Digital Repository Infrastructure Vision for European Research

DRIVER Guidelines 2.0

Guidelines for content providers - Exposing textual resources with OAI-PMH

[November 2008]

[Guidelines for Repository Managers and Administrators on how to expose digital scientific resources using OAI-PMH and Dublin Core Metadata, creating interoperability by homogenising the repository output.]

Abstract

For communication in general it is important that person B is able to understand what person A is saying. For this common understanding one needs a common ground, a basic lexicon with an awareness of the meaning of things. From this point on one can start reasoning. In order to support scholarly communication with the use of repositories, repositories should speak the same language and it is therefore essentialto create a common ground.

In technical terms we create a common ground by conducting *"interoperability"*. Interoperability can be managed at different layers. In the DRIVER Guidelines we basically try to reach interoperability on two layers, syntactical (Use of OAI-PMH & Use of OAI_DC) and semantic (Use of Vocabularies).

Table of Contents

Introduction

Acknowledgements & Contributors (version 1.0)

Martin Feijen, Maurice Vanderfeesten, Wolfram Horstmann, Friedrich Summann, Muriel Foulonneau, Karen Van Godtsenhoven, Patrick Hochstenbach, Paolo Manghi, Bill Hubbard

Acknowledgements & Contributors (version 2.0)

The creation of the DRIVER Guidelines 2.0 relies on the expertise of many people. All these people are experts and repository managers. This group has worked together to achieve interoperability in an way that can be implemented practically. The people below therefore endorse and support the DRIVER Guidelines 2.0.

Editors

- Maurice Vanderfeesten , (SURFfoundation, the Netherlands)
- Friedrich Summann, (University Bielefeld, Germany)
- Martin Slabbertje , (Utrecht University, the Netherlands)

Experts & Reviewers

- Stefania Biagioni , (CNR, Italy)
- Paolo Manghi, (CNR, Italy)
- Maria Bruna Baldacci, (CNR, Italy)
- Friedrich Summann, (University Bielefeld, Germany)
- Martin Slabbertje , (Utrecht University, the Netherlands)
- Thomas Place , (Tilburg University, the Netherlands)
- Benoit Pauwels , (Universite Libre de Bruxelles, Belgium)
- Patrick Hochstenbach , (Ghent University, Belgium)
- Karen van Godtsenhoven, (Ghent University, Belgium)

status: final 2008-11-13

- Niamh Brennan, (Trinity College Dublin, Ireland)

- Phil Cross , (Intute and the Intute Repository Search project, United Kingdom)

- Mikael Karstensen Elbæk , (Danish Technical University (DTU), Denmark)

- Maurice Vanderfeesten , (SURFfoundation, the Netherlands)

- Susanne Dobratz , (Humbolt University, Berlin, Germany)

- Frank Scholze, (Stuttgart University Library, Germany)

- Wolfram Horstmann , (University Bielefeld, Germany)

- Barbara Levergood , (University Goettingen, CACAO project)

- Eloy Rodrigues , (Universidade do Minho, Portugal)

- Arjan Hoogenaar, (KNAW, the Netherlands)

- Armand Guicherit, (KNAW, the Netherlands)

- Ruud Bronmans, (KNAW, the Netherlands)

- Jos Odekerken, (University of Maastricht, the Netherlands)

- Alenka Kavcic-Colic, (Library Research Centre at National and University Library, Slovenia)

- Myriam Bastin, (University of Luik, Belgium)

- Birgit Schmidt, (University of Goettingen, Germany)

About DRIVER

What DRIVER is

DRIVER, the "Digital Repository Infrastructure Vision for European Research" project is conducted by an EC funded consortium that is building an organisational and technological framework for a pan-European data-layer, enabling the advanced use of content-resources in research and higher education. DRIVER develops a service-infrastructure and a data-infrastructure. Both are designed to orchestrate existing resources and services of the repository landscape.

DRIVER as data-infrastructure

The data-infrastructure relies on locally hosted resources such as scientific publications that are collected in digital repositories of institutions and research organisations. These resources will be harvested by DRIVER and aggregated at the European level. In order to ensure a high quality of the aggregation, DRIVER will provide any means possible to harmonise and validate it. DRIVER will respect the provenance of resources by "branding" them with information of the local repository. DRIVER will further point to the local repository when a resource is downloaded instead of providing the resource itself. DRIVER will make its data available for re-use via OAI-PMH to all partners in the DRIVER network of content providers.

The current DRIVER information space

The starting phase of DRIVER has laid the cornerstones for a rich and ambitious pan-European repository infrastructure. The landscape of digital repositories is multifaceted with respect to different countries, different resources such as text, data or multimedia, different technological platforms, different metadata policies etc. But there is also a common ground that applies to large parts of this landscape: the major resource-type provided by digital repositories is text and the major approach for offering these textual resources is the Open-Archives-Initiative Protocol for Metadata-Harvesting. Therefore, the current phase of DRIVER is focusing on textual resources that can be harvested with OAI-PMH.

Challenges

What researchers expect

Researchers and other users of digital information systems have high expectations for provision of digital content. Retrieval should be fast, direct (within a few clicks) and versatile. The current culture in the landscape of digital repositories does not fully support these expectations. While many valuable services have been established to search and retrieve bibliographic records (metadata), the resource itself is sometimes hidden behind several intermediate pages, obscured by authorization procedures, not fully presented or not retrievable at all. Optimal scholarly communication, however, would require the full resource being just one click away. Moreover, an easy retrieval

of full-text and metadata facilitates the machine-based exploitation of content. Neither the harvested bibliographic record nor the crawled full-text on their own can enable the development of integrated, advanced services such as subject-based search combined with browsing through classifications, citation analysis and the like, but instead only the combination of both can enable this.

The full-text challenge

Fostering the direct access to textual resources has been identified as a major challenge within the DRIVER test-bed. While the DRIVER consortium dedicates any effort possible to approach this challenge technologically by processing the aggregated data, hosts of digital repositories can support DRIVER locally by offering content in a specific manner. The DRIVER Guidelines presented here will provide an orientation for local content providers how they should offer their content.

What's next?

Retrieval of full-text with bibliographic data is a basic but necessary step forward to approach rich information services based on digital repositories. Future DRIVER Guideline versions related to the DRIVER II activities will elaborate on further steps with respect to other information types such as primary data or multimedia and on more complex information objects that are made up of several resources.

About the DRIVER Guidelines

Why use the DRIVER Guidelines?

The "DRIVER Guidelines for Content Providers: Exposing textual resources with OAI-PMH" will provide orientation for managers of new repositories to define their local data-management policies, for managers of existing repositories to take steps towards improved services and for developers of repository platforms to add supportive functionalities in future versions.

How to comply with the DRIVER Guidelines? (validation)

DRIVER offers to local repositories in the near future means to check the degree of conformance with the guidelines via web-interfaces.[1] DRIVER also offers web-support (see below "Is there support?"). If the mandatory characteristics of the DRIVER Guidelines are met, a repository receives the status of being a validated DRIVER provider. If recommended characteristics are met, a repository receives the status of a future-proof DRIVER provider. Validated DRIVER repositories can re-use DRIVER data for the development of local services. They become part of the DRIVER network of content providers.

What if I don't comply?

Not conforming to all mandatory or recommended characteristics of the DRIVER Guidelines does not necessarily mean that contents of a repository will not be harvested or aggregated by DRIVER. But, depending on the specific services offered through the DRIVER infrastructure, contents of these repositories might simply not be retrievable. A search service, for example, that promises to list only records that provide a full-text link cannot process all contents of a repository that offers metadata-only records or obscures full-texts by authorization procedures. The DRIVER Guidelines shall help to differentiate between those records. The DRIVER Guidelines will, of course, not prescribe which records should be held in a local repository.

Is there support?

DRIVER offers support to local repositories to implement the DRIVER Guidelines on an individual basis. Support can be delivered through the internet[2] or can be personal[3]. DRIVER is committed to any possible solution that can be realised by central data-processing. But the sustainable, transparent and scalable road to improved services goes through the local repositories.

---

[1] For the Validation of the 1.0 guidelines see:
http://validator.driver.research-infrastructures.eu/

[2] DRIVER Support website: http://www.driver-support.eu

[3] See document "Advice for implementation of the DRIVER guidelines",
www.driver-support.eu/documents/Advice_for_implementation_of_the_DRIVER_guidelines.pdf

Scope of the DRIVER Guidelines

Are the DRIVER Guidelines a standard?

No. Although the use of standards like OAI-PMH certainly does provide a solid base to build a network like DRIVER, there is a need for additional DRIVER Guidelines. The main reason is that the standards still leave room for local interpretation and local implementation. Without that, a standard could not exist. But this openness becomes a hurdle to achieve high quality services when different implementations are combined.

Are the DRIVER Guidelines the same as cataloguing rules?

No. The guidelines are an instrument to map (or translate) the metadata used in the repository to the Dublin Core metadata as harvested by DRIVER. They are not meant to be used as data entry instructions for metadata input in your repository system.

Do the DRIVER Guidelines contain scientific quality level instructions?

No. The guidelines do not tell you what resources have the required quality level for the scientific content and which ones do not. We assume that this distinction has already been made at the repository's institutional level. In other words, we assume that the quality of the resources exposed through harvesting is good enough.

What are the main components of the DRIVER Guidelines?

The DRIVER Guidelines basically focus on five issues: collections, metadata, implementation of OAI-PMH, best practices and vocabularies and semantics.

- With respect to collections within the repository the use of "sets" that define collections of open full-text is mandatory. If all resources in the repository are textual, include not only metadata but also full-text and all resources are accessible without authorization, the use of sets is optional.
- With respect to the OAI-PMH protocol some mandatory and some recommended characteristics have been defined in order to rule out problems arising from the different implementations in the local repository.

13/137   status: final 2008-11-13

- With respect to metadata some mandatory and some recommended characteristics have been defined in order to rule out semantic shortcomings arising from heterogeneous interpretations of DUBLIN CORE.

Who stands behind the DRIVER Guidelines?

The DRIVER Guidelines have been compiled by people who have years of experience with the construction and maintenance of similar networks of interlinked repositories such as HAL in France, DARE in the Netherlands, DINI in Germany, SHERPA in the UK and they involve expertise from experienced service providers such as BASE and community organizations such as the OAI Best-Practice group.

What do you mean with textual resources?

In this phase of DRIVER we focus on textual resources. As working definitions we use the following:

- **A textual resource**: scientific articles, doctoral theses, working papers, e-books and similar output of scientific research activities
- **Open Access**: access without any form of payment, licensing, access control with password etc, technical access control with IP etc

Many repositories are used to depositing different types of resources, for example, articles, e-books, photographs, video, datasets and learning materials. These resources have metadata records that describe them. Usually the resources are in a digital form (but not always) and these digital files are usually stored within a database that is part of the repository system (but not always). Access to the resources is usually open (but not always). Within DRIVER we focus on a subset of the vast domain of resources in European repositories: we focus on textual resources in digital form that are open access.

Research shows that in doing this we will cover more than 80% of all available resources. For this reason the first mandatory guideline of Part A states: "the repository contains digital textual resources". This doesn't mean that your repository might not include other materials and non-digital items also. The statement is an

expression of the DRIVER focus on textual resources. A complete list of the textual resources is presented in element dc:type in the metadata guidelines in chapter *"Use of Vocabularies and Semantics"* section *"Publication type"*. For the implementation in dc:type see chapter "Use of Metadata OAI_DC" section "Type". Or to map with currently known type mappings see section "DRIVER-TYPE Mappings" in the chapter "Use of Best Practices for OAI_DC".

What do you mean by "sets"?

Sets are a standard component of the OAI-PMH protocol and they are used to focus (filter) specific parts of a repository. When your repository contains also non-textual items, or non-digital items, or toll gate access items or metadata only items, you can use the "set" mechanism to filter out these items when offering your content to DRIVER.

Further Resources

What else should I consider?

Existing resources have been used as input for these DRIVER Guidelines and much care has been taken to avoid special solutions. In this way, one could say that the DRIVER Guidelines utilize practical experience and worldwide existing guidelines.

- DRIVER is modelled after established and operational, distributed networks of content providers, particularly DARE in the Netherlands. The guidelines for DARE serve as a model for DRIVER. Rather than providing multiple references to guidelines scattered worldwide, DRIVER has initially made use of the DARE Guidelines and enhanced these guidelines by adopting best practises from repository managers and experts all over the European continent. The following documents have been an especially important starting point of, and essential to, the DRIVER Guidelines:
  - o The document "USING SIMPLE DUBLIN CORE TO DESCRIBE EPRINTS", by Andy Powell, Michael Day and Peter Cliff of UKOLN, University of Bath (Version 1.2), which has been adapted for specific requirements by the DARE programme historically known as "DRIVER Use of Dublin Core"

(Version 2, November 2006), has been extended in the DRIVER Guidelines 2.0 with the aid from repository managers - see chapter *"Use of Metadata OAI_DC"*

o The Open Archives Initiative Protocol for Metadata Harvesting, Protocol version 2.0, which also has been adapted by DARE for specific requirements and is available as the "DRIVER use of OAI-PMH guidelines" (Version 2, December 2006) has been extended in the DRIVER Guidelines 2.0 with the aid from repository managers - see chapter "Use of OAI-PMH"

o The DINI-Certificate "Document and Publication Services 2007" (Version 2, September 2006)[4] provides a solid basis for what to consider when operating a repository. Since DRIVER looks at repositories from the perspective of an aggregator, the DRIVER Guidelines do not cover the aspects described in the DINI-Certificate that is designed for guiding the overall local operation of a repository. Instead, the DRIVER Guidelines are based on the assumption that the criteria of the DINI certificate are considered in the operation of a repository.

o The document "Use of MODS for institutional repositories"[5] was created by the Metadata expert group of the SURFshare programme and used by the Dutch repositories. These guidelines provide a practical list of Publication types that ensures greater interoperability. The Publication types are based on the dc:type Publication list from the "DARE use of DC" document, combined with e-prints types and Publication types used in METIS in the wide spread Dutch Current Research Information System (CRIS).

---

[4] http://www.dini.de/documents/dini-zertifikat2007-en.pdf

[5]

https://www.surfgroepen.nl/sites/oai/metadata/Shared%20Documents/Use%20of%20MODS%20for%20institutional%20repositories-version%201.doc

o The Version Identification Framework[6] delivered a simple and practical Version taxonomy[7] for journal articles and more. This formed an addition to describe the Publication types even better in the scholarly workflow.

Is there a working solution that solves many problems at once?

Yes, see chapter "Use of MPEG-21 DIDL (xml-container) - Compound object wrapping". Within the SURF DARE programme it has proven useful to implement an "XML-Container" for each resource that allows resource harvesting within OAI-PMH, provides an unambiguous link to the resource (not via a jump off page), supports full text indexing and enables the representation of complex documents consisting of several PDF files. The XML-Container is based on the Digital Item Declaration Language (MPEG21-DIDL)[8]. Other solutions based on DIDL have also been developed (e.g. aDORe[9], METS profiles[10]) and further to be published in the future (e.g. OAI-ORE [11]).

Outline – DRIVER Guidelines Summary

The following outline summarises the basic DRIVER settings for the basic topics textual resources, metadata usage and OAI-PMH protocol implementation. The elaborated details can be found in the following chapters.

PART A - Textual Resources

mandatory

---

[6] http://www.lse.ac.uk/library/vif/Framework/Essential/taxonomies.html

[7] http://www.lse.ac.uk/library/versions/

[8] http://xml.coverpages.org/mpeg21-didl.html

[9] http://african.lanl.gov/aDORe/projects/adoreArchive/

[10] http://www.loc.gov/standards/mets/mets-profiles.html

[11] http://www.openarchives.org/ore/

- The repository contains digital textual resources (see explanation "What do you mean with textual resources?" on page 14)
- Textual resources have popular and widely-used formats (PDF, TXT, RTF, DOC, TeX etc.)
- Textual resources are open access, available directly from the repository for any user worldwide without restrictions such as authorisation or payment
- Textual resources are described by metadata records
- Metadata plus textual resource are linked together in such a way that an end user can access the textual resource through an identifier (usually a URL) in the metadata record
- The URL of a resource once encoded in the metadata record is permanently addressable and is never changed or re-assigned
- A unique identifier identifies the metadata record and the textual resource (no pointers to external systems such as a national library system or a publisher)

recommended

- Transparent verification of the integrity of a textual resource
- Quality (of the scientific content) assurance measures for the textual resources exposed such as a limitation to those textual resources included in the yearly scientific report (or equivalent)
- The URL of the textual resource as encoded in the metadata record is based on a persistent identifier scheme such as DOIs, URNs, ARKs
- The use of the DIDL XML-container for exposing textual resources (chapter "Use of MPEG-21 DIDL (xml-container) - Compound object wrapping")

PART B - Metadata

mandatory

- Metadata are structured as Unqualified Dublin Core (ISO 15836:2003)
- Individual elements of DC are to be used according to the chapter "Use of Metadata OAI_DC" on page 52

recommended

- Preferably use Metadata that is structured according to more comprehensive schemes such as Qualified Dublin Core or MODS. (Guidelines for these comprehensive schemas will follow in the future version of the DRIVER Guidelines.[12])
- Recommended language is English
- Recommended language for an abstract (including an abstract is optional) of the article is English

PART C - OAI-PMH Implementation

mandatory

- The repository must be OAI-2.0 compliant and must conform to the specification on chapter "Use of OAI-PMH" on page 35
- Existence of a repository identifier and use of the OAI identifier scheme
- If (and only if) the repository contains resources other than those which are mandatory in "PART A - Textual Resources", an OAI-set is defined as that which identifies the collection of digital textual resources accessible in Open Access (see explanations "DRIVER Set naming", "DRIVER Set Content definitions" and "Set Location" on pages 42-44)

recommended

- Provisions for the change of Base-URL
- Completeness of Identify Response, including use of the optional Description statement
- Use persistent of Transient deleting strategy
- Use a batch size with corresponding resumption token expiration time.

