SEVENTH FRAMEWORK PROGRAMME

CAPACITIES

Research Infrastructures

INFRA-2007-1.2.1 Research Infrastructures

DRIVER II

Grant Agreement 212147

"Digital Repository Infrastructure Vision for European Research II"

Information Space Report

Deliverable Code: D5.1

Document Description

Project	
Title:	DRIVER, Digital Repository Infrastructure Vision for European Research II
Start date:	1st December 2007
Call/Instrument:	INFRA-2007-1.2.1
Grant Agreement:	**212147**

Document	
Deliverable number:	D5.1
Deliverable title:	Information Space Report
Contractual Date of Delivery:	1st of June 2008
Actual Date of Delivery:	1st of December 2008
Editor(s):	CNR
Author(s):	Paolo Manghi
Reviewer(s):	Friedrich Summann
Participant(s):	
Workpackage:	WP5
Workpackage title:	Infrastructure Sustainability and Maintenance
Workpackage leader:	CNR
Workpackage participants:	NKUA, CNR, DTU,UMINHO, NUK, UON, UGENT
Distribution:	Public
Nature:	Deliverable
Version/Revision:	0.2
Draft/Final:	Final
Total number of pages: (including cover)	17
File name:	D5.1.pdf
Key words:	Index, MDStore, orchestration

Disclaimer

This document contains description of the DRIVER II project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of DRIVER consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 25 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

DRIVER-II is a project funded by the European Union

Table of Contents

Table of Figures

Summary

This deliverable introduces the notion of Service orchestration used in the DRIVER infrastructure to construct and maintain Information Spaces. More specifically, it illustrates how it is used to maintain the DRIVER Information Space of European Open Access publications.

1 Introduction

Infrastructure systems are maintained by one *responsible organization* (RO), which provides support to other *participating organizations* (POs) willing to provide and integrate their resources and constructing their applications with such resources.

The DRIVER Infrastructure adopts the component oriented approach of *Service Oriented Architectures* (SOA) [1], where services are implemented as Web Services [2]. In SOA, applications consist of a set of distributed services that interact to deliver the functionalities expected by their users. Services support specific sets of functionalities in isolation and can be combined into workflows of actions to model arbitrary complex data computation processes. Most importantly, Services can be shared between different applications.

The DRIVER infrastructure enriches the principles of SOA with the notion of "orchestration". More specifically, an application is not defined as a predetermined set of Services, but as a "declaration" of all service functionalities needed for the application to work. At run-time, infrastructure orchestration mechanisms will form PO applications by enabling and guiding the interaction between those Services that are sharable and match the relative declaration constraints. POs can therefore contribute in augmenting the quality-of-services of all running applications by offering new hardware and deploying DRIVER service instances; the orchestration framework will automatically exploit the new resources by reusing them in the context of needing applications.

The infrastructure framework governs all applications by means of special *Enabling Services*, available 24/7 and administered by the infrastructure RO; e.g. the European DRIVER Infrastructure for Open Access publications is administered by the *DRIVER Consortium* RO. All services need to *register* to the *Information Service* their *profile*, i.e. information into about their web location, the functionality they expose and their current status – to be continuously updated by services during their life-time. The Information Service is therefore the keeper of the "infrastructure map", being constantly updated on the latest status of all system resources.

In DRIVER, orchestration is delegated to special services, named Manager Services. Such Services are delegated the execution of actions, as a consequence of the occurrence of certain events; in the Data Layer, they are in charge of orchestrating the available resources in order to prepare the environment for the harvesting, aggregation, storage and indexing of metadata records from OAI-PMH repositories. The Manager Service orchestrates services by dynamically *discovering* through the Information Service where the services it requires are located and by combining their interaction. In the following sections we shall illustrate how the Manager Service orchestrate the available Data Layer resources to accomplish Information Space construction.

2 Information Space construction

The Data Layer of the infrastructure consists of an arbitrary number of instances of MDStore Services, Index Services, Aggregator Services. Other kind of services are part of the layer, such as Text Engine Services, Collection Services, but these are not concerned with the orchestration issues to be discussed here.

