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

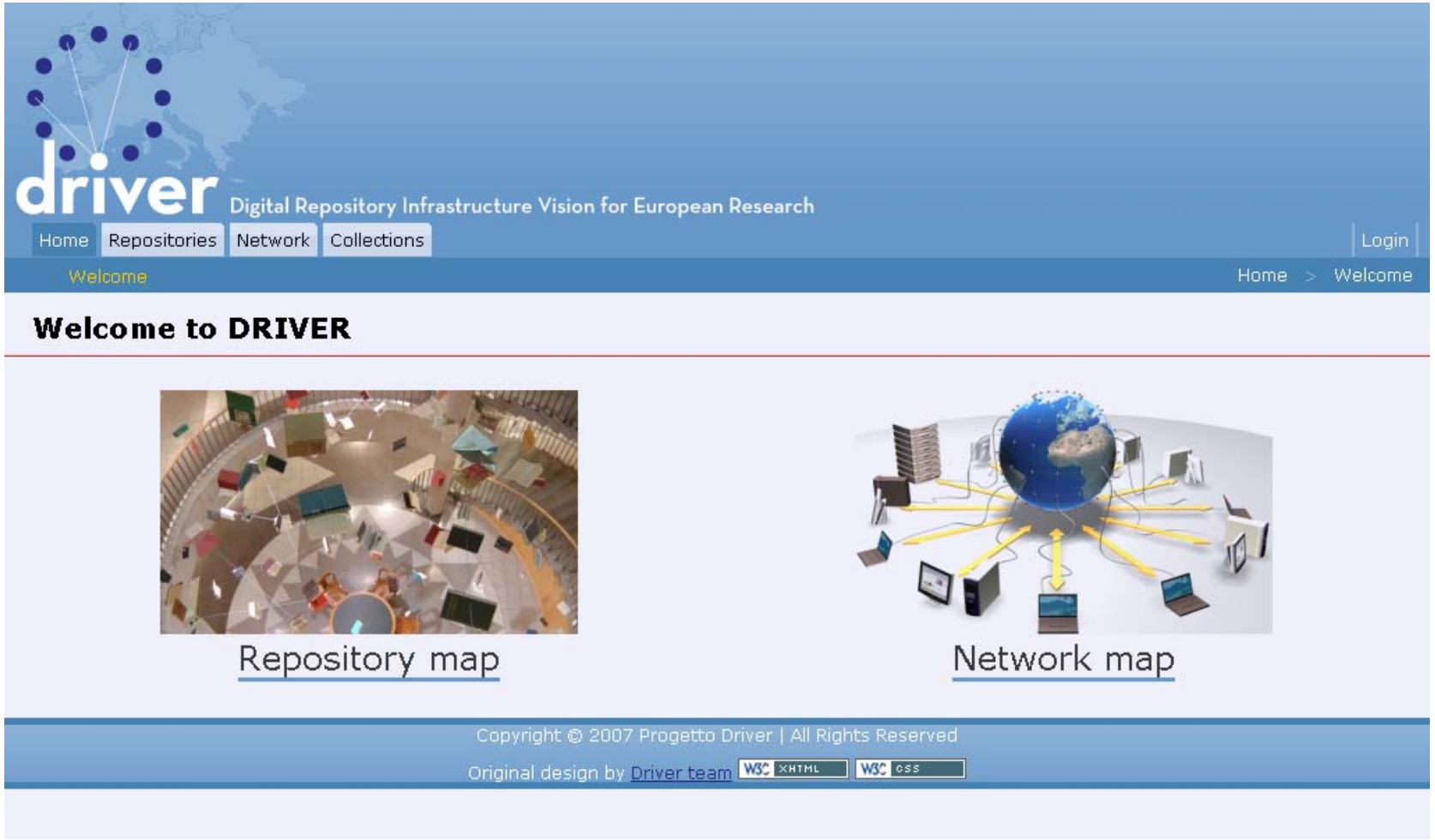
- Europäisches Repository Netzwerk
  - komplementär zum DINI-Zertifikat
  - Standpunkt des Aggregators (nicht lokaler Operator)