---

[12] Preview of the MODS guidelines

https://www.surfgroepen.nl/sites/oai/metadata/Shared%20Documents/Use%20of%20MODS%20for%20institutional%20repositories-version%201.doc

What's New

Chapter 1: Use of OAI-PMH

DRIVER Set naming

Added information to answer questions about "Recommended Set names for "Open Access" *and* "Embargoed/Delayed Access" *subcollections* –

Explanation: Recommended for hybrid repositories with a mixture of metadata-only and metadata-with-full-text to use a DRIVER set with records that contain the full text openly available. Also the DRIVER set should not contain Delayed Access records, this only leads to confusion at the end-user's side when he thinks to find Open Access material.

There should be not be separate DRIVER recommendations on sets for eTheses.

Explanation: DRIVER Guidelines are there for a bigger community. Harvested eTheses should be recognised through the terms used in the Publication type vocabulary.

Harvest batch size

Increase the recommended batch size from 100-200 records per batch, to 100-500 records per batch. See: Harvest batch size on page 41.

Explanation: The experience is that problems with breaks in a OAI ListRecords communication happen quite rarely. The topscore of records per response found up to now was around 6500 records. The positive consequence of a hugh batch size is that the harvesting activity is very quick and thus those repositories have a high throughput.

Resumption token lifespan

Beter explanation why the recommendation of the Resumption token lifespan is needed. See: Resumption token on page 40.

Explanation: There is a relation between the lifespan, batch size and throughput. If the throughput is slow and the batch size is small, the life span of the resumption token should increase. Otherwise the harvester keeps receiving only the first batch over and over again.

Deleted records strategy

The DRIVER Guidelines text explains clearer now why a persistent/transient strategy is valuable for both repository and service provider.

Explanation: The advantage for the repository to keep track of deletions is that a service provider will not display records which are not available anymore in the repository. Besides that, this strategy allows harvesters to avoid re-loading the full repository each time and makes the harvesting process more efficient.

See: Deleted records on page 39.

Chapter 2: Use of Metadata OAI_DC

Identifier

How to handle other identifiers that are in the repository. Are OAI identifiers allowed? Where should the identifier point to? How should they be exposed?

Explanation The Identification of a resource has been broadened. The repository can use any identifier that is necessary to identify the resource. However, there must be at least one **actionable** identifier that points to the jump-off page with the full text document or directly to the full text document. In case of more than one actionable identifier, the service provider will use, by default, the first actionable identifier in the list to direct the end-user to. See: Identifier on page 73.

Date

What to do when the date recommended in the DRIVER Guidelines (date of creation) is not available in the repository?

In the DRIVER Guidelines: "Use the DC element 'date' for the value [of the refinement": > date Published. The Preferred date is date Published, because this is the most meaningful and useful date for the end-user. If no date Published is available, use any other date available. It is better to use one date then no date at all." See: Date on page 66.

Explanation: Two changes have occurred:

1. The date created has changed to date published; because this is the most meaningful for the end user
2. If this does not apply, use the next best or most appropriate date to use; better some date then no date at all!

What to do with multiple date fields?

In case of OAI-DC, only use one date field, preferrably the publication date. **Explanation**: more then one date fields create ambiguity since simple DC cannot hold qualifiers. By default a service provider uses the first date in the list to use for processing, indexing and presentation.

See: Date on page 66.

Rights

Explanation on how to use the dc:rights field. See: Rights on page 79.

Language

The encoding recommendation has changed to ISO 639-3. Plus reassurance that ISO 639-1 and -2 are still allowed, since they can be mapped properly.

Explanation: ISO 639-3 encoding has many more languages then ISO 639-1, even historical languages and sub-region languages. This makes it better to explain certain publications. ISO 639-2 has two encoding types (b and t), which makes it ambiguous when used in OAI-DC. The latter does not provide an attribute that notifies which of the two encoding scheme has been used.

See: Language on page 76.

Creator

According to the DRIVER Guidelines: "Usage instruction When initial and full name are *both available* use this formatting: <dc:creator> Janssen, J. (John)</dc:creator>"

COMMENT: In the usage instruction context, what does *both available* mean?

Changed *full* name and *fore* name to *first* name.

Explanation: It is recommended to use a standardized writing style for names, so use the writing style used by the publisher in the first place. When that is not applicable use the APA bibliographic writing style as in a reference list when applicable. When both the initial(s) and first name(s) (referring to that initial) of a person is/are available, use the formatting where the first name is written between curved brackets after the APA styled name. The syntax should then be: {surname}, {initials} ({first name})

For example

- John Kennedy becomes: *Kennedy, J. (John)*
- John F. Kennedy becomes: *Kennedy, J.F. (John)*
- John Fitzgerald Kennedy becomes: *Kennedy, J.F. (John, Fitzgerald)*
- and J.F. Kennedy becomes: *Kennedy, J.F.* because the full first name was not available.

See: Creator on page 59.

Source

Broken link in *Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata*. Changed http://epub.mimas.ac.uk/DC/dc-citation-guidelines/ to http://dublincore.org/documents/dc-citation-guidelines/

Type

vocabulary change

Due to the ongoing confusion in the international repository community about the terms for the Publications types, DRIVER Guideline experts have developed two

separate vocabularies. One that explains the *naked* Publication type and one that explains the versions used in scholarly communication. The version types can be added to the Publication types to create more depth that explains the publication even more.

The Publication types are well thought-of types that do not explain the type of document, but the type of publication. These publications have been used in common scholarly processes. The terms are chosen to create a balance between not too specific (that it only applies to one research community) and not too generic.

Another thing that was lacking is a namespace that creates a level of authority of a controlled vocabulary. The URI info:eu-repo namespace has been especially been granted by the authorities to be used for this purpose.

By these criteria the DRIVER vocabulary for Publication types has been made.

See: Publication type vocabulary on page 115.

For the Version types see: Version vocabulary on page 120.

discussion on terms

Difference between Conference report and Conference lecture?

Explanation: Differences have been removed by abstracting to a more general term "Conference Object".

Map public project deliverables into External Research Report, technical reports into Research paper, editorials into Article?

Explanation: Mappings have been made. See: DRIVER-TYPE Mappings on page 83. Descriptions of the terms have been provided.

Format

Explanation: on the limitations of the list of formats. This list is just a subset of all common formats that could be used in this field. We have added Open Document Text: vnd.oasis.opendocument.text. A more extensive list can be found on http://www.iana.org/assignments/media-types/

See Formaton page 71.

Chapter 3: Use of Best Practices for OAI_DC

DRIVER-TYPEMappings

Explanation: how to map [x] Local categories to [y] DRIVER categories.

See: DRIVER-TYPE Mappings on page 83.

DRIVER-VERSION Mappings

Explanation: how to use the different status/versions of Publication and to map [x] Local categories to [y] DRIVER (version) categories.

See DRIVER-VERSION Mappings on page 86.

Use of OAI_DC with Theses

Explanation: how to use OAI_DC with e-Theses and Dissertations without losing interoperability. See Use of OAI_DC with Theses on page 87.

DC:SOURCE and DC:RELATION

status: final 2008-11-13

Explanation: how to use the DC:source and dc:relation fields with respect to scholarly communication and repositories.

See: DC:SOURCE and Citation information on page 89 and DC:RELATION and Linking related objects on page 90.

Chapter 4: Use of Compound Object Wrapping

Several major important changes have been made

- Wrong DIDL schema location, validation not possible
- Modify reference of info:eu-repo namespace
- Modifications are also put in the example
- Changes to meet future transport of Author Identifiers

Add namespace and change to valid namespace location

```
<didl:DIDL
xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:diext="http://library.lanl.gov/2004-04/STB-RL/DIEXT"
xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="
urn:mpeg:mpeg21:2002:02-DIDL-NS                http://purl.lanl.gov/STB-
RL/schemas/2004-08/DIDL.xsd              urn:mpeg:mpeg21:2002:01-DII-NS
http://purl.lanl.gov/STB-RL/schemas/2003-09/DII.xsd
http://library.lanl.gov/2004-04/STB-RL/DIEXT   http://purl.lanl.gov/STB-
RL/schemas/2004-04/DIEXT.xsd">
```

Becomes:

```
<didl:DIDL
xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
xmlns:dip="urn:mpeg:mpeg21:2005:01-DIP-NS"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="
urn:mpeg:mpeg21:2002:02-DIDL-NS
        http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/did/didl.xsd
urn:mpeg:mpeg21:2002:01-DII-NS
        http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/dii/dii.xsd
urn:mpeg:mpeg21:2005:01-DIP-NS
        http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/dip/dip.xsd">
```

Changes of container element to create beter semantic interpretation

```
<didl:DIDL>
<didl:Container>
                <didl:Item>…</didl:Item>
                <didl:Item>…</didl:Item>
                <didl:Item>…</didl:Item>
</didl:Container>
</didl:DIDL>
```

Becomes:

```
<didl:DIDL>
<didl:Item>
                <didl:Item>…</didl:Item>
                <didl:Item>…</didl:Item>
                <didl:Item>…/didl:Item>
</didl:Item>
</didl:DIDL>
```

Changes of Object type declaration per aggregated item

```
<didl:Descriptor>
        <didl:Statement mimeType="text/plain">metadata</didl:Statement>
</didl:Descriptor>
```

Becomes:

```
<didl:Descriptor> <!-- ObjectType of Item -->
        <didl:Statement mimeType="application/xml">
                <dip:ObjectType>info:eu-
repo/semantics/descriptiveMetadata</dip:ObjectType>
        </didl:Statement>
</didl:Descriptor>
```

- 'object' *becomes* 'objectFile'
- 'Jump-off-Pageâ′ *becomes* 'humanStartPage'

Text convention is *camelCase* that starts with small caps.

Use of Persistent Identifier in DIDL

This explains the position of the Persistent Identifier and the "Location to be used for Resolution mechanisms".

status: final 2008-11-13

At the top level Item Element a Component/Resource Element must be added that refers to the actionable URL of this DIDL document without the OAI-PMH elements. When this is not applicable right now, just use the URL of the Human Start Page.

```
<didl:DIDL>


<didl:Item>
<didl:Descriptor>
                <didl:Statement mimeType="application/xml">
                        <dii:Identifier>urn:NBN:nl:ui:10-
1705/6748398729821</dii:Identifier>
                </didl:Statement>
        </didl:Descriptor>
...
        <didl:Component>
                <!-- Actual resource of Item -->
                <didl:Resource mimeType="application/xml"
                                ref="http://localhost/xmlContainer-
v2.3.xml"/>
        </didl:Component>


        <didl:Item>...</didl:Item>
        <didl:Item>...</didl:Item>
        <didl:Item>...</didl:Item>

</didl:Item>


</didl:DIDL>
```

Generic metadataPrefix in OAI-PMH

This explains the real DIDL is used and not a derived scheme.

```
<request metadataPrefix="dare_didl"
```

Becomes:

```
<request metadataPrefix="didl"
```

Chapter 5: Use of Vocabularies and Semantics

Two vocabularies have been made to de-ambiguify the concepts and terms used in scholarly communication in Europe.

Several more issues therefore have been solved:

- Document type : Preprint and Postprint versioning
- Document type: What is the difference between "external research report" and "internal report"?
- Improve Document type vocabulary
- Question if *bookChapter* in the info:eu-repo vocabulary should be more generic for improved interpretation of Service providers - to a combination of terms e.g. *chapter* and *partOf* ? Answer: NO.
- Versioning of Journals - improved model

A chapter on the usage of classification information has been added.

It is recommended to deliver information on the classification usage in a repository in the Identify response and to transport the classification in the element subject "URI-fied" using an authorative namespace. If no specific slassification scheme is used, DRIVER recommends the Dewey Decimal Classification.

Chapter 6: Annex: Use of Quality labels

See Annex: Use of Quality Labels on page 124 for a starting document.

The DRIVER Guidelines 2.0 provides basic information on the importance of Quality, and Interoperability. Quality labels can be used to assure stable and reliable repositories that last longer than the hype, and have also an archival purpose for long term preservation.

Examples of Quality labels can be: the Data Seal of Approval and the DINI Certificate.

Chapter 7: Annex: Use of Persistent Identifiers

See Annex: Use of Persistent Identifiers on page 125 for a starting document.

Persistent Identifiers for web resources are needed to create a stable and reliable infrastructure. This does not concern technicalities, but mainly agreements on an organisational level.

The DRIVER Guidelines could make some recommendations on the implementation for repository managers. At the basis lies the Report on Persistent Identifiers of the PILIN project.

An implementation plan has been provided.

Chapter 8: Annex: Use of Usage Statistics Exchange

See Annex: Use of Usage Statistics Exchange on page 131 for a starting document.

In order to see the value of Open Access and offer extra services to your authors, repositories should think about aggregating usage statistics.

Two projects will gain insights and help develop guidelines for the exchange of usage statistics: PIRUS and OA-Statistik

Chapter 9: Annex: Use of Intellectual Property Rights (IPR)

See Use of Intellectual Property Rights (IPR) on page 136 for a starting document.

This addresses an important issue on Usage Rights and Deposit Rights. In practice this must be implemented. The DRIVER Guidelines should tell something on how Usage Rights and Access rights should be exposed and formatted in metadata.

Use of OAI-PMH

Introduction

This chapter explains how to use OAI-PMH in a way so that repositories and service providers can seamlessly work together by creating interoperability on a protocol level.

Remark:

The examples used for DIDL; do NOT use them literally! For the precise use of the DIDL document see the current version of the DIDL document specification. That document will overrule all DIDL examples mentioned here.

Acknowledgements

This document is largely based on discussions between repository managers and SURF. They have offered their experience and suggestions to create the DRIVER Guidelines as presented in this document.

Source material

The DRIVER Guidelines are based on and refer to, the Open Archives Initiative Protocol for Metadata Harvesting, Protocol version 2.0.

See: http://www.openarchives.org/OAI/openarchivesprotocol.html

The order of presentation of the DRIVER Guidelines is the same as in the protocol text. When useful, the protocol text is quoted. When the text has been changed, e.g. bold added to highlight some part of the text, this has been indicated between brackets.

Definitions and concepts: item, record and unique identifier

Item and Record

It is important to make a distinction between Item and Record. The protocol text states:

"…An item is conceptually a container that stores or dynamically generates metadata about a single resource in multiple formats, each of which can be harvested as records via the OAI-PMH …A record is metadata expressed in a single format. A record is returned in an XML-encoded byte stream in response to an OAI-PMH request for metadata from an item…"[bold added by MF]

Within DRIVER it is recommend to construct the XML-encoded stream according to the XML- Container specifications. These specifications are given below.

Identifier

The Unique Identifier identifies an item within a repository. Do not confuse this identifier with the element dc:identifier in Dublin Core. The OAI identifier has a

different function: it is used to extract metadata, whereas the DC identifier is used to extract the resource. Schematically:

```
                        ┌──────────────────────────┐
                        │ Item with Unique Identifier │
    ↑                   │                          │
    │   ┌──────────┐    │                          │
    │   │ Inside    │   │                          │
    │   │ repository │  └───────┬──────────┬───────┘
    │   └──────────┘           │          │
 ───┼──────────────────────────┼──────────┼─────────────
    │   ┌──────────┐   ┌────────▼──────┐  ┌▼──────────────┐
    │   │ Outside   │   │ Record with   │  │ Record with   │
    │   │ respository│  │ XML-encoded   │  │ XML-encoded   │
    │   └──────────┘   │ metadata,     │  │ metadata,     │
    ↓                  │ e.g. in simple │  │ e.g. in MARC-21│
                       │ DC            │  │               │
                       └───────┬───────┘  └───────┬───────┘
                               │                  │
                       ┌───────▼───────┐  ┌───────▼───────┐
                       │               │  │               │
                       │  Harvester A  │  │  Harvester B  │
                       │               │  │               │
                       └───────────────┘  └───────────────┘
```
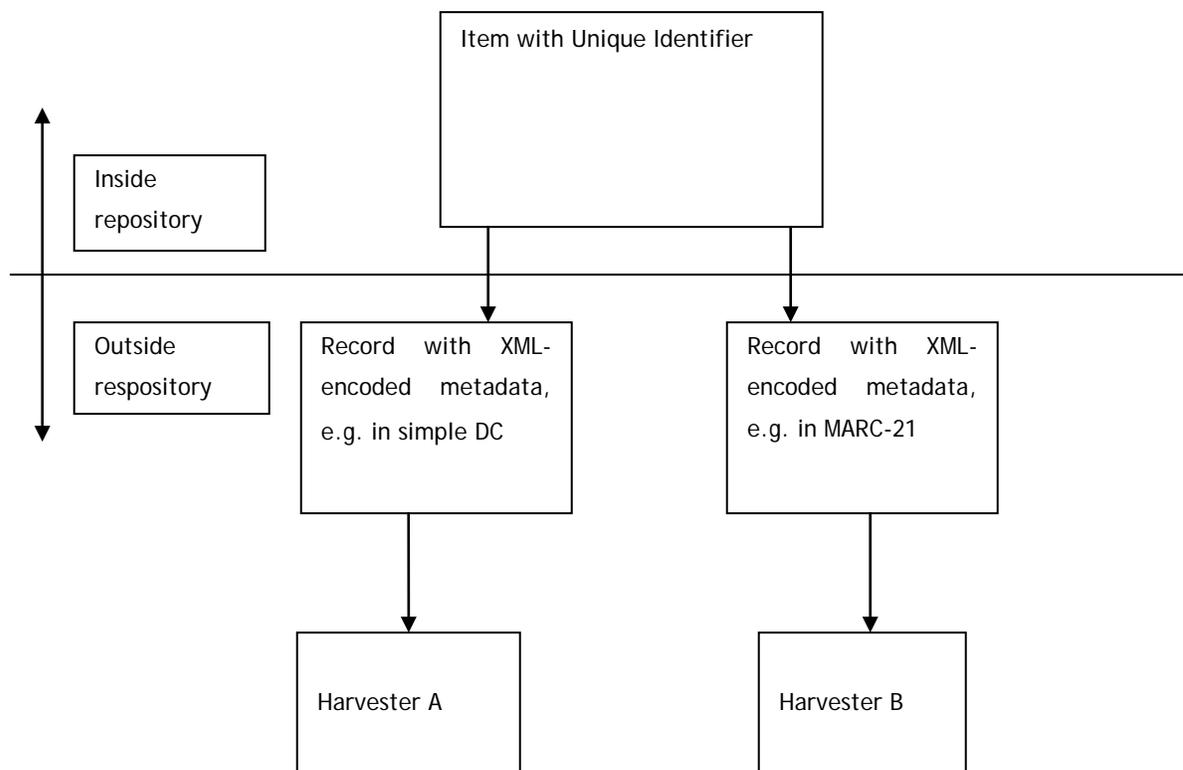
MetadataPrefix naming

See:

http://www.openarchives.org/OAI/openarchivesprotocol.html#MetadataNamespaces

OAI-PMH supports the dissemination of records in multiple metadata formats from a repository. The `ListMetadataFormats` request returns the list of all metadata formats. metadataPrefix arguments are used in `ListRecords`, `ListIdentifiers`, and `GetRecord` requests the retrieval of records, or the headers of records that include metadata in the format specified by the metadataPrefix. For purposes of interoperability, repositories must disseminate Dublin Core, without any qualification. Therefore, the protocol reserves the metadataPrefix ‘`oai_dc`’, and the URL of a

metadata   schema   for   unqualified   Dublin   Core,   which   is
http://www.openarchives.org/OAI/2.0/oai_dc.xsd. The corresponding XML namespace
URL is http://www.openarchives.org/OAI/2.0/oai_dc/.