As shown in Figure 1, the data flow in DRIVER obeys the following path: Aggregator Services are used to transfer Dublin Core (DC) metadata records from OAI-PMH repositories into MDStore Services, and then transform the incoming DC records into DRIVER Metadata Format (DMF), to be stored again in MDStore Services. Finally, DMF records are fed to Index Services. The Index Services, fed with records, constitute the DRIVER Information Space, i.e. a searchable space of uniform metadata records. Search Services reply queries over the Information Space, received from portals (user interfaces), by forwarding them to the "best" Index Service they discover in the infrastructure.

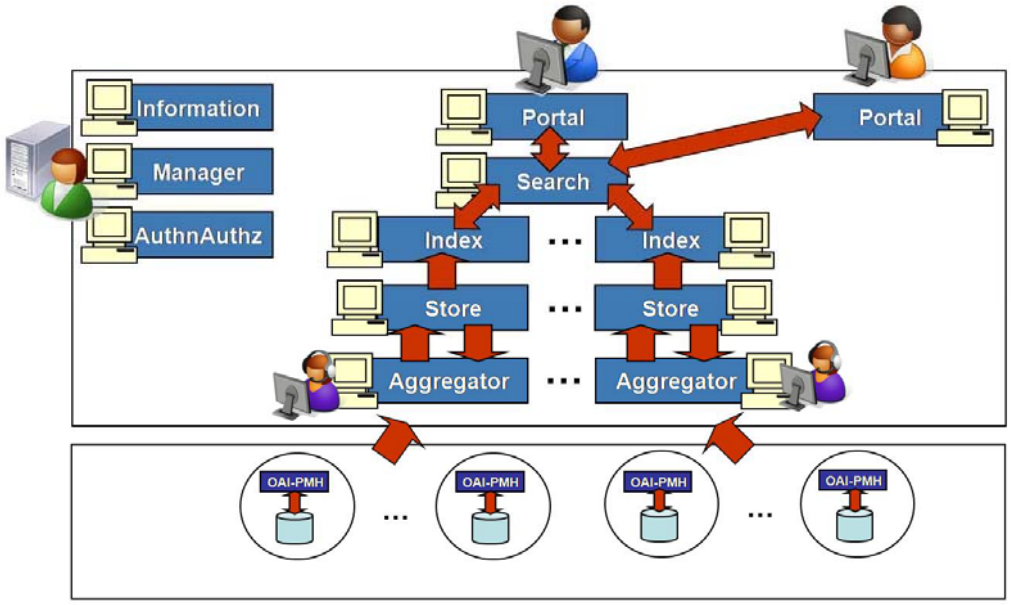

Figure 1 – Information Space construction: Data-Flow in DRIVER

2.1 Data processing

In DRIVER, the data flow described above is performed semi-automatically by the Manager Service, which orchestrates Repositories, MDStore, Index and Aggregator Services available to the infrastructure, and by Aggregator Managers, i.e. administrators responsible for the transformation phase.

2.1.1 Repository Service

OAI-PMH Repositories are registered to the system by repository managers. A Repository profile is initially created into the Information Service, specifying the oai URL of the repository, the name of the repository, contact information of the repository managers, country of origin, and other relevant information. The status of the repository, visible in the profile, is *pending*, i.e. waiting to be *enabled*. A Repository is enabled by a DRIVER administrator when it has been validated, which means that quality of service and content has been certified.

2.1.2 MDStore and Index Service

MDStore Services and Index Services are so-called *factory services*, which means their running instances are capable of managing special system resources, called *Data Structures*. In particular, factory services offer two levels of functionality:

- Data Structure management: creation, delete and update of data structures (from which the adjective “factory”);
- Data Structure interaction: execute an operation over a given data structure, created by the given service.

