

Aspekte der Datenqualität, Adressierung und Auszeichnung von Dokumenten

Workshop: Technische Aspekte
des DINI-Zertifikats 2007

Wolfram Horstmann / Friedrich Summann

Hintergrund

***Datenqualität, Adressierung und
Auszeichnung mögen lokal korrekt
erscheinen, können aber
Probleme in nationalen und
internationalen
Netzwerken bereiten***

Programm

- DINI-Zertifikat: ausgewählte Aspekte
- DRIVER-Richtlinien
- Netzwerke von Repositorien >> DEMO
- Praktische Erfahrungen bei der Aggregation
- Schluss

Eckpunkte

- Jeder OAI-Eintrag führt zu Dokumenten
- Der URI (URL/URN) ist maschinenlesbar
- Dokumente sind klassifiziert
- Aggregation, Suche und Navigation wird verbessert

DINI-Zertifikat: Sicherheit ...

- ...
- 2.5.2 Dokumente (Mindeststandard)
 - Verwendung von Persistent Identifiers, dazu zählen Systeme, die einen Resolver-Dienst besitzen, z. B. urn:nbn oder DOI.
- ...

DINI-Zertifikat: **Erschließung** ...

- 2.6.1 Sacherschließung (Mindeststandard)
 - ...
 - Verbale Sacherschließung durch freie Schlagwörter oder klassifikatorische Erschließung wird durchgeführt.
 - Dewey-Dezimalklassifikation (DDC)
 - ...

DINI-Zertifikat: **Metadatenexport**

- 2.6.2 Metadatenexport (Mindeststandard)
 - Metadaten werden frei zugänglich angeboten
 - Metadaten sind nach Dublin Core Simple (ISO 15836:2003) strukturiert.

DINI-Zertifikat: Schnittstellen

- 2.6.3 Schnittstellen (Mindeststandard)
 - ...
 - OAI-PMH 2.0 entsprechend den ***DINI-OAI-Empfehlungen*** wird unterstützt