DIDL document

The DRIVER community supports the implementation of the metadataPrefix `oai_dc`
and the metadataPrefix `didl`. Every DRIVER repository that uses the XML container
must support this `didl` metadata schema. The specification of the `didl`
XMLcontainer can be found in chapter Use of MPEG-21 DIDL (xml-container) -
Compound object wrapping on page 91.

```
<OAI-PMH ...>
    <...>
       <record>
         <metadata>
          <didl:DIDL>
             <didl:Item>...</didl:Item>
```

Datestamp

According to the protocol, each record contains a header with a datestamp with *"the
date of creation, modification or deletion of the record for the purpose of selective
harvesting."*

The protocol also explains the selective harvesting as follows:

- "...*modification* - the response must include records, corresponding to the
  metadataPrefix argument, which have changed within the bounds of the
  from and until arguments

status: final 2008-11-13

- *creation* - the response must include records, corresponding to themetadataPrefix argument, that have become available from the repository within the bounds of the from and until arguments

- *deletion* - depending on the level at which a repository keeps track of deleted records, the response may include headers of records, corresponding to the metadataPrefix argument, which have been withdrawn from the repository within the bounds of the from and until arguments. Deleted status is indicated via the status attribute of the header element and no metadata is included…"▫

It is very, very important to take great care in implementing the datestamp according to the protocol specifications as quoted above. Experience has taught that many harvesting errors that occur with incremental harvesting have their origin in misinterpretation of the datestamp.

Datestamp syntax

See: http://www.openarchives.org/OAI/openarchivesprotocol.html#Datestamp , http://www.openarchives.org/OAI/openarchivesprotocol.html#Dates and http://www.w3.org/TR/NOTE-datetime

The value of datestamps in both requests and responses must comply with the specifications for UTCdatetime in that document. The DRIVER agreement supports the use of optional granularity which involves the time with seconds YYYY-MM-DDThh:mm:ssZ.

This value complies with the specifications for the UTCdatetime in sections 3.3.1 in the OAI-PMH document. Datestamps are encoded using ISO8601 and are expressed in UTC.

```
<OAI-PMH ...>
   <...>
      <GetRecord>
         <record>
            <header>
               <datestamp>2001-12-14T12:01:45Z</datestamp>
```

A repository that supports YYYY-MM-DDThh:mm:ssZ should indicate this in the Identify response.

```
<OAI-PMH ...>
   <...>
   <Identify>
      <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
   <...>
```

Deleted records

See: http://www.openarchives.org/OAI/openarchivesprotocol.html#DeletedRecords

If a record is no longer available then it is said to be deleted. Repositories must declare one of three levels of support for deleted records in the deletedRecord element of the Identify response:

- *no* - the repository does not maintain information about deletions. A repository that indicates this level of support must not reveal a deleted status in any response
- *persistent* - the repository maintains information about deletions with no time limit. A repository that indicates this level of support must persistently keep track of the full history of deletions and consistently reveal the status of a deleted record over time
- *transient* - the repository does not guarantee that a list of deletions is maintained persistently or consistently. A repository that indicates this level of support may reveal a deleted status for records

The DRIVER Guidelines request the DRIVER repositories to use the option `'transient'`. `'persistent'` can also be used. This option makes the harvester do an easier job to detect deleted records.

The advantage of the repository keeping track of deletions is that a service provider will not display records which are not available anymore in that repository. Besides that, this strategy allows harvesters to avoid re-loading the full repository each time and makes the harvesting process more efficient.

Use of transient: When a record is deleted, the repository must indicate the deletion for at least a month. In this period of time most harvesters have updated their database incrementally (without a full re-harvest).

If a repository does keep track of deletions, then the datestamp of the deleted record **must** be the date and time that it was deleted. Responses to `GetRecord` and `ListRecords` requests for a deleted record **must** then include a `header` with the attribute `status="deleted"`. Incremental harvesting will thus discover deletions from repositories that keep track of them.

Resumption token

See: http://www.openarchives.org/OAI/openarchivesprotocol.html#Idempotency Repositories that implement resumptionTokens **must** do so in a manner that allows harvesters to resume a sequence of requests for incomplete lists by re-issuing a list request with the most recent resumptionToken. The purpose of this is to allow harvesters to recover from network or other errors that would otherwise mean that the list request sequence would have to be started again.

The protocol does not mention the life span of a token. A token life span is the time a repository keeps the token stored in memory, along with the resume information. When the life span is too short, the repository does not give the harvester a reasonable time to return to complete the harvest. When this happens the repository

does not comply with the protocol - see above: "*must do so in a manner that allows harvesters to resume…*".

Best practice: a reasonable time for a token to be kept alive is at least twenty four (24) hours. This depends on the size of the repository and the speed of the loading process and thus the resumption token life span should hold for long enough to transport the batch within that period of time.

Along with this life span there is an optimal batch size - see section "Harvest batch size".

Another aspect of the resumption token usage is the optional completeListSize attribute. This should deliver the total size of documents of the response and thus this information can be used during the harvesting process and could be compared with the total result size for control reasons (for example, is the harvest complete or broken?). Besides that, the information could be useful for maintaining the harvesting process in order to estimate the time needed.

A resumption token in an OAI response could look like this (the attributes expirationDate, completeListSize and cursor are optional):

```
<resumptionToken expirationDate="2008-07-14T23:00:24Z"
completeListSize="983" cursor="0">514284267</resumptionToken>
```

Harvest batch size

The batch size is the number of records a repository delivers to the harvester for one resumption token and determines how many request processes have to be executed.

The agreement is that DRIVER repositories must set the batch size between 100 and 500 records.

Using this batch size for all DRIVER repositories will make the harvester operate at optimal performance.

DRIVER Set naming

See: http://www.openarchives.org/OAI/openarchivesprotocol.html#Set

The OAI-PMH document states: Repositories may organize items into sets. Set organization may be flat, i.e. a simple list, or hierarchical.

The DRIVER agreement is that **hybrid** DRIVER repositories that contain metadata-only and metadata-with-full-text resources must support at least one DRIVER set. The DRIVER set is flat and does not have any hierarchical structure. The content of the DRIVER set is **Open Access, Freely available resources**. Delayed Access resources or Embargoed resources must *not* be in this list to avoid confusion at the end-user side. The table below shows the preferred setName and setSpec that can be used to create a DRIVER set.

	setName	setSpec *
The DRIVER set	`Open Access DRIVERset`	`driver`

*A harvester only uses the `setSpec` request to perform selective harvesting. The letters must be in smallcaps.

DRIVER Set Content definitions

The specific content of the 'driver' set is determined at the local repository. A DRIVER repository using this kind of sets must conform to the following rules when inserting a record into the DRIVER set:

- The *DRIVER set contains* records that must contain open access digital textual resources
  - Must contain Full text objects, not metadata-only.
  - Content is Open Accessible

- Content is Not Firewalled

- Content is accessible also Outside the University Campus

- Content is not behind toll-gated websites

-

The picture below shows that is is possible to place one record in different sets. The records below, represented by a blue dot, exist also in the 'driver' set. Two records exist in all three sets. The biochemistry set, the neurophysics set and the driver set. The first two are sets that indicate a subject, the driver set indicates a type (open access). The header of a record can contain zero or more setSpecs. An OAI record might look like this.

```
<record>
  <header>
    <identifier>oai:repository:it/0112017</identifier>
    <datestamp>2002-02-28</datestamp>
    <setSpec>biochemistry</setSpec>
    <setSpec>neurophysics</setSpec>
    <setSpec>driver</setSpec>
  </header>
  <metadata>
    <oai_dc:dc xmlns:oai_dc="http ....
</record>
```

Illustration:
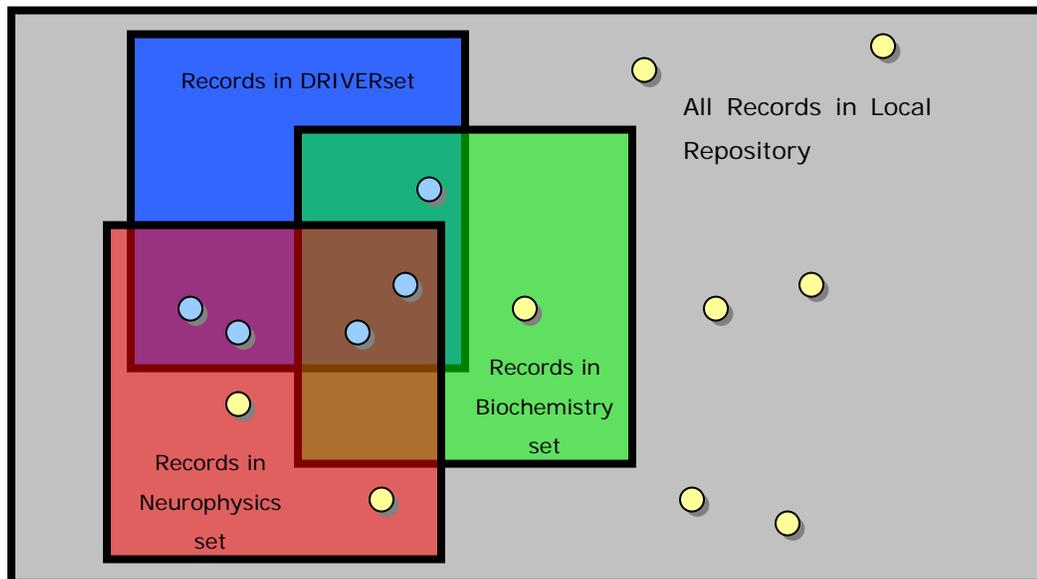
Set Location

The DRIVER set and the other sets can be located at a different locations/baseURLs.

adminEmail for error logging feedback

See: http://www.openarchives.org/OAI/openarchivesprotocol.html#Identify   The repository must provide an administrator e-mail address in the Identify request.

In the near future we want the harvester to give immediate response to the Repository Administrator to inform about the errors this DRIVER repository is creating. See table below for an example of an Identify response which includes the administrators e-mail address.

```
<OAI-PMH ...>

   <...>

   <Identify>

     <adminEmail>somebody@loc.gov</adminEmail>

     <adminEmail>anybody@loc.gov</adminEmail>

     <...>
```

The use of an adminEmail in the `Identify` request is mandatory, and is also dictated by the OAI-PMH protocol. See below:

"The `Identify` verb is used to retrieve information about a repository."

"The response must include one or more instances of the following element:

- adminEmail : the e-mail address of an administrator of the repository."

Descriptive Provenance Information

The description container of the Identify response may be used to deliver additional information on the repository. Service providers may look for this and improve their data processing and the services based on the metadata and their quality.

Best practice: Use this container to describe as many common information about the repository as possible in detail with added examples. This includes used classification schemas(in which format in which element), used vocabularies (type, language), policies and background information.

While the Identify response deals with the repository level, the record level can hold additional information in the `about` element. To allow the service providers to assign harvested material the `provenance` sub-element can be used.

Best practice: Use the `provenance` element in the `about` tag of the metadata to relate to the original document deliverer.

Example:

```
<about>
<provenance xmlns="http://www.openarchives.org/OAI/2.0/provenance"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/provenance

http://www.openarchives.org/OAI/2.0/provenance.xsd">
<originDescription harvestDate="2002-02-02T14:10:02Z" altered="true">
  <baseURL>http://the.oa.org</baseURL>
  <identifier>oai:r2.org:klik001</identifier>
  <datestamp>2002-01-01</datestamp>

<metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadata
Namespace>
  <originDescription harvestDate="2002-01-01T11:10:01Z" altered="false">
    <baseURL>http://some.oa.org</baseURL>
    <identifier>oai:r2.org:klik001</identifier>
    <datestamp>2001-01-01</datestamp>

<metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadata
Namespace>
  </originDescription>
</originDescription>
</provenance>
</about>
```

Prefix & namespace declaration

See: http://www.openarchives.org/OAI/openarchivesprotocol.html#Record

namespace declarations -- the declarations of the namespaces used within the metadata part, each of which is prefixed with xmlns. Namespace declarations within the metadata part fall into two categories:

- metadata format specific namespace(s) - every metadata part must include one or more xmlns prefixed attributes that define the correspondence between a metadata format prefix -- e.g. didl -- and the namespace URI (as defined by the XML namespace specification) of the respective metadata format. Some metadata formats employ tags from multiple namespaces, requiring multiple xmlns prefixed attributes -- in the example below under 'XML validation', there are declarations for both oai_dc and dc.

- xml schema namespace - every metadata part must include the attribute xmlns:xsi, the value of which must always be the URI shown in the example, which is the namespace URI for XML schema.

- xsi:schemaLocation -- the value of which is a "URI, URL" pair; the first is the namespace URI (as defined by the XML namespace specification) of the metadata that follows in this part, and the second is the URL of the XML schema for validation of the metadata that follows.

The recommended use of prefixes and namespaces is that these entities should be declared on the first element of that namespace. This prevents "operational difficulties", as described in http://www.w3.org/TR/REC-xml-names/#ns-using .

"Using prefixes may lead to operational difficulties in the case where the namespace declaration attribute is provided, not directly in the XML document entity, but via a default attribute declared in an external entity."

 Example of the recommended use of prefixes and namespaces.

```
<OAI-PMH
        xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="
http://www.openarchives.org/OAI/2.0/
   http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
>
   <...>
   <metadata>
      <didl:DIDL
         xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
         xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
         xmlns:dcterms="http://purl.org/dc/terms/"
         xsi:schemaLocation="
urn:mpeg:mpeg21:2002:02-DIDL-NS
http://standards.iso.org/.../didl.xsd
urn:mpeg:mpeg21:2002:01-DII-NS
http://standards.iso.org/.../dii.xsd"
      >
         <...>
      </didl:DIDL>
   </metadata>
   </...>
<OAI-PMH>
```

Another argument is that for example a DIDL document is considered an autonomous entity that can exist outside a OAI record. When making a snippet from this DIDL document it should be valid according to a XML validator on its own. Thus does not need any namespace declaration texts that was left in the OAI-PMH xml.

According to the proclamation in the same document

(http://www.w3.org/TR/REC-xml-names/#ns-using), the DRIVER agreement will be that it is also possible to declare prefixes and namespaces in the ancestors of the document.

"The namespace prefix, unless it is xml or xmlns, MUST have been declared in a namespace declaration attribute in either the start-tag of the element where the prefix is used or in an ancestor element (i.e. an element in whose content the prefixed markup occurs)."

Example of the optional uses of prefixes and namespaces.

```
<OAI-PMH
        xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
        xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xsi:schemaLocation="
http://www.openarchives.org/OAI/2.0/
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd
urn:mpeg:mpeg21:2002:02-DIDL-NS
http://standards.iso.org/.../didl.xsd
urn:mpeg:mpeg21:2002:01-DII-NS http://standards.iso.org/.../dii.xsd
"
>
    <...>
    <metadata>
        <didl:DIDL>
          <...>
        </didl:DIDL>
    </metadata>
    </...>
<OAI-PMH>
```

XML validation

The XML that the repository provides will be validated automatically during the DRIVER repository registration process and the DRIVER harvesting process. A DRIVER

repository must provide a valid XML according to all XML schemas used (OAI-PMH, DIDL, oai-dc etc)

Validation can be tested using an XML validator (for example, from altova. www.altova.com) by saving the repository output as an xml document and opening it in the validator.