MDStore Services manage MDStore data structures (see Figure 2), which means they can be requested to allocate some storage space capable of storing metadata records of a given metadata format. In the Figure, metadata formats are represented by the different colors – “interpretation”, not relevant to our discussion, is a tag used to differentiate records of the same metadata format but associated to different domains. Any consuming service, e.g. the Manager Service, authorized to do so, can therefore discover an MDStore Service available and send it a request for the creation of an MDStore Data Structure that matches its needs. Once the MDStore Data structure is created, consuming services, e.g. the Aggregator Service, can interact with the relative MDStore Service so as to add, remove or retrieve records within the MDStore.

Index Services are factory services managing Index Data Structures. Similarly to MDStore Services, Index Services can create or delete Index Data Structures capable of indexing records of a given metadata format. Besides, consuming services can feed records to the Index Data Structures, empty them, or query them, through SRU/SRW interfaces.

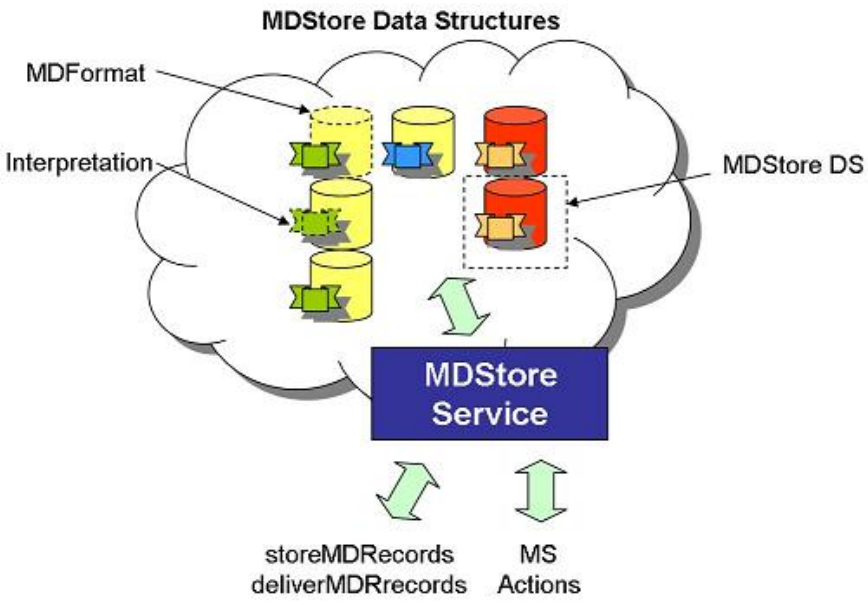

Figure 2 – MDStore “factory” Service

2.1.3 Aggregator Service

Aggregator Services are concerned with Repository content management, thus of the data flowing from the repositories onto the DRIVER Data Layer. An infrastructure can host a number of Aggregator Service instances and each of them is in charge of managing a separate list of repositories. Each instance is in turn operated through a user interface by

an Aggregator Manager, an administrator responsible for performing two main actions over the assigned repositories:

- Harvesting: fetching metadata records from the OAI-PMH repository interface and store them into a Dublin Core MDStore Data Structure;
- Transforming: fetching the data from the Dublin Core MDStore Data Structure and transform them into DMF records to be then stored into a DMF MDStore Data Structure. Transformation consist in applying the mappings from Dublin Core format onto DMF, to be specified by the aggregator manager, to the harvested Dublin Core records.

Accordingly, in order to perform the aggregation process relative to a Repository the Aggregation Service needs to know:

- the identifier of the Repository profile: to fetch its oai URL and start harvesting;
- the identifier of the Dublin Core MDStore Data Structure profile: to know the location of the relative MDStore Service and deliver into the MDStore Data Structure the harvested DC records;
- the identifier of the DMF MDStore Data Structure profile: to to know the location of the relative MDStore Service and deliver into the MDStore Data Structure the DMF records obtained from the transformation;
- the mappings from the Dublin Core formats onto the DMF format.

Harvesting and transformation can be fired manually by the Aggregator Manager or, once the mapping has been specified, be scheduled for automatic execution at given intervals in time (daily, weekly, monthly) by the same person.