- 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

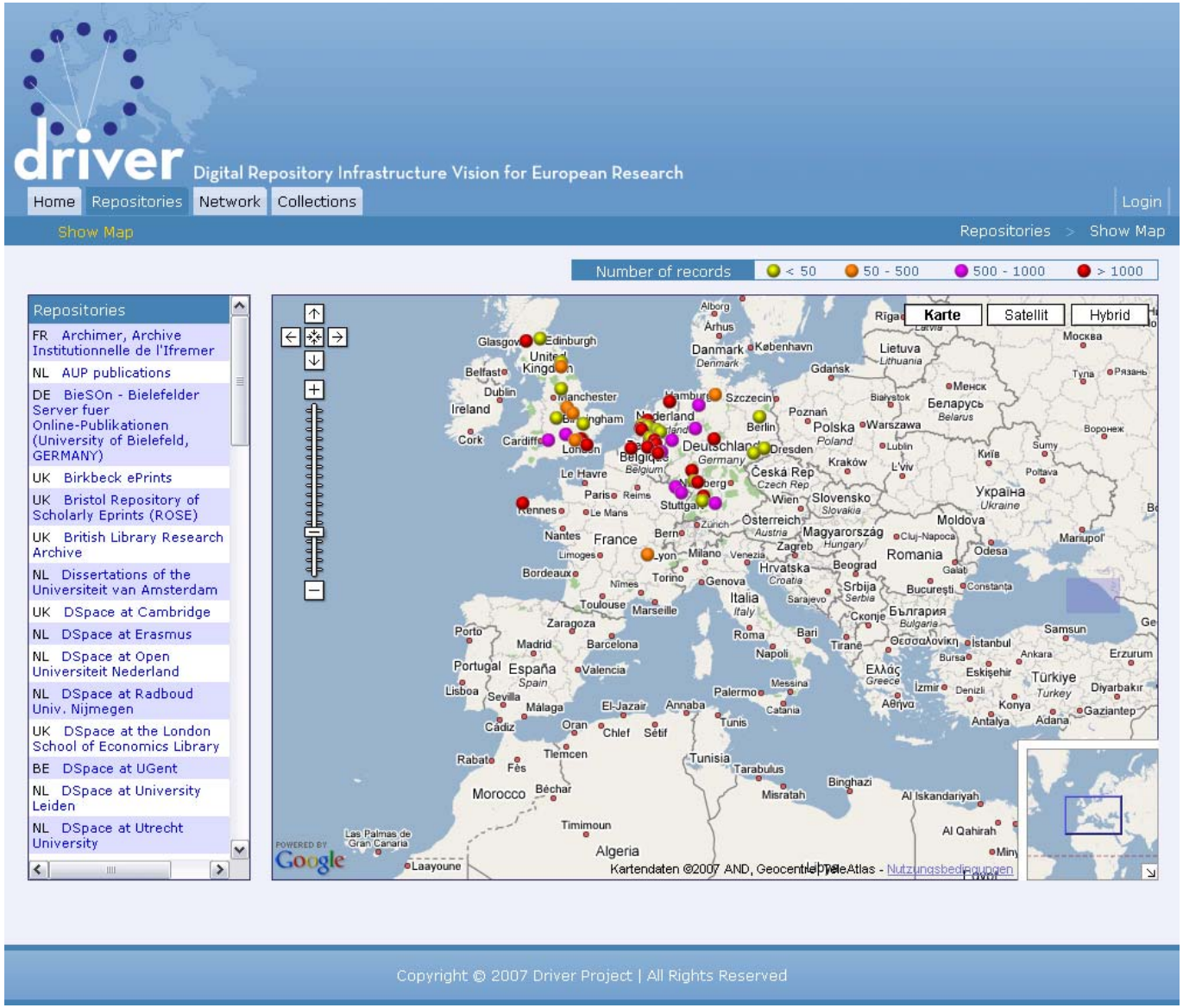
Netzwerke

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

DEMO OF TEST-VERSION



DEMO OF TEST-VERSION



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_UmVwb3NpdG9yeVNIcnZpY2VSZXNvdXJjZXMvUmVwb3NpdG9yeVNIcnZpY2VSZXNvdXJjZVR5cGU=**
Harvesting instance	**111-9de560c8-0d36-11dc-9ade-000347f19e46_SGFydmVzdGluZ0luc3RhbmNIRFNSZXNvdXJjZXMvSGFydmVzdGluZ0luc3RhbmNIRFNSZXNvdXJjZ**
	Status: active

DRIVER Harvesting/Aggregating

Based on BASE Harvesting expertise

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

DRIVER Aggregatíng 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 seperately
- Links to the Document adress 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, Addressierung 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