For a validator to validate an XML document, inside the document the xsi:schemaLocation(s) must be used.

```
For the <OAI-PMH> schema use:


<OAI-PMH

            xmlns="http://www.openarchives.org/OAI/2.0/"

            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"


            xsi:schemaLocation="
  http://www.openarchives.org/OAI/2.0/
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
>
```

For the <oai_dc:dc> schema use:

```
<oai_dc:dc

            xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"

            xmlns:dc="http://purl.org/dc/elements/1.1/"

            xsi:schemaLocation="
http://www.openarchives.org/OAI/2.0/oai_dc/
  http://www.openarchives.org/OAI/2.0/oai_dc.xsd
http://purl.org/dc/elements/1.1/
  http://dublincore.org/schemas/xmls/simpledc20021212.xsd"
>
```

For the <didl:DIDL> schema use:

```
<didl:DIDL
        xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
        xmlns:dip="urn:mpeg:mpeg21:2005:01-DIP-NS"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="
  urn:mpeg:mpeg21:2002:02-DIDL-NS
http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/did/didl.xsd

  urn:mpeg:mpeg21:2002:01-DII-NS
http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/dii/dii.xsd

  urn:mpeg:mpeg21:2005:01-DIP-NS
http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/dip/dip.xsd">
```

For other schemas use the same logic; keep the metadata independent of the OAI-PMH protocol.

Communication for Repository modification

Modification to baseURL, setSpec, metadataPrefix, or metadata schema's

When a DRIVER repository modifies either the baseURL, setSpec, metadataPrefix or metadata schemas which influence the DRIVER content cycle, then the concerning repository administrator **must** report this to the DRIVER community and the DRIVER harvester administrator in particular.

(http://helpdesk.driver.research-infrastructures.eu/)

status: final 2008-11-13

Use of Metadata OAI_DC

This chapter describes the way DRIVER envisions interoperability for scholarly communication. This means qualitative correct metadata of the records based on the use of standards.

Acknowledgements

This document is largely based on the recommendations for the use of Unqualified (simple) Dublin Core metadata as described in: USING SIMPLE DUBLIN CORE TO DESCRIBE EPRINTS, by Andy Powell, Michael Day and Peter Cliff, UKOLN, University of Bath, Version 1.2

See: http://www.intute.ac.uk/publications/eprints-uk/simpledc-guidelines.html

Additional information, descriptions, explanations, comments, usage instructions and best practices have been carefully provided with the aid of all DRIVER Guidelines contributors in order to create syntactic and semantic interoperability that will be appropriate for most European repositories.

Definitions

status: final 2008-11-13

"An institutional repository is a facility, consisting of hardware, software, data and procedures, that contains digital resources representing any type of scientific output…"

"digital resources = any bit stream, independent of content or format, which has been marked as scientific output by an approved person…"

Within this document we use the word "resource" to describe the instance of scientific output, and the word "object" to refer to the digital bit stream.

When "Requirement" is used we mean the following: "1 something required; a need. 2 something specified as compulsory.[13]"

When "Recommendation" is used we mean: "1 put forward with approval as being suitable for a purpose or role. 2 advise as a course of action. 3 make appealing or desirable.[13]"

Introductory remarks

Scope

The DRIVER Guidelines are written *primarily to facilitate the exchange of metadata* between DRIVER content providers and DRIVER services, in compliance with the DCMI definitions for Unqualified (simple) Dublin Core as specified in the OAI-PMH specifications.[14] Basically these DRIVER Guidelines describe the **mapping** from an internal format to Unqualified (simple) DC to support harvesting. They are **not to be used as cataloguing instructions.**

---

[13] Compact Oxford Dictionary of Current English third edition

[14] OAI-PMH specifications "For purposes of interoperability, repositories must disseminate Dublin Core, without any qualification." http://www.openarchives.org/OAI/openarchivesprotocol.html#MetadataNamespaces

In these DRIVER Guidelines Repository Managers have to accept the fact that not everything can be expressed with Unqualified DC, these guidelines therefore concentrate on the most important information in the perspective of the end-user who is not a librarian.

Minimal requirements

- Metadata are structured as Unqualified Dublin Core (ISO 15836:2003)
- Individual elements of DC are to be used according to the guidelines as presented in this appendix
- The use of Unicode is mandatory
- The values (i.e. actual content) of the DC-elements given below must not contain any HTML (or XML) markup. They may contain LaTeX commands, but there is no mechanism for explicitly indicating that LaTeX is being used.

Recommendations

- Represent Metadata in a higher granular structure such as Qualified Dublin Core or MODS. (Future work, additions to the DRIVER Guidelines)
- The DRIVER metadata guidelines only refer to metadata as exchange format. They do not hard code the recommendations made in the DRIVER Guidelines nor use a mapping between the locally implemented high granular metadata structures and the DRIVER recommendations.
- Recommended language for descriptive information is English, in order for the end-user to reach knowledgable documents that are normally "locked in" an national context.

Editions & difference in intellectual content

Only one metadata record should be used for different manifestations of a digital object (for example a postscript and a pdf version), unless the intellectual content is different. Common practice is to create a new metadata record when the intellectual content is different. This happens for instance when a new "edition", with modifications in the intellectual content, is created. In that case the recommended

best practice is to use the relation element to link the more recent version to the older one.

Classification schemas & Review policies

In some cases, **additional information** on local review policies, the use of metadata elements dc:subject and dc:type on local classification schemas or controlled keyword vocabulary, may be useful for the harvesting party and service provider. A content provider typically releases this type of information via the 'Identify request' on IR level; not on the metadata level. See for instance: **3. Guidelines for Optional Containers** at: http://www.openarchives.org/OAI/2.0/guidelines.htm and: http://arXiv.org/oai2?verb=Identify for best practices. On dc-element level this can be done by adding an URI to a term. For classification schemes that do not already have a namespace adding a sub-namespace to the info-uri namespace might be helpful. (see www.info-uri.info)

Dumbing down & Qualifiers

Some words on the use of **refinements (qualifiers)**: When mapping to Unqualified DC the content provider has to make choices when the internal format is "richer" than unqualified DC. This means that during the mapping process all refinements are simply dropped (the DCMI dumbing down principle). The effect of the dumbing down principle is that the simple form of the element, i.e. without the refinement, is the default one. E.g. when the internal format distinguishes between main title and Sub-title this would show as follows in DC:

Internal format
`245 $aMain title$sSub-title`
Qualified DC
`<dc:title>Main title</dc:title>`
`<dcterms:alternative>Sub- title</dcterms:alternative>`
Unqualified DC
`<dc:title>Main title:Sub-title </dc:title>`

Default dc-elements interpretations

However, within DRIVER the following values are selected as the default values for oai_dc

```
dc:description ->    default "abstract"
dc:date        ->    default "published"
dc:audience    ->    default "education level"
```

Within DRIVER this means that the date element always pertains to the date published etc. It is advised that all content providers supply this information to external harvesters as information about their repository (in the OAI-PMH Identify response).

Table 1: example of notifying the service provider on the default interpretation of the dc-element fields.

```
<OAI-PMH>
  <Identify>
    <description>
      <eprints>
        <metadataPolicy>
          oai_dc:dc:description(default "abstract");
          oai_dc:dc:date(default "published");
          oai_dc:dc:audience(default "education level");
        </metadataPolicy>
      </eprints>
    </description>
 </Identify>

</OAI-PMH>
```

The Elements: short description

                                               status: final 2008-11-13

Within DRIVER the *use of elements* is either:

- **mandatory** *(M)* = the element must always be present in the metadata record. An empty element is not allowed.
- **mandatory when applicable** *(MA)* = when the element can be obtained it must be present in the metadata record

- **recommended** *(R)*= the use of the element is recommended
- **optional** *(O)*= it is not important whether the element is used or not

The recommended status is made primarily to encourage users to input certain elements when creating a metadata record to enhance services.

Unqualified DC: oai_dc

Basic element	*Status*	*Encoding schemes*
Title	M	None, free text
Creator	M	APA bibliographic writing style as in a reference list. Syntax: surname, initials (first name) [http://en.wikipedia.org/wiki/Apa_style#Reference_list]
Subject	MA	Choice of keywords and classifications can be free text (preferably in English) and defined by an URI scheme (preferably info:eu-repo/classification).
Description	MA	None, free text. Recommended practice is to include an abstract in English. "Abstract" is the default interpretation to the value for dc:description
Publisher	R	None
Contributor	O	APA bibliographic writing style as in a reference list. Syntax: surname, initials (first name) [http://en.wikipedia.org/wiki/Apa_style#Reference_list]
Date	M	Date	ISO 8601 W3C-DTF - "Published" is the default interpretation to the value for dc:date
Type	M	Publication type and Version type can be free text (preferably

		in English) and defined by an URI scheme (preferably info:eu-repo/semantics).
Format	R	IANA registered list of Internet Media Types (MIME types) [http://www.iana.org/assignments/media-types/]
Identifier	M	URI scheme, linking to persistent identifier (URN, handle, DOI), full text document or human start page.
Source	O	Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata [http://dublincore.org/documents/dc-citation-guidelines/] as in dcterms:bibliographicCitation
Language	R	ISO 639-3
Relation	O	None
Coverage	O	"Period" is the default interpretation to the value for dc:coverage. Encoding: DCMI Period [http://dublincore.org/documents/2000/07/28/dcmi-period/] For more ncoding schemas see Chapter 5 Use of vocabularies and semantics.
Rights	R	None
Audience	O	None. "Eduction level" is the default value for dc:audience.

If no default interpretations are mentioned in the oai_dc elements in the table above, please describe the specific use of the oai_dc elements in the Identify section of your IR. See for instance: 3. Guidelines for Optional Containers at: http://www.openarchives.org/OAI/2.0/guidelines.htm and: http://arXiv.org/oai2?verb=Identify

The Elements: full description

Below full descriptions of the elements are provided.

DCMI definitions come from the DCMI guidelines document "Using Dublin Core - The Elements" see http://dublincore.org/documents/usageguide/elements.shtml .

Title

Element name	Title
DCMI definition	A name given to the resource. Typically, a Title will be a name by which the resource is formally known.
Usage	Mandatory
Usage instruction	Preserve the original wording, order and spelling of the resource title. Only capitalize proper nouns. Punctuation need not reflect the usage of the original. Subtitles should be separated from the title by a colon. . This instruction would result in Title:Subtitle (i.e. no space). If necessary, repeat this character for multiple titles.
Do not confuse with	(n.a.)

Examples	`<dc:title>Main title:Sub-title </dc:title>`
	`<dc:title>Dewey Classificatie in Archief systemen:Dewey Classification in Archival systems</dc:title>`
	`<dc:title>Preliminary studies for the "Philosophical Investigations", generally known as the blue and brown books </dc:title>`

Creator

Element name	Creator
DCMI definition	An entity primarily responsible for making the content of the resource. Typically, the name of a Creator should be used to indicate the entity.
Usage	Mandatory

     status: final 2008-11-13

Usage instruction	Examples of a Creator include a person, an organization, or a service. If necessary, repeat this element for multiple authors.
	Use *inverted name,* so the syntax will be the following: "surname", "initials" ("first name") "prefix" For example Jan Hubert de Smit becomes `<dc:creator>    Smit,    J.H.(John)    de</dc:creator>`
	Within the scope of Unqualified DC it is recommended to use a standardised writing style for names, use the writing style used by the publisher when this is available. When that is not available use the encoding of the APA bibliographic writing style as in a reference list when applicable. (outside the scope of Unqualified DC more precise and granular formatting methods are available.)
	When initials and first name are both available use this formatting: `<dc:creator>    Janssen,    J.    (John)</dc:creator>`
	Generational suffixes (Jr., Sr., etc.) should follow the surname. When in doubt, give the name as it appears, and do not invert. Omit titles (like "dr", "ir" etc.) For example: "Dr. John H. de Smit Jr." becomes `<dc:creator> Smit Jr., J.H. (John) de </dc:creator>`
	In the case of an organization name which clearly includes an organizational hierarchy, list the parts of the hierarchy from largest to smallest, separated by full stops. For example: `<dc:creator> Utrecht University. Department of Computer Sciences </dc:creator>` If it is not clear whether there is a hierarchy present, or unclear which is the larger or smaller portion of the body, give the name as it appears in the resource. Only encode organisations in this element to indicate corporate

	authorship, not to indicate the affiliation of an individual. The inclusion of personal and corporate name headings from authority lists constructed according to local or national thesaurus files is optional. It is recommended to encode thesauri with an URI, for service providers to recognise the thesaurus schema. For example: `<dc:creator> urn:NationalOrgThesaurus:nl/234 </dc:creator>` In cases of lesser responsibility, other than authorship, use dc:contributor. If the nature of the responsibility is ambiguous, recommended best practice is to use dc:publisher for organizations, and dc:creator for individuals.
Do not confuse with	• Contributor (see also User instruction above). • Publisher. The DC element 'creator' describes the name(s) of the creator(s) of the resource, as mentioned in the resource, whereas the DC element 'contributor' describes the scientist(s) that has/have made contributions to the given scientific output, not as a primary creator or (commercial) publisher.

Examples	`<dc:creator>Evans, R.J.</dc:creator>`
	 `<dc:creator>Walker Jnr., John</dc:creator>`
	 `<dc:creator>International Human Genome Sequencing Consortium</dc:creator>`
	 `<dc:creator>Loughborough University. Department of Computer Science</dc:creator>`

Subject

Element name	Subject
DCMI definition	The topic of the resource. Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe the intellectual content of the resource.

Usage	Mandatory when applicable
Usage instruction	In the DC subject element two kinds of values are possible: encode either a keyword or a classification. When both are available use separate occurrences of this element.
	Use the first occurrence of the DC element 'subject' for a human readable keyword.
	In general, choose the most significant and unique words for keywords, avoiding those too general to describe a particular resource. If the subject of the resource is a person or an organization, use the same form of the name as you would if the person or organization were an author, but do not repeat the name in the dc:creator element.
	For keywords/keyphrases that are not controlled by a vocabulary or thesaurus either encode multiple terms with a semi-colon separating each keyword/keyphrase; or repeat the element for each term. There are no requirements regarding the capitalization of keywords though internal (within archive) consistency is recommended.
	Where terms are taken from a standard classification schema: encode each term in a separate element. Encode the complete subject descriptor according to the relevant scheme. Use the capitalisation and punctuation used in the original scheme.
	It is recommended to use an URI when using classification schemes or controlled vocabularies especially when codified schemes are used DDC or UDC. Service providers can recognise encoding schemas more easy when the schema is "URI-fied" by an authority namespace. When the classification scheme is codified, use a human readable text of the code, preferably in English, directly below the codified element. For example:
	`<dc:subject>info:eu-repo/classification/ddc/641</dc:subject>` `<dc:subject>Anatomy</dc:subject>`

	If no specific classification scheme is used we recommend the Dewey Decimal Classification (DDC). The first 1000 terms is called the Dewey Decimal Classification Summary and can be downloaded at http://www.oclc.org/dewey/resources/summaries/ if one agrees with the following terms and conditions: http://www.oclc.org/research/researchworks/ddc/terms.htm
Do not confuse with	• Type DC element 'subject' describes the topic(s) of an resource; DC element 'type' describes the kind of academic output / Publication Type the resource is a representation of.
Schema	More on subject classification, see the section *Subject classification* on page 114 in chapter "Use of Vocabularies and Semantics".

Examples	`<dc:subject>polar oceanography; boundary current; mass transport; water masses; halocline; mesoscale eddies</dc:subject>`
	`<dc:subject>Germany--History--1933-1945</dc:subject>`
	`<dc:subject>info:eu-repo/classification/ddc/641</dc:subject>`
	`<dc:subject>Anatomy</dc:subject>`

Description

Element name	Description
DCMI definition	An account of the content of the resource. Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
Usage	Mandatory if applicable
Usage instruction	This element is used for a textual description of the content. When a resource consists of several separate physical object files, do not use dc:description to list the URL's of these files.

status: final 2008-11-13

	Default = abstract
Do not confuse with	(n.a.)

Examples	`<dc:description>`Foreword [by] Hazel Anderson; Introduction; The scientific heresy: transformation of a society; Consciousness as causal reality [etc]`</dc:description>`
	
	 `<dc:description>`A number of problems in quantum state and system identification are addressed. `</dc:description>`

Publisher

Element name	Publisher
DCMI definition	An entity responsible for making the resource available. Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Usage	Mandatory if applicable
Usage instruction	The (commercial or non-commercial) publisher of the resource; not the (sub)institution the author is affiliated with. Publisher is used only in the bibliographic / functional sense, not an organisational one. Use only the full name of the given (commercial) publisher, not the name of an organization or institute that is otherwise [in a broader sense] associated with the creator. With university publications place the name of the faculty and/or research group or research school after the name of the university. In the case of organizations where there is clearly a hierarchy present, list the parts of the hierarchy from largest to smallest, separated by full stops. If it is not clear whether there is a hierarchy present, or unclear which is the larger or smaller portion of the body, give the name as it appears in the eprint. The use of publisher names from authority lists constructed according

	to local or national thesaurus files is optional.
Do not confuse with	• Contributor • Creator In most cases the publisher and the creator are not the same.