2.2 Data Layer Orchestration

The Manager Service (MS) orchestrates the resources available to the data layer, thus including data structures and services, to ensure the data workflow depicted in the previous section. In particular, orchestration consists on the execution of a number of actions in correspondence of certain resource-related events. Functionality for subscription and notification of events are supported by the Information Service. Services can contact this service to subscribe and be notified to events of the form:

- Registration/de-registration of resources of a certain type; e.g. creation of an MDStore Data Resource, deployment of an Index Service.
- Modification of the status of a given resource (by identifier); e.g. repository status from “pending” to “enabled”; MDStore Data Structure has been fed with new records.

Information Space construction is performed by having the MS supporting Aggregator Managers in their mapping definition and execution. To his aim, the MS first prepares the environment to the DC harvesting and DMF transformation of a Repository, then waits for the Aggregator Manager to perform both actions, and finally indexes the DMF records. More specifically:

- Occurrence of the event **enabling of a repository**: the MS is subscribed to all events of this form; when notified, the MS prepares the environment to host the records of the new repository, which means it needs to create Dublin Core and DMF MDStore Data Structures. In particular, to ensure robustness, the MS aims at generating K replicas of both kinds of metadata stores. To this aim:

- the MS discovers K MDStore Services available to allocate a pair of DC and DMF MDStores Data Structures and sends to all of them the double request.
- When K×2 MDStores have been created in the K Services, the MS creates and registers to the Information Service a new resource, called *Harvesting Instance Data Structure*, which contains: repository identifier, DC MDStore identifiers, DMF MDStore identifiers.
- The Manager Service subscribes to be notified of the event **DMF MDStore fed with new content** for all the newly created DMF MDStore Data Structures; the subscription is needed to orchestrate Index feeding whenever new DMF records are created;
- The Harvesting Instance is then sent to an available Aggregator Service, where an Aggregator Manager has to manually set up the harvesting and transformation phases relative to the repository.

- The Aggregator Service selected in the previous step, receives the Harvesting Instance and alerts the responsible Aggregator Manager that a new Repository is ready for harvesting. The administrator accesses the Aggregator Service user interface and sets up the mapping to be applied to the incoming Dublin Core records. When the mapping is ready and configured, the Aggregator Manager can proceed with Dublin Core harvesting and transformation. The second step of transformation, results in storing the new DMF records into the DMF MDStores, thereby firing the following MS-handled event.
- Occurrence of the event **DMF MDStore fed with new content**: when notified of this event, the MS has to move the new DMF records, relative to a specific repository, into an Index Service. Moreover, in order to ensure robustness and enforce scalability, the MS feeds the records to K Index Services, in K dedicated Index Data Structures. To this aim:
  - the MS discovers K Index Services available to allocate a new DMF Index Data Structures and sends to all of them the request.
  - When the Index Data Structures have been created, the MS extracts the new records from the DMF MDStore Data Structures and feeds them to the Index Data Structures.

3 DRIVER Information Space

Currently, the DRIVER Information Space is maintained by replicating the harvested Dublin Core records and the corresponding DMF records three times, in three different MDStore Services, located at different remote sites. At the moment, three MDStore Services are running in the infrastructure, respectively at CNR, UNIBI and ICM.

DMF records are then indexed by creating one Index Data Structure for each DMF MDStore, thus one Index Data Structure for each repository. In particular, the MS ensures that each Index Service instance in the infrastructure has one replica of each Index Data Structure created in the environment. At the moment, three Index Service instances are running in the infrastructure, respectively at CNR, NKUA and ICM. Such policies ensures that each Index Service cover the whole DRIVER Information Space and is therefore available to reply queries from any consuming application.

Today, 1st of December 2008, 166 repositories (see Table 1) distributed across 21 countries (see Figure 3) are already part of the Information Space for a total of 763,018 open access records. This number is due to increase quickly in the next months. DC content is available and stored in three MDStore Data Structure replicas as well as the corresponding DMF records. Similarly, DMF records are indexed three times under different Index Services, accessible through three Search Services, which can be queried in turn, by five portals.