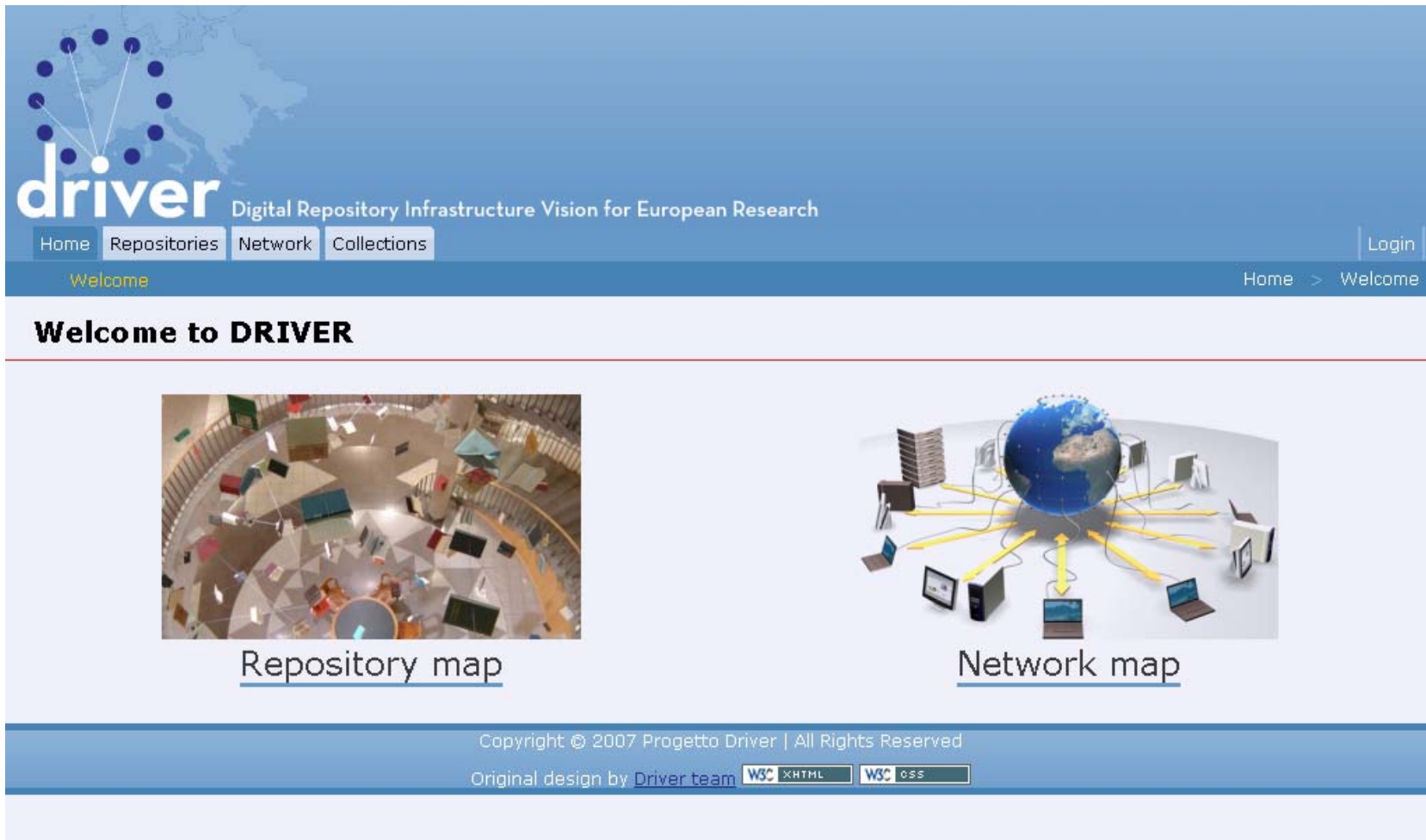
DRIVER-Richtlinien

- Europäisches Repository Netzwerk
 - komplementär zum DINI-Zertifikat
 - Standpunkt des Aggregators (nicht lokaler Operator)
- Das DINI-Zertifikat setzt den Maßstab
 - DRIVER-Richtlinien bilden nur einen Teil des DINI-Zertifikats ab
 - DINI-zertifizierte Repositorien werden automatisch in DRIVER aufgenommen
 - DRIVER-validierte Repositorien haben es nicht mehr weit zum DINI-Zertifikat

Netzwerke

- Am Beispiel DRIVER
 - Landschaft der Repositorien >> DEMO1; >> DEMO2
- Praktische Erfahrungen
 - Aggregation von Repositorien >> DEMO

DEMO OF TEST-VERSION



The screenshot shows the DRIVER website interface. At the top left is the DRIVER logo, which consists of a stylized globe with blue dots and lines, and the text "driver" in a bold, lowercase font. To the right of the logo is the text "Digital Repository Infrastructure Vision for European Research". Below the logo and text is a navigation menu with buttons for "Home", "Repositories", "Network", and "Collections". On the right side of the navigation menu is a "Login" button. Below the navigation menu is a blue banner with the text "Welcome" on the left and "Home > Welcome" on the right. Below the banner is the heading "Welcome to DRIVER". The main content area is divided into two columns. The left column features a 3D rendering of a large, modern building interior with a curved ceiling and many colorful, geometric shapes, representing a repository map. Below this image is the caption "Repository map". The right column features a 3D rendering of a globe with several yellow arrows pointing outwards to various computer monitors and laptops, representing a network map. Below this image is the caption "Network map". At the bottom of the page is a blue footer with the text "Copyright © 2007 Progetto Driver | All Rights Reserved" and "Original design by Driver team". To the right of the footer text are two small logos: "W3C XHTML" and "W3C CSS".

driver Digital Repository Infrastructure Vision for European Research

Home Repositories Network Collections Login

Welcome Home > Welcome

Welcome to DRIVER

Repository map

Network map

Copyright © 2007 Progetto Driver | All Rights Reserved
Original design by Driver team W3C XHTML W3C CSS

DEMO OF TEST-VERSION

The screenshot displays the DRIVER website interface. At the top, the logo features a circular arrangement of blue dots connected by lines, with the word "driver" in white lowercase letters. To the right of the logo is the text "Digital Repository Infrastructure Vision for European Research". Below this, a navigation bar includes "Home", "Repositories", "Network", and "Collections" buttons, along with a "Login" link. A "Show Map" link is also present. The main content area is divided into two sections. On the left is a vertical list of repositories, each with a country code and name. On the right is a map of Europe with colored dots indicating repository locations. A legend at the top of the map shows four color-coded categories for the number of records: green for "< 50", orange for "50 - 500", purple for "500 - 1000", and red for "> 1000". The map includes navigation controls like a compass, zoom in/out buttons, and a scale bar. A "Karte" button is visible in the top right of the map area. At the bottom of the page, a blue footer contains the text "Copyright © 2007 Driver Project | All Rights Reserved".

driver Digital Repository Infrastructure Vision for European Research

Home Repositories Network Collections Login

Show Map Repositories > Show Map

Number of records: ● < 50 ● 50 - 500 ● 500 - 1000 ● > 1000

Repositories

- FR Archimer, Archive Institutionnelle de l'Ifremer
- NL AUP publications
- DE BieSON - Bielefelder Server fuer Online-Publikationen (University of Bielefeld, GERMANY)
- UK Birkbeck ePrints
- UK Bristol Repository of Scholarly Eprints (ROSE)
- UK British Library Research Archive
- NL Dissertations of the Universiteit van Amsterdam
- UK DSpace at Cambridge
- NL DSpace at Erasmus
- NL DSpace at Open Universiteit Nederland
- NL DSpace at Radboud Univ. Nijmegen
- UK DSpace at the London School of Economics Library
- BE DSpace at UGent
- NL DSpace at University Leiden
- NL DSpace at Utrecht University