Examples	`<dc:publisher>Loughborough University. Department of Computer Science</dc:publisher>`
	`<dc:publisher>University of Cambridge. Department of Earth Sciences</dc:publisher>`
	`<dc:publisher>University of Oxford. Museum of the History of Science</dc:publisher>`
	`<dc:publisher>University of Reading. Rural History Centre</dc:publisher>`
	`<dc:publisher>University of Exeter. Institute of Cornish Studies</dc:publisher>`
	`<dc:publisher>European Bioinformatics Institute</dc:publisher>`
	`<dc:publisher>John Wiley & Sons, Inc. (US)</dc:publisher>`

Contributor

Element name	Contributor
DCMI definition	An entity responsible for making contributions to the content of the resource. Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Usage	Optional
Usage instruction	Examples of contributors are: a supervisor, editor, technician or data collector. **Personal names should be listed as**: see instructions under Creator. A "promotor", i.e. a professor supervising a student's work for a doctor's degree - is considered a contributor of a dissertation in his or her role

	as promotor / examiner. In less-rich Unqualified DC it is difficult to express all roles in different contexts. In the PhD thesis as a document, the key figures are the author and the supervisor. In the overall PhD process other roles are involved, such as committee members and the Master of Ceremonies, but in Unqualified these roles have to be sacrificed. **In the case of organizations** : see instructions under Creator The inclusion of personal and corporate name headings from authority lists constructed according to local or national thesaurus files is optional.
Do not confuse with	• Creator • Publisher The DC element "contributor" describes the scientist(s) that has/have made contributions to the given scientific output, not as a primary creator or (commercial) publisher.)

Examples	`<dc:contributor>Sulston, John E.</dc:contributor>`
	`<dc:contributor>Evans, R. J.</dc:contributor>`
	`<dc:contributor>International   Human   Genome   Sequencing Consortium</dc:contributor>`
	`<dc:contributor>Loughborough   University.   Department   of Computer Science</dc:contributor>`

Date

Element name	Date
DCMI definition	A date associated with an event in the life cycle of the resource. Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and follows the YYYY-MM-DD format.
Usage	Mandatory
Usage	The date should be formatted according to the W3C encoding rules for

instruction	dates and times :
	Complete date: - YYYY-MM-DD (e.g. 1997-07-16) where: - YYYY [four-digit year] is mandatory - MM [two-digit month (01=January, etc.)] is optional - DD [two-digit day of month (01 through 31)] is optional **One date field – Date of Publication:** Often repository systems have more then one date fields that serve different purposes. Date of creation, publication, modified, promotion, etc. Unqualified DC is unable to express all these dates, and for the end-user perspective it is confusing to receive more dates from the service provider. The service provider should make a choice what date-field to pick. Preferrably in the end-users perspective the most logical and meaningful date will be the date of publication. To reduce the ambiguity of having a number of date fields without qualifiers, we recommend to reduce the number of fields and present the most meaningful date to the service provider. In most cases this is the date of the publication. In other cases this is the date of promotion of a PhD degree. **No date of publication available:** If no date of publication is available, use any other date available. It is better to use one date than no date at all. **Datestamp additions:** Additions like "Zulu time" should NOT be part of the metadata. **Fuzzy dates:** For fuzzy dates use a logical year that most represents that period, e.g.

	"1650" instead of "17th century"
	To express more about that temporal period, one can use the dc:coverage field. A temporal period can be expressed in a standard way when precisely defined (see Coverage) or when "fuzzy" or uncertain by free text expressions.
	A service provider is able to sort dates based on date standards like W3CDTF. Since there is no standard for fuzzy dates for terms like "Renaissance" or "17th Century", they will simply not appear on date-based query results.
Do not confuse with	-
scheme	ISO 8601 [W3CDTF] http://www.w3.org/QA/Tips/iso-date

Examples	`<dc:date>2000-12-25</dc:date>`
	`<dc:date>1978-02</dc:date>`
	`<dc:date>1650</dc:date>`

Type

Element name	Type
DCMI definition	The type of scientific output the resource is a manifestation of. In the DC element type the kind of dissemination, or the intellectual and/or content type of the resource is described. It is used to explain to the user what kind of resource he is looking at. Is it a book or an article. Was it written for internal or external use. Etc.
Usage	DC Element 'type' is used for three purposes: 1. Mandatory: Publication type (controlled): to indicate the type of publication based on the controlled DRIVER Publication-type vocabulary, 2. Optional: Publication type (free): to indicate the type of publication based on a local repository vocabulary 3. Recommended: Version (controlled): to indicate the status in the

	publication process.
Usage instruction	**1. Publication types (controlled):** The first occurrence of the DC Element 'type' is mandatory and should be used for the type indication of the scientific output based on the DRIVER-type vocabulary. Use exact string of characters as shown in the list below. The terms are explained in detail in the chapter about vocabularies and semantics. Info:eu-repo is a namespace where the DRIVER Publication types are registered. • info:eu-repo/semantics/article • info:eu-repo/semantics/bachelorThesis • info:eu-repo/semantics/masterThesis • info:eu-repo/semantics/doctoralThesis • info:eu-repo/semantics/book • info:eu-repo/semantics/bookPart • info:eu-repo/semantics/review • info:eu-repo/semantics/conferenceObject • info:eu-repo/semantics/lecture • info:eu-repo/semantics/workingPaper • info:eu-repo/semantics/preprint • info:eu-repo/semantics/report • info:eu-repo/semantics/annotation • info:eu-repo/semantics/contributionToPeriodical • info:eu-repo/semantics/patent • info:eu-repo/semantics/other **2. Publication types (free text):** The second occurrence of the DC Element 'type' is optional and should be used for the subtype indication of the scientific output. **3. Version (controlled):**

	The last occurrence of the DC Element 'type' is recommended and should be used for the version of the scientific output based on the DRIVER-version vocabulary. Use exact text as shown in the list below. For more information about the version model see http://www.lse.ac.uk/library/versions/ • info:eu-repo/semantics/draft • info:eu-repo/semantics/submittedVersion • info:eu-repo/semantics/acceptedVersion • info:eu-repo/semantics/publishedVersion • info:eu-repo/semantics/updatedVersion **Mapping & backwards-transformability:** For mappings of the DRIVER types from the DRIVER guidelines 1.0 see *DRIVER-TYPE Mappings.*
Do not confuse with	• Format DC element 'type' describes the kind of academic output the resource is a representation of. DC element 'format' describes the media type of this resource.
Schemes	Publication types: see the section *Publication type* on page 115 in chapter "Use of Vocabularies and Semantics". Version vocabulary: See the section *Version* on page 120 in chapter "Use of Vocabularies and Semantics". Mappings: see the section *DRIVER-TYPE Mappings* on page 83 in chapter "Use of Best Practices for OAI_DC".

Examples	`<dc:type>info:eu-repo/semantics/article</dc:type>`
	`<dc:type>info:eu-repo/semantics/publishedVersion</dc:type>`
	
	`or`
	
	`<dc:type>info:eu-repo/semantics/other</dc:type>` *<!--1-->*

```
<dc:type>image</dc:type><!--2-->
<dc:type>info:eu-repo/semantics/updatedVersion</dc:type> <!-
-3-->
```

Format

Element name	Format
DCMI definition	The physical or digital manifestation of the resource. Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats).
Usage	Recommended
Usage instruction	Based on best practice, the IANA registered list of Internet Media Types (MIME types) is used to select a term from. For the full list see the scheme location below. Below will follow an example list of IANA MIME types:

	Type	Subtype
	text	- plain - richtext - enriched - tab-separated-values - html - sgml - xml
	application	- octet-stream - postscript - rtf - applefile

		• mac-binhex40
		• wordperfect5.1
		• pdf
		• vnd.oasis.opendocument.text
		• zip
		• macwriteii
		• msword
		• sgml
		• ms-excel
		• ms-powerpoint
		• ms-project
		• ms-works
		• xhtml+xml
		• xml
	image	• jpeg
		• gif
		• tiff
		• png
		• jpeg2000
		• sid
	audio	• wav
		• mp3
		• quicktime
	video	• mpeg1
		• mpeg2
		• mpeg3
		• avi
	If one specific resource (an instance of scientific output) has more than one physical formats (e.g. postscript and pdf) stored as different object	

	files, all formats are mentioned in the DC element 'format', for example: • <dc:format>application/pdf</dc:format> • <dc:format>application/postscript</dc:format> • <dc:format>application/vnd.oasis.opendocument.text</dc:format>
Do not confuse with with	• Type • Identifier DC element 'format' describes the media type of this resource. DC element 'type' describes the kind of academic output the resource is a representation of. Dc:identifier is used to represent manifestations of digital resources.
Scheme	the IANA registered list of Internet Media Types (MIME types) - http://www.iana.org/assignments/media-types/

Examples	`<dc:format>video/quicktime</dc:format>`
	 `<dc:format>application/pdf</dc:format>`
	 `<dc:format>application/xml</dc:format>`
	 `<dc:format>application/xhtml+xml</dc:format>`
	 `<dc:format>application/html</dc:format>`
	 `<dc:format>application/vnd.oasis.opendocument.text</dc:format>`

Identifier

Element name	Identifier
DCMI definition	An unambiguous reference to the resource within a given context.
Usage	Mandatory
Usage instruction	Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Example formal identification systems include the Uniform Resource Identifier

	(URI) (including the Uniform Resource Locator (URL), the Digital Object Identifier (DOI) and the <u>URN:NBN</u> The ideal use of this element is to use a direct link or a link to a jump-off page (persistent URL) from dc:identifier in the metadata record to the digital resource or a jump-off page. Smart practice: # use stable URL's ▪ provide every identifier one can find about the publication. o (URL, DOI, <u>URN:NBN</u>, ISBN, ISSN, etc.) ▪ place the "most appropriate" identifier in the form of a URL at the top of the list of Identifiers. In almost all cases this is the one that will be used by a service provider to let an end-user refer to. This can be a link to a jump-off page or a direct link to the file. Also this can be a direct URL, or a redirection URL, like PURL, HANDLE or other international resolution mechanisms.
Do not confuse with	• dc:relation (Use dc:relation to refer from one version of the resource to another.) • dc:source (Use dc:source for bibliographic citation of the originating resource.)
Examples	In this example the identifiers are sorted where the URL's are given first. The first URL will be considered as "most appropriate" and will be used in e.g. DRIVER to let an end-user redirect to. In this case the handle redirects to the jump-off page. A Jump-off page is a good way to refer to. The end-user has the opportunity to see more information about the object(s) he has found, see the context and enjoy the other services a local repository has to offer. `<oai_dc:dc>`

```
  ...
  <dc:identifier>http://hdl.handle.net/1234/5628
</dc:identifier>
  <dc:identifier>http://arno.unimaas.nl/show.cgi?fid=5628
</dc:identifier>
  <dc:identifier>http://n2t.info/urn:nbn:nl:ui:14-
123456789</dc:identifier>
  <dc:identifier>urn:nbn:nl:ui:13-123456789</dc:identifier>
  <dc:identifier>urn:isbn:123456789</dc:identifier>
  <dc:identifier>info:doi:10-123456789</dc:identifier>
  ...

</oai_dc:dc>
```

Source

Element name	Source
DCMI definition	A reference to a resource from which the present resource is derived.
Usage	Optional
Usage instruction	The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system. Best practice: Use only when the described resource is the result of digitization of non-digital originals. Otherwise, use Relation. Optionally metadata about the current location and call number of the digitized publication can be added. Use: Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata ([http://dublincore.org/documents/dc-citation-guidelines/]).

Do not confuse with	• dc:relation
	• dc:identifier
Examples	`<dc:source>Ecology    Letters    (1461023X)    vol.4 (2001)</dc:source>`
	`<dc:source>ISSN: 0928-0987</dc:source>`

Language

Element name	Language
DCMI definition	A language of the intellectual content of the resource.
Usage	Recommended
Usage instruction	A specific resource (an instance of scientific output) is either written in one human language or more. In these cases all used languages are used in the DC element 'language'. If a specific resource (an instance of scientific output) is written in one human language and is translated into other human languages, each translation does have its own record..
	Recommended: ISO 639-x, where x can be 1,2 or 3.
	Best Practice: we use **ISO 639-3** and by doing so we follow:
	[http://www.sil.org/ISO639-3/codes.asp]
	If necessary, repeat this element to indicate multiple languages.
	If ISO 639-2 and 639-1 are sufficient for the contents of a repository they can be used alternatively. Since there is a unique mapping this can be done during an aggregation process.
Do not	• Country codes ISO 3166-1

confuse with	http://www.iso.org/iso/country_codes/iso_3166_code_lists/ english_country_names_and_code_elements.htm
Scheme	ISO 639-3 http://www.sil.org/ISO639-3/codes.asp
Examples	`<dc:language>eng</dc:language>` `<dc:language>deu</dc:language>` `<dc:language>nld</dc:language>` `<dc:language>nld/dut</dc:language>` `<dc:language>dut</dc:language>` `<dc:language>nl</dc:language>`

Relation

Element name	Relation
DCMI definition	The reference to a related resource.
Usage	Optional
Usage instruction	Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system. The DC element 'relation' can be used to indicate different kinds of relations between several metadata records. If relations between metadata records are made visible by using metadata the following holds for the distinction between versions (author version and publisher version, preprint, postprint, etc.): • A metadata record is self-contained • Different manifestations of one and the same resource (an instance of scientific output that can be described with exactly the same bibliographic metadata, except for the DC element 'format') are linked to one single metadata record using dc:relation. Changes in the metadata other than the DC element 'format' leads to

	creating a new metadata record of this new instance of scientific output, which meets all requirements formulated in this document and has a value in the DC element 'relation'.
Do not confuse with	dc:identifier and dc:source.

Examples	```<dc:relation>http://hdl.handle.net/10 </dc:relation>```
	```The value of dc:relation is the identifer of the other```
	```document.```
	
	```Linking two documents:```
	```---Document A:---```
	```<dc:type>info:eu-repo/semantics/submittedVersion</dc:type>```
	```<dc:identifier> http://hdl.handle.net/10</dc:identifier>```
	```<dc:relation>http://hdl.handle.net/20</dc:relation>```
	