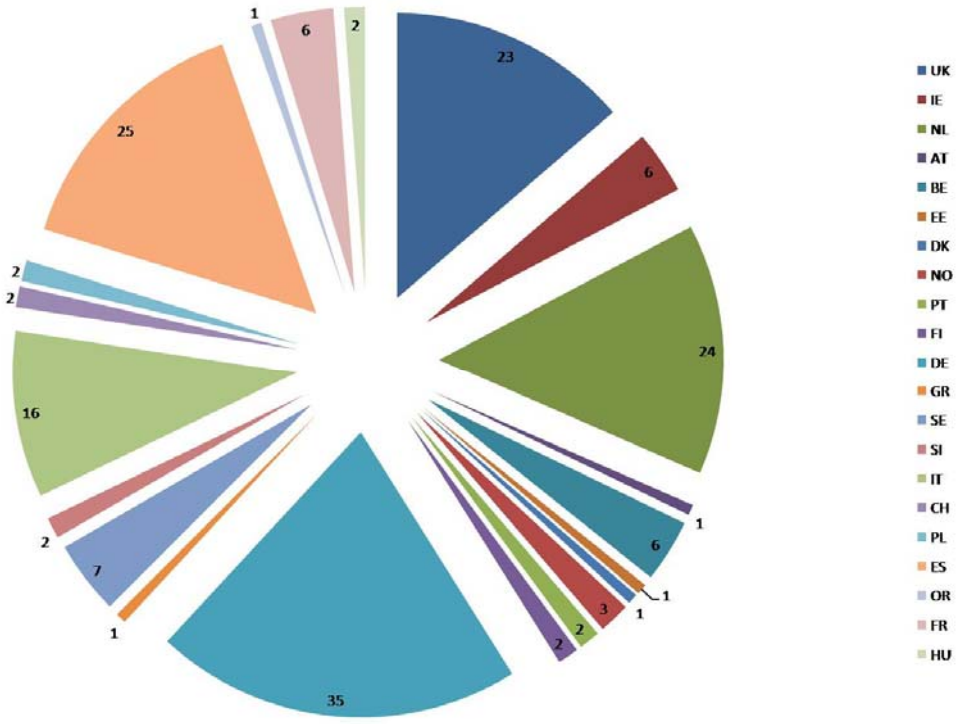

Figure 3 – Distribution of repositories per country

Repository Name	*Records*	*Country*

ARROW@DIT	378	IE
ART-Dok Publikationsplattform Kunstgeschichte	590	DE
AUP publications	64	NL
Alma Mater Digital Library	121	IT
AperTo	238	IT
ArchiMeD - Elektronische Publikationen der Universität Main	1371	DE
Archimer, Archive Institutionnelle de l'Ifremer	3943	FR
Archivio Eprints Universita di Firenze	200	IT
Archivo Digital UPM	970	ES
Arias Montano: Repositorio Institucional de la Universidad d	59	ES
Aristotle University of Thessaloniki	1592	GR
BURJC-DIGITAL Universidad Rey Juan Carlos	187	ES
Bibliotheque Numerique RERO DOC	4744	CH
Biblos-e Archivo Universidad Autonoma de Madrid	882	ES
BieSOn - Bielefelder Server fuer Online-Publikationen (Unive	1038	DE
Birkbeck ePrints	510	UK
Bristol Repository of Scholarly Eprints (ROSE)	1981	UK
British Library Research Archive	58	UK
Cranfield CERES	2767	GB
DALEA - Högskolan Dalarnas elektroniska arkiv	3007	SE
DESY Publication Database	3261	DE
DIPP NRW	772	DE
DORAS	725	IE
DSpace at Cambridge	172242	UK
DSpace at Erasmus	6684	NL
DSpace at Open Universiteit Nederland	15	NL
DSpace at Radboud Univ. Nijmegen	7279	NL
DSpace at Tartu University Library	7486	EE
DSpace at UGent	4974	BE
DSpace at University Leiden	8994	NL
DSpace at Utrecht University	22679	NL
DSpace at Vrije Universiteit Amsterdam	7446	NL
DSpaceUnipr - University of Parma	480	IT
Debreceni Egyetem elektronikus Archivumanak (DEA)	8397	HU
Digital.CSIC	6351	ES
Dissertations of the Universiteit van Amsterdam	2339	NL
Document Server@UHasselt	1401	BE
Doktori disszertációk	294	HU
Dokumenten- und Publikationsservice	443	DE
Durham Research Online	4930	UK
E-Archivo	2538	ES
E-LIB Dokumentserver	1110	DE
ECNIS Repository (Environmental Cancer Risk, Nutrition and I	132	PL
ELPUB : Digital Library	534	SI

