

# 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

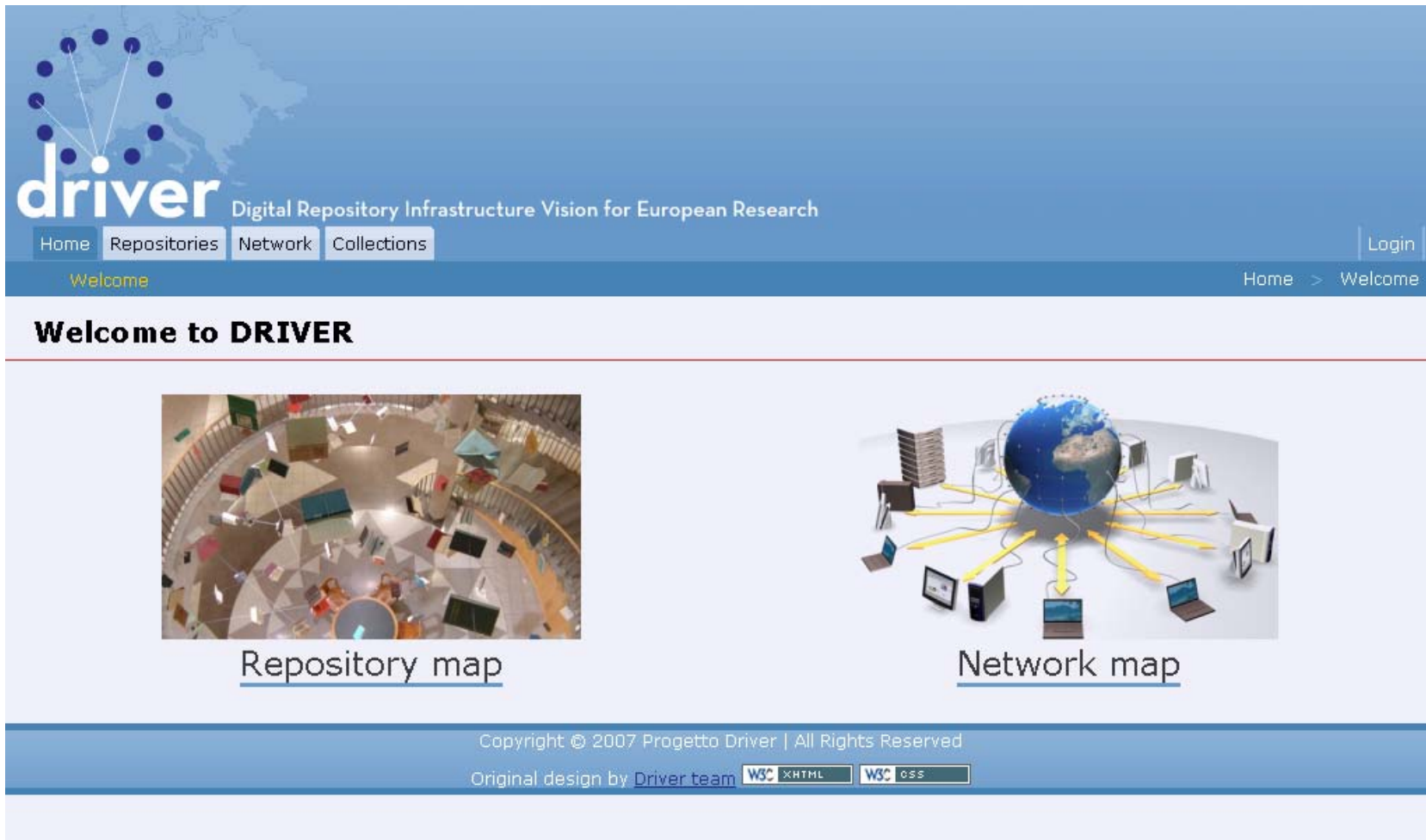
# DRIVER-Richtlinien

- 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 image shows a screenshot of the DRIVER website's home page. At the top left, there is a logo consisting of a circle of blue dots connected by lines, with the word "driver" in white lowercase letters below it. To the right of the logo, the text "Digital Repository Infrastructure Vision for European Research" is displayed. Below the logo and text, there are navigation buttons for "Home", "Repositories", "Network", and "Collections". On the right side of the top navigation bar, there is a "Login" button. Below the navigation bar, the text "Welcome" is displayed on the left, and "Home > Welcome" is displayed on the right. The main content area features the heading "Welcome to DRIVER" in bold black text. Below this heading, there are two images: on the left, a 3D architectural rendering of a large, modern building interior with a central atrium and multiple levels, labeled "Repository map"; on the right, a 3D illustration of a globe with yellow arrows pointing to various computer monitors and servers, labeled "Network map". At the bottom of the page, there is a blue footer bar containing the text "Copyright © 2007 Progetto Driver | All Rights Reserved" and "Original design by Driver team". To the right of this 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 left is the DRIVER logo, a stylized globe with the text "driver" and "Digital Repository Infrastructure Vision for European Research". Navigation tabs include "Home", "Repositories", "Network", and "Collections". A "Login" link is on the right. Below the navigation is a "Show Map" button. The main content area features a map of Europe with colored dots representing repository locations. A legend at the top of the map indicates the number of records: green for < 50, orange for 50 - 500, purple for 500 - 1000, and red for > 1000. The map includes navigation controls (directional arrows, zoom in/out, and a scale bar) and style selection buttons for "Karte", "Satellit", and "Hybrid". On the left side, a "Repositories" list is shown, containing entries such as "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", and "NL DSpace at Utrecht University". The footer contains the text "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	<a href="#">University of Technics Hamburg, GERMANY, TUBdok</a>	active	2007-05-28T23:34:44Z	REFRESH	WEEKLY
2	<a href="#">Dissertations of the Universiteit van Amsterdam</a>	inactive	2007-06-01T10:36:14Z	REFRESH	WEEKLY
3	<a href="#">SciDok, der Wissenschafts-Server der Universitaet des Saarlandes</a>	active	2007-05-28T23:45:10Z	REFRESH	WEEKLY
4	<a href="#">Royal Holloway Research Online</a>	active	2007-05-28T23:37:38Z	REFRESH	WEEKLY
5	<a href="#">University Digital Archive of the University of Groningen, The Netherlands.</a>	active		REFRESH	WEEKLY
6	<a href="#">DSpace at Open Universiteit Nederland</a>	active	2007-05-29T13:13:52Z	REFRESH	WEEKLY
7	<a href="#">DSpace at Vrije Universiteit Amsterdam</a>	inactive			
8	<a href="#">OAI-Repository SUB Goettingen</a>	active	2007-05-28T23:49:49Z	REFRESH	WEEKLY
9	<a href="#">SOAS Eprints</a>	active	2007-05-28T23:51:31Z	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)

<b>Repository:</b>	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](mailto:whorstmann@sub-goettingen.de)
  - Guidelines Helpdesk
    - <http://www.driver-support.eu>
    - [feijen@surf.nl](mailto:feijen@surf.nl)
    - ++31-30-2346600