	```---Document B:---```
	```<dc:type>info:eu-repo/semantics/acceptedVersion</dc:type>```
	```<dc:identifier> http://hdl.handle.net/20</dc:identifier>```
	```<dc:relation>http://hdl.handle.net/10</dc:relation>```

Coverage

Element name	Coverage
DCMI definition	The extent or scope of the content of the resource. Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity).
Usage	Optional
Usage instruction	Recommended best practice is to select the value from a controlled vocabulary (for example, the Getty Thesaurus of Geographic Names or TGN) and that, where appropriate, named places or time periods be used

	in preference to numeric identifiers as, for example, sets of co- ordinates or date ranges. If necessary, repeat this element to encode multiple locations or periods.
Do not confuse with	•
Scheme	• ISO 3166 [http://www.iso.ch/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/index.html] • Box [http://dublincore.org/documents/dcmi-box/] • TGN [http://www.getty.edu/research/tools/vocabulary/tgn/] • DCMI Period [http://dublincore.org/documents/2000/07/28/dcmi-period/]

Examples	`Example Spatial: ISO 3166`
	`<dc:coverage>NL</dc:coverage>`
	
	`Example Spatial: BOX`
	`<dc:coverage> name=Western Australia; northlimit=-13.5;`
	`southlimit=-35.5; westlimit=112.5;`
	`eastlimit=129</dc:coverage>`
	
	Note ad BOX: The syntax used here is provisional, and is currently under review as part of the DCMI work on recommending coordinated syntax recommendations for HTML, XML, and RDF. These recommendations and minor editorial changes in this document can be expected to take place in the near future. Point http://dublincore.org/documents/dcmi-point/

Rights

Element name	Rights
DCMI definition	Information about rights held in and over the resource.
Usage	Recommended

Usage instruction	Typically, a Rights element will contain a rights management statement for the access or use of the object, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. It is preferred to refer to a rights service where the reuse rights are made clear to the end-user by using a URL. For example the Creative Commons organisation has created URIs for their different Licences in the different Jurisdictions. This can be applied to create machine readable usage licenses.
Do not confuse with	•
Examples	``` <dc:rights>(c) University of Bath, 2003</dc:rights> <dc:rights>(c) Andrew Smith, 2003</dc:rights> ``` Using Creative Commons right services, makes the usage rights much more clear to the end user. More information see Use of Intellectual Property Rights. In this case Andrew Smith referring to http://creativecommons.org/licenses/by-sa/2.0/uk/ ``` <!-- example 1 --> <dc:rights>http://creativecommons.org/licenses/by- sa/2.0/uk/</dc:rights> ``` The URL provides the location where the license can be read. With creative common licenses the type of license can be recognized in the URL name itself. A pro for having the license point to an URL in this way, is that this is machine readable. ``` <!-- example 2 --> <dc:rights>cc-by-sa, Andrew Smith</dc:rights> ``` The string cc-by-sa provides the licence type in a rough sense. The name is the person or party where the rights apply to. ``` <!-- example 3 --> ```

	`<dc:rights>cc-by-sa, info:eu-repo/dai/nl/344568</dc:rights>`
	
	or
	
	`<dc:rights>cc-by-nc-sa, urn:isni:234562-2</dc:rights>`
	
	Also a Digital Author Identifier (DAI) or International Standard Name Identifier (ISNI) can be used to globally uniquely identify persons and organisations and relate thse names with the approprate rights.

Audience

Element name	Audience
DCMI definition	A class of entity for whom the resource is intended or useful.
Usage	Optional
Usage instruction	A class of entity may be determined by the creator or the publisher or by a third party. On the U.S. Department of Education, Metadata Reference site, an example is given of audiences: http://www.ed.gov/admin/reference/index.jsp : • Administrators • Community Groups • Counsellors • Federal Funds Recipients and Applicants • Librarians • News Media • Other • Parents and Families • Policymakers • Researchers • School Support Staff • Student Financial Aid Providers • Students

	• Teachers
Do not confuse with	•
Examples	`<dc:audience>Researchers</dc:audience>` `<dc:audience>Students</dc:audience>`

Use of Best Practices for OAI_DC

This chapter deals with common problems that repository administrators come across when installing a repository. These practices are not mandatory, but form the best possible solution to common problems. These solutions come from best practices from other repository administrators who already have dealt with these kinds of problems before. The main focus here is interoperability and the ease of implementation in terms of the scholarly communication life cycle.

DRIVER-TYPE Mappings

Mapping of other Publication type lists with the one made available in the section *Publication type* on page 115 in chapter "Use of Vocabularies and Semantics". In that section one can find details definitions of the terms used in that vocabulary in order to make custom mappings.

DRIVER v1.1 types to DRIVER v2.0 types

Below is the mapping between the document types used in the DRIVER Guidelines version 1.1 compared with the ones in version 2.0 .

DRIVER types v1.0	becomes / maps to	DRIVER types v2.0
Article	>>	article
Bachelor thesis	>>	bachelorThesis
Master thesis	>>	masterThesis
Doctoral thesis	>>	doctoralThesis
Book	>>	book
Part of book or chapter of book	>>	bookPart
not available in DRIVER types v1.1!	>>	review
Conference lecture	>>	conferenceObject
Conference report	>>	conferenceObject
Lecture	>>	lecture
Research paper	>>	preprint or workingPaper
External research report	>>	report
Internal report	>>	report
not available in DRIVER types v1.1!	>>	annotation
Contribution for newspaper or weekly magazine	>>	contributionToPeriodical
Newsletter	>>	contributionToPeriodical
not available in DRIVER types v1.1!	>>	patent
not available in DRIVER types v1.1!	>>	other

E-Print type vocabulary to DRIVER v2.0 types

Below is the mapping between the document types used in the e-print vocabulary compared with the ones in version 2.0 .

How to express an article with 2 object files, the one 'accepted', the second one being the 'published' version?

e-print type vocabulary	becomes / map to	DRIVER types v2.0	DRIVER versioning
JournalArticle	>>	article	accepted / published / updated
JournalItem	>>	article	accepted / published / updated
SubmittedJournalArticle	>>	preprint or workingPaper	submitted
Thesis (broader)	>>	bachelorThesis	
Thesis (broader)	>>	masterThesis	
Thesis (broader)	>>	doctoralThesis	
Book	>>	book	
BookItem	>>	bookPart	
BookReview	>>	review	
ConferencePaper	>>	conferenceObject	
ConferenceItem	>>	conferenceObject	
ConferencePoster	>>	conferenceObject	
not available in e-print type vocabulary	>>	lecture	
WorkingPaper	>>	workingPaper	
ScholarlyText	>>	other ??? (to generic)	
Report (broader)	>>	report	
not available in e-print type vocabulary	>>	annotation	
NewsItem	>>	contributionToPeriodical	
Patent	>>	patent	
not available in e-print type vocabulary	>>	other	

More information about the e-print type vocabulary can be found here http://purl.org/eprint/type/

DRIVER-VERSION Mappings

Below are the mappings of the DRIVER versioning scheme compared to other versioning schemes In the library and repository world. More about DRIVER versions in the section *Version* on page 120 in chapter "Use of Vocabularies and Semantics".

Eprints Version types to DRIVER Guidelines v2.0 VERSION types

Below is the mapping between the document types used in the Eprints Version types compared with the ones in the DRIVER guidelines version 2.0 .

e-print versions	becomes / maps to	DRIVER GL v2.0 VERSIONS
non-peer reviewed	>>	draft
non-peer reviewed	>>	submittedVersion
peer reviewed	>>	acceptedVersion
peer reviewed	>>	publishedVersion
peer reviewed	>>	updatedVersion

Common version terms to DRIVER Guidelines v2.0 VERSION types

Below is the mapping between the document types used in common scientific termscompared with the ones in the DRIVER guidelines version 2.0 .

traditional versions	becomes / map to	DRIVER GL v2.0 VERSIONS
Working paper	>>	draft
Pre print	>>	submittedVersion
Post print	>>	acceptedVersion
Journal article	>>	publishedVersion
Reprint	>>	updatedVersion

Journal Article Versions (JAV) Technical Working Group versions to DRIVER Guidelines v2.0 VERSION types

These recommendations provide a simple, practical way of describing the versions of scholarly journal articles that typically appear online before, during, and after formal journal publication. The Recommended Terms and Definitions for Journal Article Versions define journal articles at seven stages.

JAV	becomes / map to	DRIVER GL v2.0 VERSIONS
Author's Original	>>	draft
Submitted Manuscript Under Review	>>	submittedVersion
Accepted Manuscript	>>	acceptedVersion
Proof	>>	acceptedVersion
Version of Record	>>	publishedVersion
Corrected Version of Record	>>	publishedVersion
Enhanced Version of Record	>>	updatedVersion

More information about JAV: http://www.niso.org/publications/rp/RP-8-2008.pdf

Use of OAI_DC with Theses

This recommendation is based on the study report "A PORTAL FOR DOCTORAL E-THESES IN EUROPE; Lessons Learned from a Demonstrator Project"

This study is aiming at generic scholarly communication services harvesting OAI_DC. For context specific e-theses services we recommend to use other metadata schemas besides OAI_DC where all aspects concerning e-theses are offered.

Common practice when using OAI_DC dc:type with the content **"info:eu-repo/semantics/doctoralThesis"**, is that very close attention must be paid to following:

- The **dc:date field** must always contain the **date of publication** (not the date of the defense. The defense date is meaningful in the specific context of e-theses services)

- Use only **one date field**. More date fields will be considered ambiguous, because DC has no room to specify other types of dates.

- The **dc:contributor** field always must contain the name of the **supervisor**. (Using contributor fields with names of other roles will be considered ambiguous. DC has no room to specify other contributor roles.)

- The rest of the fields should follow the DRIVER Guidelines exactly. Please pay attention to the dc:language field that it is preferably encoded in iso639-3. Also note that the dc:identifier is the only field that contains a URL that points to a full text thesis document or intermediate page with open access to the full text thesis document. The dc:date field must be ISO8601 (YYYY-MM-DD). And the dc:creator and dc:contributor fields are formatted in "lastname, firstname" style.

Example

In this section an example is given for an electronic thesis. In this case it is a "Habilitation" a German type of thesis that is used when a person becomes a Professor. This is an academic work that is even rated higher than a PhD / Doctoral thesis in Germany. In the DRIVER Guidelines we only support the terms used in the Bologna convention, so the repository manager can use the rule "everything equal and higher then a Doctoral thesis will be put in the category doctoralThesis". In the DRIVER Guidelines it is allowed to put the extra information "habilitation" in order to keep the local levels.

For more information on the Diplom level terms see http://en.wikipedia.org/wiki/Diplom

The XML that is used could look like the following (the comments between *<!--* and *-->* should not be in the out XML, but serve as a reading aid.):

```
<oai_dc:dc >
  <dc:title>Mixing Oil and Water : </dc:title>

  <dc:creator>Stage, Jesper</dc:creator> <!-- The Author -->

  <dc:date>2003-12-02</dc:date> <!-- The Published Date, one data field
-->

  <dc:contributor>Crane, Walter</dc:contributor> <!-- The Supervisor  -
->

  <dc:type>info:eu-repo/semantics/doctoralThesis</dc:type>     <!--
DRIVER type 2.0 for Doctoral thesis, used for interoperability -->

  <dc:type>habilitation</dc:type> <!-- Local specific term. In Germany
Habilitation is the thesis a Professor has to write -->

  <dc:type>info:eu-repo/semantics/publishedVersion</dc:type>   <!--
Optional, the status of the work -->

  <dc:identifier>http://some.url.to/the_jump-off_page.html
</dc:identifier>
  ...
</oai_dc:dc>
```

DC:SOURCE and Citation information

For publications use the DC:SOURCE field for inserting information a person can use to appropriately make a citation of the record he/she has found. Preferably use the APA style of writing references.

For example

```
<dc:source>Ecology Letters (1461023X), vol.4 (2001)</dc:source>
```

DC:RELATION and Linking related objects

The DC:RELATION field can typically be used for describing relations to other expressions, or versions of the document.

For example the Published version of an article and the author version of an article. These can be referred to each other by using the "most appropriate" identifier that is actionable (URL). For example

This record with ID 1111, is a paper that has been submitted for peer reviewing. This paper has a relation with the peer reviewed article with ID 2222.

```
<oai_dc:dc >
  <de:identifier>http://hdl.handle.net/1234/1111</dc:identifier>
  <dc:type>info:eu-repo/semantics/paper</dc:type>
  <dc:type>info:eu-repo/semantics/submittedVersion</dc:type>
  <dc:relation>http://hdl.handle.net/1234/2222</dc:relation>
</oai_dc:dc>
```

The metadata record below shows the record of the article with ID 2222. This article has a relation with the submitted paper.

```
<oai_dc:dc >
  <de:identifier>http://hdl.handle.net/1234/2222</dc:identifier>
  <dc:type>info:eu-repo/semantics/article</dc:type>
  <dc:type>info:eu-repo/semantics/publishedVersion</dc:type>
  <dc:relation>http://hdl.handle.net/1234/1111</dc:relation>
</oai_dc:dc>
```

Digital Repository Infrastructure Vision for European Research

Use of MPEG-21 DIDL (xml-container) - Compound object wrapping

Introduction and Goal

This document is an addition to the existing DIDL specification document for repositories which is being used by the Dutch Universities, Koninklijke Bibliotheek, National Library of The Netherlands, and NARCIS. The goal of this document is to make the use of DIDL unambiguously clear by describing:

1. the nature of the different parts "metadata", "objects" and "jump-off-page"
2. What the identification is
3. What the modification-date is

When used correctly, this specification will create a valid XML MPEG-21 DIDL record for use with OAI-PMH responses. This specification of the DIDL document for repositories is based on decisions that were proposed early in the development of this XML format to use MPEG-21 DIDL. The proposition was a rough sketch of a wrapper format that has room for metadata, object and jump-off-page resources. This specification is a more precise workout.

status: final 2008-11-13

Background information

The DIDL XML container was originally developed within the DARE program of SURF as a first implementation of MPEG-21 DIDL. The rationale behind this development was:

- A solution for resource harvesting via OAI-PMH for transport of the digital resources (PDF's etc) from the local repository to the National Library for ingest of the resources into the E-Depot system for long term preservation
- A solution for resource harvesting via OAI-PMH for transport of the digital resources (PDF's etc) from the local repository system to a service provider (e.g. a search portal that indexes the full text of documents)
- A (partial) solution for representing complex documents; at first focused on theses that consist of multiple digital resource files
- A solution for the confusing use of `dc:identifier` in case of a link to a so called jump-off page (JOP). Many repositories place a link to a jump-off page in dc:identifier instead of a direct link to the digital resource file.

The DIDL XML container has been in use within DARE since the summer of 2006. One of the results is that the contents of all Dutch repositories are now part of the E-Depot of the Koninklijke Bibliotheek, National Library of The Netherlands.

OAI Response with a DIDL document

The DIDL document is part of an OAI-PMH response. The DIDL document will be returned within an OAI-record when using `didl` as value of the metadataPrefix verb. This enables the repository to generate this particular DIDL format that is described in the document below. Within the OAI XML structure, the DIDL resides within the metadata element. See below:

```
<OAI-PMH ...>
  ...
  <request ... metadataPrefix="didl_document">
  ...
  <record>
    <header>...</header>
    <metadata>
      <didl:DIDL
        xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
        xmlns:dip="urn:mpeg:mpeg21:2005:01-DIP-NS"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="
urn:mpeg:mpeg21:2002:02-DIDL-NS
http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/did/didl.xsd
urn:mpeg:mpeg21:2002:01-DII-NS
http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/dii/dii.xsd
urn:mpeg:mpeg21:2005:01-DIP-NS
http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/dip/dip.xsd">
          ...
        </didl:DIDL>
    </metadata>
    <about>...</about>
  </record>
  ...
</OAI-PMH>
```

Remarks:

1. Don't forget the DIDL tag in the OAI-PMH response
2. Make a declaration of the `didl` , `dii`, `dip` and `dcterms` namespaces here, in the DIDL tag. These namespaces are needed throughout the whole DIDL document. Do not create these namespaces in the <OAI-PMH> tag, because the

rationale of a DIDL document is that it can exist out of the context of OAI-PMH as an autonomous entity.

3. The `about` element is optional in OAI-PMH

DIDL as wrapper

The DIDL XML container, as defined in DRIVER, is a document with one top-level `Item` element. The `Item` contains several child `Item` elements. These child item elements appear in three different kind of types. Between the straight brackets the cardinality of the XML elements are shown:

```
<metadata>
  <didl:DIDL ...>
    <didl:Item>
      <didl:Item>...</didl:Item>
      <didl:Item>...</didl:Item>
      <didl:Item>...</didl:Item>
    </didl:Item>
  </didl:DIDL>
</metadata>
```

DIDL[1..1]
Item[1..1]
Item[1..∞] (of type 1 metadata)
Item[1..∞] (of type 2 objectFile)
Item[0..1] (of type 3 start page)
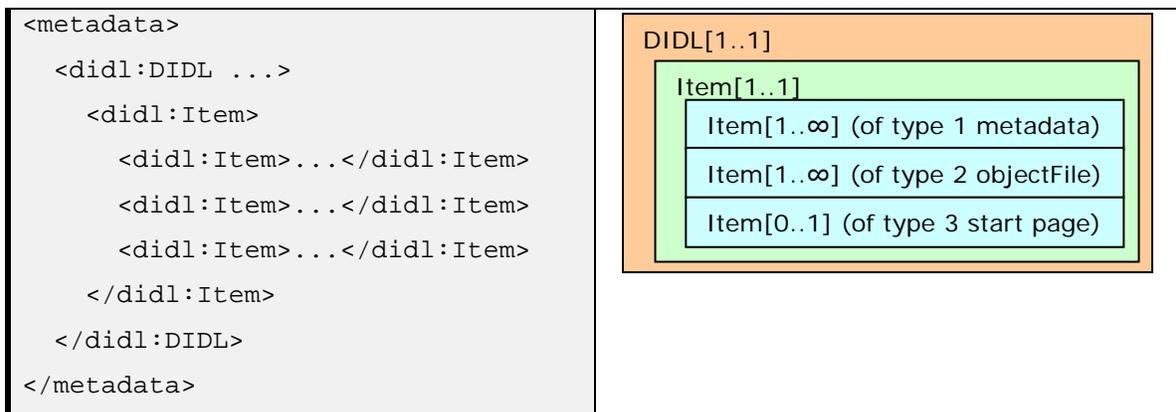
Root Element: DIDL document Identification attribute

The `DIDL` root element contains *one* `attribute`; namely `DIDLDocumentID`. This `attribute` provides information about the Identifier of the `DIDL` wrapper as an autonomous entity. This Identifier is NOT to identify the intellectual work, but to Identify the serialisation of the DIDL XML.

```
<didl:DIDL
  DIDLDocumentId="urn:nbn:nl:ui:10-15290" <!-- Identification -->
  ...
>
  ...
</didl:DIDL>
```

The `DIDLDocumentId` attribute contains the ID of the DIDL wrapper. This CAN be the same as the OAI-Identifier that is being used to get a record. The DIDL wrapper can be used as an autonomous entity out of the OAI-PMH context, therefore a DIDL is not the same 'thing' as an OAI record. There is a demand for Persistent Identifiers assigned to digital objects in the future (mandatory for the OAI-ORE project.). For libraries it is recommended to use urn:nbn:{country code}:{isil library code}[15]- {object id}. {object id} could be the database number. It is recommended to store this number in a separate field and not to auto generate from the database id because a database update in the future will change these numbers and the persistency could be lost.

Remarks

1. This DIDLDocumentId has in the first place a different Identifier than the OAI identifier for this record. The rationale behind this is that a DIDL document is an autonomous entity that can exist outside and separate of an OAI-record. However for easing the operational implementation, it is allowed to use the Identifier that is used for the OAI record when both, OAI record and DIDL document are inextricably bound together

Item Descriptor Elements (optional)

The `Item` elements can **OPTIONALLY** contain *two* or *three* `Descriptor` elements. One `Descriptor` element describes the modification date of the `Item` element. To compare similar harvested `Item` elements on modification date, an identifier must be added.

---

[15] ISO/NP 15511: International Standard Identifier for Libraries and Related Organizations (ISIL)
http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=52666

Example on level one:

```
<didl:DIDL ...>
  <didl:Item>
    <didl:Descriptor>...</didl:Descriptor> <!-- Identification -->
    <didl:Descriptor>...</didl:Descriptor> <!-- Modification date -->
    <didl:Item>...</didl:Item>
    <didl:Item>...</didl:Item>
    <didl:Item>...</didl:Item>
    ...
  </didl:Item>
</didl:DIDL>
```

Example on level two; Object type added:

```
      <didl:Item> <!-- Level 1 Root Item -->
        <didl:Item> <!-- Level 2 Child Item -->
            <didl:Descriptor>...</didl:Descriptor> <!--
Identification -->
            <didl:Descriptor>...</didl:Descriptor> <!-- Modification
date -->
            <didl:Descriptor>...</didl:Descriptor> <!-- Object type -
->
         ...
        </didl:Item>
          <didl:Item>...</didl:Item>
          <didl:Item>...</didl:Item>
          <didl:Item>...</didl:Item>
          ...
        </didl:Item>
      </didl:DIDL>
```

Descriptor Statement: Item 'Identifier'

The *first* Descriptor contains the ID of the Item elements. This is mostly used to uniquely identify the digital object (e.g. with a DOI). This ID is wrapped in a Statement with a DII Identifier element. For example:

```
<didl:Item>

  <didl:Item>

    <didl:Descriptor>

      <didl:Statement mimeType="application/xml">

        <dii:Identifier>urn:nbn:nl:ui:10-6748398729821</dii:Identifier>

      </didl:Statement>

    </didl:Descriptor>

    ...