Karte Satellite Hybrid

Copyright © 2007 Driver Project | All Rights Reserved

DEMO OF TEST-VERSION

DRIVER Aggregator Manager - Admin Control Panel

List of Repositories

Nr.	Repository	Status	Last Harvesting Date	Harvestingtype	Harvesting schedule
1	University of Technics Hamburg, GERMANY, TUBdok	active	2007-05-28T23:34:44Z	REFRESH	WEEKLY
2	Dissertations of the Universiteit van Amsterdam	inactive	2007-06-01T10:36:14Z	REFRESH	WEEKLY
3	SciDok, der Wissenschafts-Server der Universitaet des Saarlandes	active	2007-05-28T23:45:10Z	REFRESH	WEEKLY
4	Royal Holloway Research Online	active	2007-05-28T23:37:38Z	REFRESH	WEEKLY
5	University Digital Archive of the University of Groningen, The Netherlands.	active		REFRESH	WEEKLY
6	DSpace at Open Universiteit Nederland	active	2007-05-29T13:13:52Z	REFRESH	WEEKLY
7	DSpace at Vrije Universiteit Amsterdam	inactive			
8	OAI-Repository SUB Goettingen	active	2007-05-28T23:49:49Z	REFRESH	WEEKLY
9	SOAS Eprints	active	2007-05-28T23:54:31Z	REFRESH	WEEKLY

DEMO OF TEST-VERSION



DRIVER Aggregator Manager - OAI Admin Panel

Repository Form

[List of Repositories](#)

Repository Information: [Identify](#) - [ListSets](#) - [ListMetadataFormats](#)

Record Information: [View ListRecords](#) [Test Mapping](#)

Start Harvesting (Current settings)

Repository:	SciDok, der Wissenschafts-Server der Universitaet des Saarlandes
Repository Identifier:	99-83409cf6-0d36-11dc-9ade-000347f19e46_UmVwb3NpdG9yeVNlcnZpY2VSZXNvdXJjZXMvUmVwb3NpdG9yeVNlcnZpY2VSZXNvdXJjZVR5cGU=
Harvesting instance	111-9de560c8-0d36-11dc-9ade-000347f19e46_SGFydmVzdGluZ0luc3RhbmNlRfNSZXNvdXJjZXMvSGFydmVzdGluZ0luc3RhbmNlRfNSZXNvdXJjZ
	Status: active

DRIVER Harvesting/Aggregating

Based on **BASE** Harvesting expertise



- 650 OAI interfaces tested
- 550 responses analysed
- 419 indexed -> included

DRIVER Aggregating Service



- 60 OAI repositories (DINI, SHERPA, DARE, CNRS, Gent) for the testbed
- Aggregator Service (Open Source Harvester)
- Developing Cleaning, Enriching, Mapping

OAI harvesting challenges (1)

- Repositories do not response or deliver zero records
- Repositories deliver Error Messages only
(Apache, Tomcat, PHP)
- Harvesting process is slow (records per call to low) or dies
- Incremental Harvesting not supported
(delivering zero records or all records)
- Links to the Document (dc:identifier) are not included or do not work
- XML file is not well-formed
(encoding, tagging, error messages)

OAI harvesting challenges (2)

- Resumption token usage is problematic (usage of ,0', no variation, expiration date)
- Repository is only available via Aggregating Service and cannot be extracted separately
- Links to the Document address a jump-off page (prevent indexing the fulltext)
- Data contain only References without any Fulltext
- Access to fulltext often is restricted (ip control, document delivery, login)
- Open access fulltext can not be recognized
- Field content varies without any standard

DINI-Server sind besser als der Durchschnitt!

Cleaning

Set restriction

Normalizing (dates, languages, types)

Tag mapping

Removing tags (Duplicates)

Changing values

Removing xml errors

Correcting encoding

Enriching

Repository name

Repository country

Date of collection

Adding the normalized fulltext

Schluss

- **Datenqualität**, Adressierung und Auszeichnung von Dokumenten entscheidend für die **Aggregation**
- Bestimmt **Außenbild** der Repositorien
 - Bis hin zum Fehlen von Server / Dokumenten
- **DINI-Zertifikat** beinhaltet das Wesentliche
- **DRIVER** Richtlinien werden entsprochen
- **Registrierung** in ROAR/OpenDoar

Kontakt & Hilfe

- DINI
 - Anträge: <http://www.dini.de>
- DRIVER
 - Allgemein: whorstmann@sub-goettingen.de
 - Guidelines Helpdesk
 - <http://www.driver-support.eu>
 - feijen@surf.nl
 - ++31-30-2346600