ETH E-Collection	8969	CH
Edinburgh Research Archive	1998	UK
Eindwerkensite Katholieke Hogeschool Kempen	4048	BE
Electronic Research Archive	955	SE
Elektronische Dissertationen der Ruhr-Univ. Bochum	2388	DE
Elektronisches Volltextarchiv EVA	3756	DE
Estudo Geral	2970	PT
Fraunhofer-ePrints	3517	DE
GFZ Publications	786	DE
Glasgow ePrints Service	4073	UK
Goldsmiths Research Online	475	UK
HAL - Hyper Article on Line	143532	FR
Heidelberger Dokumentenserver (HeiDok)	7792	DE
Hochschulschriftenserver(HSSS) SLUB Dresden, GERMANY, Docume	2937	DE
Horizon / Pleins textes	33713	FR
Humboldt University of Berlin, GERMANY, Document Server	5485	DE
International Institute for the Study of Islam in the Modern	81	NL
JUWEL (JUelicher Wissenschaftliche Elektronische Literatur)	2879	DE
KNAW	1492	NL
Kobra - DSpace an der Universität Kassel	1259	DE
LSE Research Online	6764	UK
Les Bibliothèques Virtuelles Humanistes (BVH)	225	FR
Linköping University Electronic Press	5354	SE
MONARCH - Dokumenten- und Publikationsservice	1703	DE
NUI Maynooth ePrints and eTheses Archive	853	IE
National Research Council Genova - IBF Repository	3	IT
National Research Council Genova - IENI Repository	88	IT
National Research Council Genova - IMATI Repository	123	IT
National Research Council Genova - ISEM Repository	6	IT
National Research Council Pisa - IFC Repository	126	IT
National Research Council Pisa - IIT Repository	190	IT
National Research Council Pisa - ISTI Repository	1744	IT
National Research Council Taranto - IAMC Repository	1	IT
Naturalis Digital Academic Repository	1034	NL
Nottingham ePrints	567	UK
Nottingham eTheses	261	UK
OAI Repository of the Technische Universiteit Eindhoven (TU/	16248	NL
OAI-Repository SUB Goettingen	1755	DE
OAR, het Open Archief van VIOE-publicaties	166	BE
OPUS Publikationsserver	269	DE
ORBi (University of Liège)	632	BE
Online-Publikationsservice der Universitaet Tuebingen	3037	DE
Open Marien Archief (OMA)	3533	BE
Oulun yliopiston julkaisemia elektronisia julkaisuja	1184	FI