  </didl:Item>

  ...

</didl:Item>
```

Remarks:

1. For child `Item` elements of the root `Item` element accounts that this `Identifier` is NOT equal to the used OAI identifier or DIDL identifier
2. The `Identifier` in the root `Item` element CAN be the same as the DIDL or OAI Identifier, but this is not recommended
3. The namespace for `dii` has had to be declared in the DIDL tag
4. The Identifier in the HAS TO BE described as an URI when applicable

Descriptor Statement: Item 'modified'

The *second* `Descriptor` contains a modification date. When something changes inside an `Item`, this modification date element has to be up-dated. This modification date is being specified by the `modified` element from the `dcterms` namespace:

```
<didl:Item>

  <didl:Item>

  ...

    <didl:Descriptor>

      <didl:Statement mimeType="application/xml">

        <dcterms:modified>2006-12-20T10:29:12Z</dcterms:modified>

      </didl:Statement>

    </didl:Descriptor>

    ...
```

status: final 2008-11-13

```
  </didl:Item>
  ...
</didl:Item>
```

Remarks:

1. Declare the `dcterms` namespace in the DIDL tag
2. The format of the date is Zulu-time; which means that it can be sorted as text
3. There can be only one `Statement` element in a `Descriptor` element, which means that `dii:identifier` and `dcterms:modified` reside in separate `Descriptor` elements

Descriptor Statement: Item 'ObjectType'

The *third* `descriptor` contains the object type. This Object type appears on the second level of `Item` elements. In other words; this applies only on child Item elements of the root `Item`. This object type is being specified by the `ObjectType` element from the MPEG-21 Digital Item Processing (DIP) namespace that specifies an architecture pertaining to the dissemination of Digital Item Documents (DIDs).

```
<didl:Item>
  <didl:Item>
  ...
    <didl:Descriptor>
      <didl:Statement mimeType="application/xml">
        <dip:ObjectType>info:eu-repo/semantics/descriptiveMetadata
</dip:ObjectType>
      </didl:Statement>
    </didl:Descriptor>
    ...
  </didl:Item>
  ...
</didl:Item>
```

status: final 2008-11-13

In the section *Compound Element: representation of the complex work* the representation of the complex work this ObjectType statement will be further eleborated upon.

Remarks:

1. Declare the `dip` namespace in the `DIDL` tag
2. The `ObjectType` in the Descriptor Statement HAS TO BE described as an URI
3. The processing architecture we use for dissemination will be for General European repositories. The URI used is placed at the `info` namespace as info:eu-repo. (http://info-uri.info/) Meanwhile it is used as an un-official standard within the driver community.

Compound Element: representation of the complex work

The top-Item element contains at least *two mandatory* `Item` element ObjectTypes. These Item-ObjectTypes are expressions of the root Item: one for the metadata and one for the digital object file, e.g. a PDF, as described by the metadata. *Optionally* there can be a third `Item` element ObjectType for a jump-off-page. The jump-off page is an html intermediate page that is used for human readable presentations when an Item has more than one digital object file. This situation typically occurs with theses that have separate object files (for example, when the thesis consists of a set of previously published articles). It also occurs when the content provider has a PDF, MS Word DOC and a HTML version of the same article.

```
<didl:DIDL ...>
  <didl:Item>
    <didl:Item>...</didl:Item> <!-- metadata -->
    <didl:Item>...</didl:Item> <!-- objects -->
    <didl:Item>...</didl:Item> <!-- jump-off-page -->
  </didl:Item>
</didl:DIDL>
```

The first `Item` contains the metadata as Unqualified Dublin Core (DC) (mandatory) which is normally used in the OAI_DC format according to the DRIVER metadata

guidelines that belongs to a Digital Item Processing architecture. The second `Item(s)` contain links to the digital objects, and the third `Item` contains a link to a jump-off page.

```
<didl:Item>
  <didl:Item> <!--one or many occurrences-->
    <didl:Descriptor>
      <didl:Statement mimeType="application/xml">
        <dip:ObjectType>info:eu-
repo/semantics/descriptiveMetadata</dip:ObjectType>
      </didl:Statement>
    </didl:Descriptor>
    ...
  </didl:Item>
  <didl:Item>     <!--one or many occurrences-->
    <didl:Descriptor>
      <didl:Statement mimeType="application/xml">
        <dip:ObjectType>info:eu-
repo/semantics/objectFile</dip:ObjectType>
      </didl:Statement>
    </didl:Descriptor>
    ...
  </didl:Item>
  <didl:Item> <!-- zero or one occurrences -->
    <didl:Descriptor>
      <didl:Statement mimeType="application/xml">
        <dip:ObjectType>
            info:eu-repo/semantics/humanStartPage</dip:ObjectType>
      </didl:Statement>
    </didl:Descriptor>
    ...
  </didl:Item>
</didl:Item>
```

The URI's will be processed case un-sensitive. It is recommended to use camelCase writing. It is VERY important to use the exact combinations of characters, otherwise

status: final 2008-11-13

automatic processing will not be possible. To make it very clear the following URI's are used:

- info:eu-repo/semantics/**descriptiveMetadata**
  (This Item occurs 1 or many times)
- info:eu-repo/semantics/**objectFile**
  (This Item occurs 1 or many times)
- info:eu-repo/semantics/**humanStartPage**
  (This Item occurs 0 or 1 time)

Remarks:

- The info:eu-repo namespace is used with the following syntax: info:eu-repo/_type_/_identifier_
  For more information see

  http://info-uri.info/registry/OAIHandler?verb=GetRecord&metadataPrefix=reg&identifier=info:eu-repo/

- The semantics of the ObjectTypes mean for example that this Item states that the first sub-Item has or contains Descriptive Metadata.

ObjectType: Metadata Item

The first `Item` ObjectType element contains the metadata. The metadata is put in a `Resource` element. Every `Resource` element contains the namespace of a metadata format that has been used. This way the format will be recognised by service providers. According to the OAI protocol it is mandatory to use '`oai_dc`'. For ease of implementation one can use the OAI_DC as metadata, since OAI_DC is a basic requirement of OAI-PMH. Every metadata item can *optionally* have its own `Identifier` and `modified` element in a `Descriptor` element:

```
<didl:Item>
<didl:Descriptor>
<didl:Statement                              mimeType="application/xml">
<dip:ObjectType>
```

status: final 2008-11-13

	`info:eu-repo/semantics/descriptiveMetadata</dip:ObjectType>`
	`</didl:Statement>`
	`</didl:Descriptor>`

1	`<didl:Descriptor> <!-- This metadata instance has its own ID number -`
	`->`
	`<didl:Statement                                mimeType="application/xml">`
	`<dii:Identifier>info:doi/10.1705/74836724783</dii:Identifier>`
	`</didl:Statement>`
	`</didl:Descriptor>`

2	`<didl:Descriptor> <!-- This record has its own Modification date -->`
	`<didl:Statement                                mimeType="application/xml">`
	`<dcterms:modified>2006-12-20T10:29:12Z</dcterms:modified>`
	`</didl:Statement>`
	`</didl:Descriptor>`

	`<didl:Component>`

3	`<didl:Resource  mimeType="application/xml">  <!--  the  DC  data  -->`
	`<oai_dc:dc`
	`xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"`
	`xmlns:dc="http://purl.org/dc/elements/1.1/"`
	`xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"`
	`xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/`
	`http://www.openarchives.org/OAI/2.0/oai_dc.xsd">`
	`<dc:creator>...</dc:creator>`
	`<dc:creator>...</dc:creator>`
	`<dc:title>                       ...                       </dc:title>`
	`...`
	`</oai_dc:dc>`
	`</didl:Resource>`

	`</didl:Component>`
	`</didl:Item>`

Remarks:

1. (Mandatory when applicable) It is recommended to identify every separate component, for future reference or re-assemble purposous. This metadata set has its own identifier, which is NOT the same as the DIDL identifier.

 status: final 2008-11-13

2. If the date of the metadata has been changed, make sure the modification date of the root level Item is also being changed.

3. Declare the dc namespace in the start-tag of the Resource element where you use Dublin core.

ObjectType: Object Item

The second Item ObjectType contains a link to one digital object. This is always "by-reference" to limit the file size, when used for metadata transfer purpouses. ("by-value" is possible but increases the file size and touches the issueon ownership, use base64 encoding, not exampled here), and the `Item` element has an ObjectType statement with an `info:eu-repo/semantics/objectFile` URI. An objectFile Item can occur more than once. See the following:

```
<didl:Item>
...
<!-- Below this line one can find links to one or more digital objects
-->
```
```
<didl:Item> <!-- First Item for a File/Bitstream -->
<didl:Descriptor>
<didl:Statement mimeType="application/xml">
<dip:ObjectType>info:eu-repo/semantics/objectFile</dip:ObjectType>
</didl:Statement>
</didl:Descriptor>
...
<didl:Component>
<didl:Resource mimeType="application/pdf"
ref="http://my.server.nl/report.pdf"/></didl:Component>
</didl:Item>
```
```
<didl:Item> <!-- Second Item for a File/Bitstream -->
<didl:Descriptor>
<didl:Statement mimeType="application/xml">
<dip:ObjectType>info:eu-repo/semantics/objectFile</dip:ObjectType>
</didl:Statement>
</didl:Descriptor>
...
```

status: final 2008-11-13

```
<didl:Component>
<didl:Resource mimeType="application/pdf"
ref="http://my.server.nl/appendix.pdf"/><didl:Component>
</didl:Item>
```
```
<didl:Item> <!-- Third Item for a File/Bitstream -->
<didl:Descriptor>
<didl:Statement mimeType="application/xml">
<dip:ObjectType>info:eu-repo/semantics/objectFile</dip:ObjectType>
</didl:Statement>
</didl:Descriptor>
...
<didl:Component>
<didl:Resource mimeType="application/pdf"
ref="http://my.server.nl/datasheets.xls"/><didl:Component>
</didl:Item>
```
```
</didl:Item>
```

As you can see in the above example, the Resource locations do not appear in several components within one Item, but each Resource location is wrapped in an Item element. The rationale behind this is that each Bit stream of file can have its own Identifier. On the three dots "…" (given in the examples) one may place the Identifier and modified tags, which is similar to the metadata Item.

Remarks:

1. The order of the object components should be in a logical reading order! The Item with chapter 1 should be followed by the next sibling Item element that contains chapter 2, etc… This way the service provider can make a better presentation. Making the order explicit by placing sequence numbers is being specified in the next version of the specification.

2. If there are important modification dates for the Resource element, propagate these date changes upwards though out the parent `Item` elements that encapsulate the modified child `Item` element.

3. Only add Identifiers when there actually are any

4. If there are no Identifiers for the ObjectType Item elements, the Identifier of the DIDL element will be used by the service provider.

 status: final 2008-11-13

5. Use for a modified or Identifier element a separate `<Descriptor>` `<Statement>` element construction

6. The rule of thumb is that if a Bitstream or file has its own identifier, the wrapper is an Item element. To keep the possibility open for a Bitstream to have an Identifier, we use the Item element as default to wrap a resource location.

ObjectType: Jump-off-page Item

The third ObjectType Item element contains a link to the jump-off page or intermediate page. This is done in the same way as for the Object Item element. Currently this is restricted to 1 Item of this type; there are no identifier elements, nor modification date elements present. This Item element is optional:

```
<didl:Item>
...
<!-- Below this line; an Item with a link to one optional Intermediate
page -->
<didl:Item>
<didl:Descriptor>
<didl:Statement                          mimeType="application/xml">
<dip:ObjectType>
info:eu-repo/semantics/humanStartPage
</dip:ObjectType>
</didl:Statement>
</didl:Descriptor>
...
<didl:Component>
<didl:Resource                           mimeType="application/html"
ref="http://my.server.nl/mypub.html"/></didl:Component>
</didl:Item>
</didl:Item>
```

Example of a DIDL embedded in OAI-PMH

```
<?xml version="1.0" encoding="UTF-8"?>
```

status: final 2008-11-13

```
<?xml-stylesheet type="text/xsl" href="DIDL_documentHTML.xsl"?>
<OAI-PMH
xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="
     http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2006-12-20T10:29:11Z</responseDate>
   <request identifier="oai:dspace.library.uu.nl:1874/15290" metadataPrefix="didl" verb="GetRecord">
         http://dspace.library.uu.nl:8080/dspace-oai/request
   </request>
   <GetRecord>
      <record>
         <header>
            <identifier>oai:dspace.library.uu.nl:1874/15290</identifier>
            <datestamp>2006-12-06T19:00:49Z</datestamp>
            <setSpec>hdl_1874_69</setSpec>
            <setSpec>hdl_1874_12233</setSpec>
         </header>
         <metadata>
            <!-- Introducing the DIDL document.  -->

            <!-- Implementation Version 2.3. used in the SURFshare (nl)  and DRIVER (eu) context-->
            <!--
               <didl:DIDL> is the wrapper or container that can be seen as an autonomous entity
               that can exist outside the OAI-PMH context.

               The DIDLDocumentId attribute (optional) is the DIDL identifier
               and it CAN be the same as the record Identifier!
               Leave it out if you have no dedicated DIDL identifier.
              -->
            <didl:DIDL DIDLDocumentId="urn:NBN:nl:ui:10-6748398729821"
            xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
            xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
            xmlns:dip="urn:mpeg:mpeg21:2002:01-DIP-NS"
            xmlns:dcterms="http://purl.org/dc/terms/"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="
                   urn:mpeg:mpeg21:2002:02-DIDL-NS
                      http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/did/didl.xsd

                   urn:mpeg:mpeg21:2002:01-DII-NS
                      http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/dii/dii.xsd

                   urn:mpeg:mpeg21:2005:01-DIP-NS
                      http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/dip/dip.xsd">
               <!-- The Item is the autonomous compound complex entity that is a representation of a
work-->
               <didl:Item>
                  <didl:Descriptor>
                     <didl:Statement mimeType="application/xml">
                        <dii:Identifier>urn:NBN:nl:ui:10-6748398729821</dii:Identifier>
```

```
                    </didl:Statement>
                </didl:Descriptor>
                <didl:Descriptor>
                    <didl:Statement mimeType="application/xml">
                        <dcterms:modified>2006-12-20T10:29:12Z</dcterms:modified>
                    </didl:Statement>
                </didl:Descriptor>
                <didl:Component>

                    <!-- Actual resource of Item; Location of the DIDL document -->
                    <didl:Resource mimeType="application/xml"
                            ref="http://dspace.library.uu.nl:8080/dspace-oai/request?verb=GetRecord

&amp;metadataPrefix=didl&amp;identifier=oai%3Adspace.library.uu.nl%3A1874%2F15290"/>
                </didl:Component>
                <!-- Introducing the area for metadata  -->
                <didl:Item>
                    <didl:Descriptor>
                        <!-- ObjectType of Item -->
                        <didl:Statement mimeType="application/xml">
                            <dip:ObjectType>info:eu-
repo/semantics/descriptiveMetadata</dip:ObjectType>

                        </didl:Statement>
                    </didl:Descriptor>
                    <didl:Component>
                        <!-- Actual resource of Item -->
                        <didl:Resource mimeType="application/xml">
                            <oai_dc:dc xmlns="http://www.w3.org/2001/XMLSchema"
                            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                            xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
                            xmlns:dc="http://purl.org/dc/elements/1.1/" xsi:schemaLocation="
                                http://www.openarchives.org/OAI/2.0/oai_dc/
                                        http://www.openarchives.org/OAI/2.0/oai_dc.xsd
                                http://purl.org/dc/elements/1.1/
                                        http://dublincore.org/schemas/xmls/simpledc20021212.xsd">
                            <dc:title>Neonatal Glucocorticoid Treatment and Predisposition
                                    to Cardiovascular Disease in Rats</dc:title>
                            <dc:creator>Bal, M.P.</dc:creator>

                            <dc:subject>Geneeskunde</dc:subject>
                            <dc:subject>glucocorticoid</dc:subject>
                            <dc:subject>dexamethasone</dc:subject>
                            <dc:subject>
                                <!--etc...-->
                            </dc:subject>
                            <dc:subject>cellular hypertrophy</dc:subject>

                            <dc:subject>contractile proteins</dc:subject>
                            <dc:description>The    present    thesis    describes    the    issue    of
&quot;neonatal glucocorticoid
                                treatment and predisposition to cardiovascular disease in rats&quot;.
</dc:description>

                            <dc:publisher>Utrecht University</dc:publisher>
```

```xml
                    <dc:date>2006-12-12</dc:date>
                    <dc:type>Doctoral thesis</dc:type>


                    <dc:format>image/jpeg</dc:format>
                    <dc:format>image/pdf</dc:format>
                    <dc:format>image/pdf</dc:format>
                    <dc:format>
                        <!--etc...-->
                    </dc:format>
                    <dc:identifier>
                      http://igitur-archive.library.uu.nl/dissertations/2006-1206-
200250/UUindex.html

                    </dc:identifier>


                    <dc:language>en</dc:language>
                    <dc:rights>(c) Bal, M.P., 2006</dc:rights>
                </oai_dc:dc>
            </didl:Resource>
        </didl:Component>
    </didl:Item>
    <!-- Introducing the area for MODS metadata  -->
    <didl:Item>


        <didl:Descriptor>
            <!-- ObjectType of Item -->
            <didl:Statement mimeType="application/xml">
                <dip:ObjectType>info:eu-
repo/semantics/descriptiveMetadata</dip:ObjectType>
            </didl:Statement>
        </didl:Descriptor>
        <didl:Component>
            <didl:Resource mimeType="application/xml">


