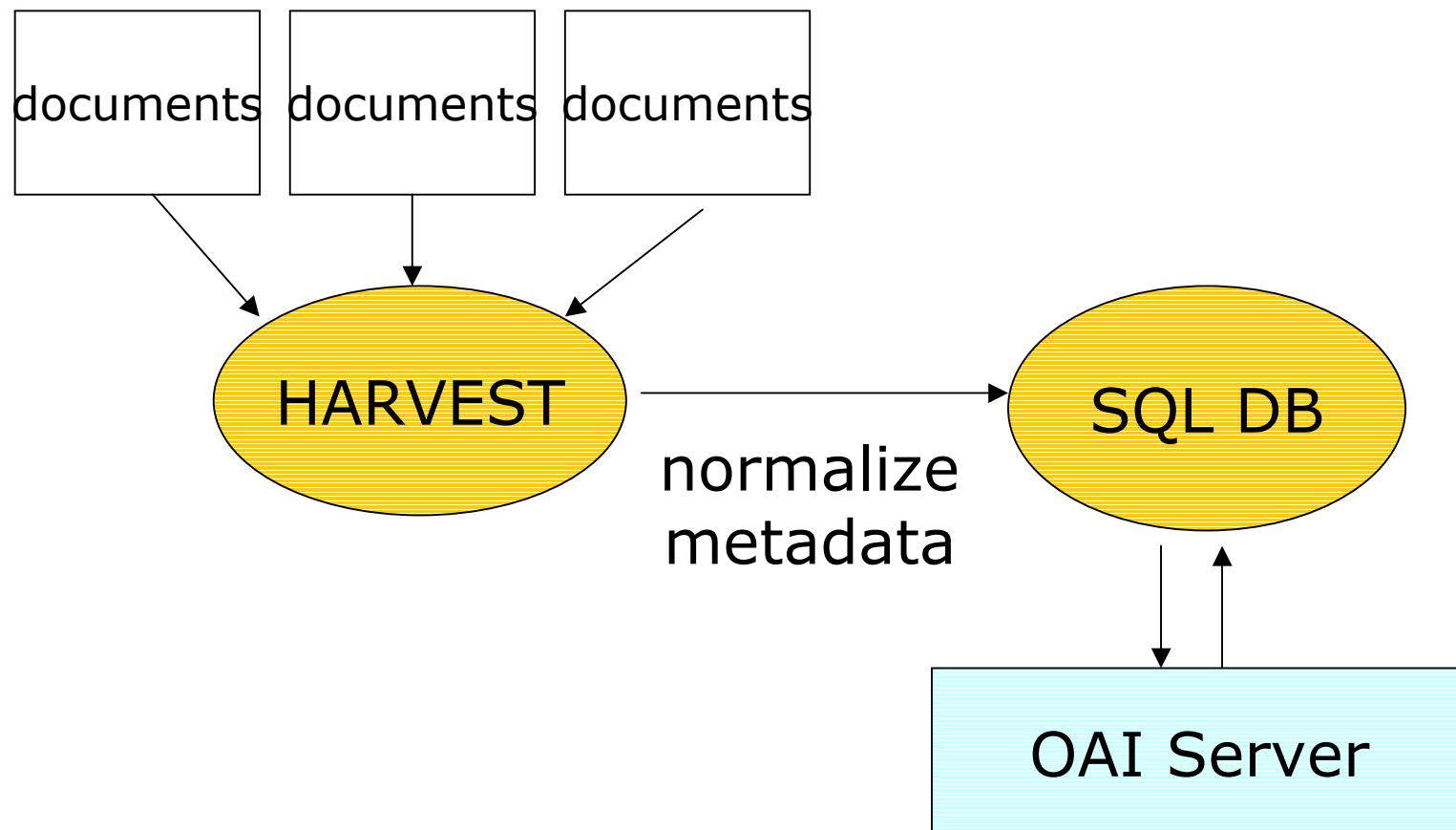
# PhysDoc

- PhysDoc is itself a distributed document database network, started in 1995
- Especially for individual authors, smaller departments and institutions
- HARVEST gatherer collects documents and its metadata from linklists of document collections
- 40000 documents used in alpha test

# OAI Implementation

- modified HARVEST holds SOIF and DC metadata in local text files
- storage size no problem
- decision to convert data offline and store structured data in SQL database (mysql)
- use DC when possible, otherwise map SOIF to DC

# OAI Implementation



# OAI Implementation

- software written in PHP
- protocol
  - easy because it uses modified implementation of HU Berlin
- metadata converter
  - maps SOIF to DC
  - converts different DC representations to one common one

# Metadata quality

- good quality important
- currently 988 documents out of 40000 provide DC metadata
- DC is not DC...
  - different representations for languages
  - GER ↔ de
  - different representations for dates
  - 26.2.2001 ↔ 2001-02-26

## Future work

- improve metadata converter
  - improve summarizers
  - closer look at different DC representations
- tell people to use metadata
  - OAI workshops
- ease production of metadata

# OAI in Germany

- Supported by DINI
- First Implementation Workshop
  - June 2001
  - by
    - Humboldt-University Berlin, Computing Centre (Peter Schirmbacher)
    - University Library of Oldenburg (Han Wätjen)
    - Institute for Science Networking, Univ. Oldenburg (Prof. Hilf)

# Thank You!

- OAI-project page at RZ HU Berlin:  
<http://dochost.rz.hu-berlin.de/oai>  
/OAI-Script
- OAI at Institute of Science Networking,  
Oldenburg:  
<http://physnet.uni-oldenburg.de/oai/oai.php>
- [stamer@uni-oldenburg.de](mailto:stamer@uni-oldenburg.de)
- [dobratz@rz.hu-berlin.de](mailto:dobratz@rz.hu-berlin.de)