Oxford Eprints	555	UK
PADIS	355	IT
Padua@research	413	IT
Pharmacy Eprints	880	UK
Poznan Supercomputing and Networking Center (PSNC): Digital	106	PL
Propylaeum-DOK Publikationsplattform Altertumswissenschaften	147	DE
PsyDok, Psychologie Open Access Server der Saarlaendischen U	1613	DE
Publications of the Universiteit van Amsterdam	8393	NL
Publikasjoner fra Norges teknisk-naturvitenskapelige univers	1839	NO
Publikationer fran Karlstads Universitet	1610	SE
Publikationer från Stockholms universitet	2696	SE
Publikationer från Uppsala universitet	3973	SE
Publikationsserver der Universitaet Potsdam	2471	DE
RUDAR - Roskilde University Digital Archive	3263	DK
Repositorio Institucional da FCT-UNL	380	PT
Royal Holloway Research Online	1015	UK
SOAS Eprints	4920	UK
SavifaDok Publikationsplattform fuer die Sozialwissenschafte	157	DE
SciDok, der Wissenschafts-Server der Universitaet des Saarla	1249	DE
Scientific production, Tilburg University	8874	NL
Spir@l - Imperial College Digital Repository	1187	UK
Surrey Scholarship Online	836	UK
TARA	6592	IE
TEORA - Telemark Open Research Archive	483	NO
TKK Teknillisen korkeakoulun elektroniset vaeitoeskirjat	1232	FI
TU Delft repository	10995	NL
The Depot	61	UK
The Univ. of Birmingham: The ePrints Archive	69	UK
Theses database, Tilburg University	4318	NL
Thèses de l'Ecole nationale des chartes	306	FR
UCL Eprints	5582	UK
UDCDspace	458	ES
UNESCO-IHE Institute for Water Education	21	NL
Univ. Duesseldorf: Duesseldorfer Dokumenten- und Publikation	2304	DE
Univ. Girona: Tesis Doctorals en Xarxa (TDX)	328	ES
Univ. Konstanz: Konstanzer Online-Publikations-System (KOPS)	6134	DE
Univ. Maastricht	9268	NL
Univ. Murcia: Tesis Doctorals en Xarxa (TDX)	151	ES
Univ. Tromsoe: Munin Open Research Archive	1425	NO
Univ. do Minho: RepositoriUM	7276	PT
Universidad de Cantabria: Tesis Doctorals en Xarxa (TDX)	57	ES
Universidad de Oviedo: Tesis Doctorals en Xarxa (TDX)	148	ES
Universidade da Coruña: Tesis Doctorals en Xarxa (TDX)	23	ES
Universitaet Stuttgart	3400	DE

Universitat Autònoma de Barcelona: Tesis Doctorals en Xarxa	2065	ES
Universitat Jaume I: Tesis Doctorals en Xarxa (TDX)	158	ES
Universitat Oberta de Catalunya: Tesis Doctorals en Xarxa (T	3	ES
Universitat Politecnica de Catalunya	811	ES
Universitat Pompeu Fabra: Tesis Doctorals en Xarxa (TDX)	203	ES
Universitat Ramon Llull: Tesis Doctorals en Xarxa (TDX)	88	ES
Universitat Rovira i Virgili: Tesis Doctorals en Xarxa (TDX)	378	ES
Universitat de Lleida: Tesis Doctorals en Xarxa (TDX)	160	ES
Universitat de València: Tesis Doctorals en Xarxa (TDX)	566	ES
Universitat de Vic: Tesis Doctorals en Xarxa (TDX)	2	ES
Universiteit voor Humanistiek	23	NL
University Digital Archive of the University of Groningen, T	6156	NL
University of Mannheim, GERMANY, MADOC	2036	DE
University of Newcastle Library E-Print Archives	5368	UK
University of Technics Hamburg, GERMANY, TUBdok	458	DE
University of Twente Publications	7162	NL
University of Ulm, GERMANY, VTS Publication Service	1336	DE
Wageningen University and Researchcenter Publications	14952	NL
Waterford Inst. of Technology: WIT Institutional Repository	334	IE
Webbased Archive of RIVM Publications	2201	NL
Whiterose Research Online	3366	UK
ask23 repository (lkw, hfbk hamburg, germany)	91	DE
e-spacio UNED	14150	ES
eDoc Server Max Planck Institute for Plasma Physics	985	DE
ePrints.FRI	217	SI
ePub WU	811	AT
e_buah Biblioteca digital de la Universidad de Alcala	1148	ES
fedOA	2044	IT
mediaTUM	3700	DE
memSIC : Memoires en Sciences de l'Information et de la Comm	155	FR
miami	3580	DE
pub.fontys.nl	1555	NL
research_online@UCD	628	IE
xerxes	7741	SE

Table 1 – Repositories in the Information Space

4 References

[1] Ali Arsanjani, *Service-oriented modeling and architecture*, IBM developerWorks, http://www-128.ibm.com/developerworks/webservices/library/ws-soa-design1

[2] *W3C Web Services Activity web site*, http://www.w3.org/2002/ws