            <mods version="3.2"
            xmlns="http://www.loc.gov/mods/v3"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
            xsi:schemaLocation="http://www.loc.gov/mods/v3
        http://www.loc.gov/standards/mods/v3/mods-3-2.xsd">
                <titleInfo xml:lang="en">
                    <title> Neonatal Glucocorticoid Treatment and Predisposition
                            to Cardiovascular Disease in Rats </title>
                </titleInfo>
                <name type="personal" ID="n1">
                    <namePart type="family"> Bal </namePart>
                    <namePart type="given">M.P.</namePart>

                    <role>
                        <roleTerm authority="marcrelator" type="code">aut</roleTerm>
                    </role>
                </name>
                <name type="personal" ID="n2">
                    <namePart type="family">Winter, de</namePart>
                    <namePart type="given">R.J.</namePart>
```

status: final 2008-11-13

```
                            <role>
                                <roleTerm authority="marcrelator" type="code">aut</roleTerm>
                            </role>
                        </name>
                        <extension>
                            <daiList  xmlns:dai="info:eu-repo/dai"  xsi:schemaLocation="info:eu-
repo/dai
                    http://www.surfgroepen.nl/sites/oai/metadata/Shared%20Documents/dai-
extension.xsd">
                                <identifier          IDref="n2"          authority="info:eu-
repo/dai/nl">157455590</identifier>
                                <identifier          IDref="n1"          authority="info:eu-
repo/dai/nl">123456678</identifier>

                            </daiList>
                        </extension>
                    </mods>
                </didl:Resource>
            </didl:Component>
        </didl:Item>
        <!-- Introducing the area for digital fulltext objects  -->
        <!--Bitstream no: [0] -->
        <didl:Item>

            <didl:Descriptor>
                <!-- ObjectType of Item -->
                <didl:Statement mimeType="application/xml">
                    <dip:ObjectType>info:eu-repo/semantics/objectFile</dip:ObjectType>
                </didl:Statement>
            </didl:Descriptor>
            <didl:Component>
                <!-- Actual resource of Item -->

                <didl:Resource mimeType="text/html"
ref="https://dspace.library.uu.nl:8443/bitstream/1874/15290/18/index.htm"/>
            </didl:Component>
        </didl:Item>
        <!--Bitstream no: [1] -->
        <didl:Item>
            <didl:Descriptor>
                <!-- ObjectType of Item -->
                <didl:Statement mimeType="application/xml">
                    <dip:ObjectType>info:eu-repo/semantics/objectFile</dip:ObjectType>

                </didl:Statement>
            </didl:Descriptor>
            <didl:Component>
                <!-- Actual resource of Item -->
                <didl:Resource mimeType="image/jpeg"
                    ref="https://dspace.library.uu.nl:8443/bitstream/1874/15290/16/bal.jpg"/>
            </didl:Component>
        </didl:Item>
        <!--Bitstream no: [2] -->
```

status: final 2008-11-13

```xml
        <didl:Item>

            <didl:Descriptor>
                <!-- ObjectType of Item -->
                <didl:Statement mimeType="application/xml">
                    <dip:ObjectType>info:eu-repo/semantics/objectFile</dip:ObjectType>
                </didl:Statement>
            </didl:Descriptor>
            <didl:Component>
                <!-- Actual resource of Item -->

                <didl:Resource mimeType="application/pdf"
                    ref="https://dspace.library.uu.nl:8443/bitstream/1874/15290/15/c1.pdf"/>
            </didl:Component>
        </didl:Item>
        <!--Bitstream no: [3] -->
        <didl:Item>
            <didl:Descriptor>
                <!-- ObjectType of Item -->
                <didl:Statement mimeType="application/xml">
                    <dip:ObjectType>info:eu-repo/semantics/objectFile</dip:ObjectType>

                </didl:Statement>
            </didl:Descriptor>
            <didl:Component>
                <!-- Actual resource of Item -->
                <didl:Resource mimeType="application/pdf"
                    ref="https://dspace.library.uu.nl:8443/bitstream/1874/15290/14/c2.pdf"/>
            </didl:Component>
        </didl:Item>
        <!--Bitstream no: [etc...] -->
        <!-- Introducing the intermediate page -->

        <didl:Item>
            <didl:Descriptor>
                <!-- ObjectType of Item -->
                <didl:Statement mimeType="application/xml">
                    <dip:ObjectType>info:eu-repo/semantics/humanStartPage</dip:ObjectType>
                </didl:Statement>
            </didl:Descriptor>
            <didl:Component>

                <!-- Actual resource of Item -->
                <didl:Resource mimeType="text/html"
                    ref="http://igitur-archive.library.uu.nl/dissertations/2006-1206-
200250/UUindex.html"/>
                </didl:Component>
            </didl:Item>
        </didl:Item>
    </didl:DIDL>
  </metadata>
 </record>
</GetRecord>
```

```
</OAI-PMH>
```

 status: final 2008-11-13

Use of Vocabularies and Semantics

info:eu-repo – A namespace for URI-fying un-URIfied Schema's and Identifiers

The namespace info:eu-repo is registered at http://info-uri.info

This name space is an authoritive placeholder for semantic terms, controlled vocabularies and identifiers.

By using this namespace all the terms used have a "web presence". Therefore it is no longer an arbitrary string, but contains meaning. This utilisation makes it future-proof.

Author Identification

(this information is cited and modified from the European NEEO project[16])

---

[16] Network of European Economists Online (NEEO): project information see http://www.nereus4economics.info/neeo.html. For the DAI information see specifications: http://homepages.ulb.ac.be/~bpauwels/NEEO/WP5/WP5%20Technical%20guidelines.pdf

Building dynamic publication lists per author requires that these authors are unambiguously identified. This is best done through a unique identifier that is assigned to each author of a work. Such an author identifier is called a DAI (Digital Author Identifier).

A DAI can be assigned to authors on a national level (like in the Netherlands where each author gets a unique identifier in the METIS system), or on an institutional level. It is the sole responsibility of each IR to ensure that an author can be identified through a DAI and that each assigned DAI is unique within an IR.

Format of a DAI

Every IR can deliver its DAI's in the format it wants, as long as the authoritive party that acts as a Registration Agency can be recognised in the scheme. However it is recommended to use the International Standard for Name Identification (ISNI)[17] number. All DAI's MUST be globally unique. This is accomplished by combining the DAI with its authority (value of the authority attribute of the identifier element) or by making the DAI a complete URI that is unique. Some examples of valid encodings of a DAI:

```
info:eu-repo/dai/nl/12456454
http://staff.university.eu/19262
urn:isni:1234567-2
```

Persistence of a DAI

DAI's should be Persistent Identifiers: a change of DAI for an author could effectively result in incoherent results for service providers worldwide and publication lists could become incomplete. For example, part of a publication list would be allocated to DAI X, another part to DAI Y, both DAI's referring to the same author. Statistics on downloads of publications per author would also become incorrect. If an institution needs to change the DAI's of its authors, for whatever reason, a complete re-harvest of

---

[17] (ISNI): Standard in development, No Registration Agencies set-up so far. The project finishes in 2009. The DAI numbers in the Netherlands are ISNI compliant due to involvement via OCLC. http://www.collectionscanada.gc.ca/iso/tc46sc9/docs/sc9n429.pdf

the IR should be operated by all service providers and link resolvers on a global scale, in order, for example, to get the publication lists right again. Errors in usage statistics services would probably be irrecoverable. The advice is clearly that DAI's shouldn't change, once they are assigned to authors.

Subject classification

Metadata delivered via OAI-PMH contain a broad range of subject headings and classification information. The used classification and subject heading systems and the presentation formats vary broadly. In most cases this information appears in simple dc format in the subject element. Classification information is often used for groupng a repository into items under discipline orientated aspects. Therefore such information appears frequently in the OAI setSpec element. EPrints repositories (LoC classification) and DINI-certificated repositories (DDC) are examples for this approach.

Most frequent used classification schemes in OAI context are

- Library of Congress Classification[18]
- Dewey Decimal Classification (DDC)[19]
- Universal Decimal Classification[20]

Frequently used subject headings systems in OAI context are

- Library of Congress Subject Headings (LCSH)
- Schlagwortnormdatei (SWD)

Besides this, OAI metadata contain discipline-related classification codes from schemes such as the Mathematics Subject Classification (MSC) and the Medical Subject Headings (MeSH) but also different local classification information.

---

[18] http://www.loc.gov/catdir/cpso/lcco/

[19] http://www.oclc.org/dewey/

[20] http://www.udcc.org/

status: final 2008-11-13

Currently, services based on this information have serious problems to extract the information from the delivered data in an apparopriate way. The first step to improve the situation should focus on making the used technique and classification scheme transparent to the service provider.

DRIVER recomends that the repository should transport the information related to the usage of classification and subject headings in the description element of the Identify response. When a classification is used for structuring the repository via sets, the classification part should be repeated in the subject element.

Best practice is to transport the classification in the element subject "URI-field" using an authoritative namespace in order to support recognizing the classification scheme. Based on this information service providers can use it for establishing services as classification browsing. This includes substituting classification codes by English terms, translating terms to different languages or doing a merge of classification codes using mapping rules.

It is recommended to use an URI when using classification schemes or controlled vocabularies especially when codified schemes are used DDC or UDC. Service providers can recognise encoding schemas more easily when the schema is "URI-fied" by an authority namespace. When the classification scheme is codified, use a human readable text of the code, preferably in English, directly below the codified element. For example:

```
<dc:subject>info:eu-repo/classification/ddc/641</dc:subject>
<dc:subject>Anatomy</dc:subject>
```

If no specific classification scheme is used we recommend the Dewey Decimal Classification (DDC). The first 1000 terms are called the Dewey Decimal Classification Summary and can be downloaded at http://www.oclc.org/dewey/resources/summaries/ if one agrees with the following terms and conditions: http://www.oclc.org/research/researchworks/ddc/terms.htm

Publication type vocabulary

The Publication type vocabulary listed below has a deep history from within the European repository community. It is a combination of the types DARE uses from DC guidelines, types listed in the DINI certificate and the e-Prints publication types[21]. Based on these authoritative guidelines, improved guidelines have been made for DRIVER in "Use of MODS for institutional repositories"[22] which is in line with publication types used by commons Current Research Information Systems (CRIS) like METIS. This document was the basis for the Publication types listed below.

These Publication types below have a strong focus on European Interoperability amoung repositories for exchange purpouses only. The Publication types are used to close the semantic gap by creating a common ground and provide meaning for the different types. The terms and descriptions are chosen in a way that will cover the types used in scholarly communication, diverse enough to distinguish between the different items used in scholarly communication, generic enough for repository managers to fit a suitable mapping and not too specific that they only will apply to one community.

Remark: The Publication types below are developed for exchanging metadata towards service providers aiming at scholarly communication in general, and are not meant for internal repository usage. One should map internal publication types with the ones listed below. The descriptions are carefully assembled with the aid of metadata experts and repository administrators. These descriptions will help the mapping process of the local repository.

For the publication types a special namespace is used in order for humans and machines to recognise the vocabulary that is used. This namespace is the *"info:eu-repo/semantics/"* namespace (see the first column of the following table). The URI is used as a prefix to the term that represents a Publication type. For example, the URI for articles is "*info:eu repo/semantics/article*". The third column contains the

---

[21] Vocabulary of the Eprints Application Profile (Scholarly Works Application Profile - SWAP) http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Type_Vocabulary_Encoding_Scheme
[22]

https://www.surfgroepen.nl/sites/oai/metadata/Shared%20Documents/Use%20of%20MODS%20for%20institutional%20repositories-version%201.doc

descriptions of the Publication types. This should ease the mapping descisions that have to be made at the local repositories.

The second column contains the versions that describe the status of the document. This makes it able to describe the Publication type without mixing the terms with version or status information. The term "PeerReviewedArticle" is split in for example *info:eu repo/semantics/***article** and *info:eu repo/semantics/***accepted**.

info:eu-repo/semantics/	Version allowed	Description
article	accepted / published / updated	Article or an editorial published in a journal
bachelorThesis	accepted / published / updated	Lowest level of a thesis (normally after three years of study). See also http://en.wikipedia.org/wiki/Diplom
masterThesis	accepted / published / updated	Intermediate level of a thesis (normally after four or five years of study). See also http://en.wikipedia.org/wiki/Diplom This also refers to theses of the pre-Bologna period for degrees that are at the same level as what now is known as a master degree.
doctoralThesis	accepted / published / updated	Highest level of a thesis normally after more than four or five years of study. See also http://en.wikipedia.org/wiki/Diplom Also everything equal and higher then a Doctoral thesis, that does not follow the "Bologna Convention", will be put in the category doctoralThesis. A free text field will Provide the opportunity to specify this further.

book	accepted / published / updated	Book or monograph
bookPart	accepted / published / updated	Part or chapter of a book
review	draft / submitted / accepted / published / updated	Review of a book or article
conferenceObject	draft / submitted / accepted / published / updated	All kind of documents related to a conference, p.e. conference papers, conference reports, conference lecture, papers published in conference proceedings, conference contributions, reports of abstracts of conference papers and conference posters.
lecture	draft / submitted / accepted / published / updated	Lecture or presentation presented during an academic event, e.g., inaugural lecture. Excluded is a conference lecture (see conferenceItem).
workingPaper	draft / submitted	a preliminary scientific or technical paper that is published in a series of the institution where the research is done. Also known as research paper, research memorandum or discussion paper. **The difference with a preprint is that a workingPaper is published in a institutional series.** Examples are: working papers, research papers, research

118/137   status: final 2008-11-13

		memoranda and discussion papers.
preprint	draft / submitted	like a workingPaper this is a preliminary scientific or technical paper, **but it is not published in a institutional series**. The paper is intended to be published in a scientific journal or as a chapter in a book.
report	draft / submitted / accepted / published / updated	This is a more or less a rest category and covers commission reports, memoranda, external research reports, internal reports, statistical report, reports to funding agency, technical documentation, project deliverables etc. Excluded are conference reports (See conferenceItem).
annotation	draft / submitted / accepted / published / updated	Note to a legal judgment
contributionToPeriodical	draft / submitted / accepted / published / updated	Contribution to a newspaper, weekly magazine or another non-academic periodical
patent	draft / submitted / accepted / published / updated	Patent
other	draft / submitted / accepted / published / updated	Especially meant for non-publication data like research data, audio-visual materials, animations etc.

Derived from

- the e-print type vocabulary http://purl.org/eprint/type/

Usage examples with the complete string including the URI info:eu-repo:

```
<dc:type>info:eu-repo/semantics/article</dc:type>
<dc:type>info:eu-repo/semantics/accepted</dc:type>
```

The string "info:eu-repo" is always attached to the term. It therefore sets the authority of the used controlled vocabulary.

The namespace info:eu-repo is registered at http://info-uri.info

More about the usage of DC:type with versioning see section *Type* on page 68 in chapter "Use of Metadata OAI_DC"

Version vocabulary

This section is about the versions that describe the status of the document. We have introduced version information to make it possible to describe the Publication type without mixing the terms with version or status information. For example, the term "PeerReviewedArticle" can be split into *info:eu repo/semantics/*article and *info:eu repo/semantics/*accepted.

The version vocabulary is derived from http://www.lse.ac.uk/library/versions/, which is a JISC funded project called VERSIONS (Versions of Eprints – a user Requirements Study and Investigation Of the Need for Standards). This project addresses the issues and uncertainties relating to versions of academic papers in digital repositories. VERSIONS aims to help build trust in open access repository content among all stakeholders and has developed a toolkit that can be found at: http://www.lse.ac.uk/library/versions/VERSIONS_Toolkit_v1_final.pdf

info:eu-	Description

repo/semantics/	
draft	Early version circulated as work in progress
submittedVersion	The version that has been submitted to a journal for peer review
acceptedVersion	The author-created version that incorporates referee comments and is the accepted for publication version
publishedVersion	The publisher created published version
updatedVersion	A version updated since publication

Encoding schemes

The DRIVER Guidelines use the following encoding schemes:

Name	*Field*	*Scheme*
Author	dc:creator	APA bibliographic writing style as in a reference list. Syntax: surname, initials (first name) [http://en.wikipedia.org/wiki/Apa_style#Reference_list]
Contributor	dc:contributor	APA bibliographic writing style as in a reference list. Syntax: surname, initials (first name) [http://en.wikipedia.org/wiki/Apa_style#Reference_list]
Languages	dc:language	ISO 639-3 Syntax: 3 characters [http://www.sil.org/ISO639-3/codes.asp]
Dates	dc:date	ISO 8601 [W3CDTF] Syntax: YYYY-MM-DD , MM and DD are optional [http://www.w3.org/QA/Tips/iso-date]
Formats	dc:format	IANA registered list of Internet Media Types (MIME types) [http://www.iana.org/assignments/media-types/]
Territory	dc:coverage	ISO 3166 (Countries) [http://www.iso.ch/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/index.html]
Area	dc:coverage	Box [http://dublincore.org/documents/dcmi-box/]
Geographic names	dc:coverage	TGN [http://www.getty.edu/research/tools/vocabulary/tgn/]

Time period	dc:coverage	DCMI Period [http://dublincore.org/documents/2000/07/28/dcmi-period/]
Citation info	dc:source	Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata [http://dublincore.org/documents/dc-citation-guidelines/] as in dcterms:bibliographicCitation

Annexes: Future Points of Interest

Annex: Use of Quality Labels

The DRIVER Guidelines 2.0 provides basic information on the importance of Quality, and Interoperability. Quality labels can be used to assure Stable and reliable repositories that last longer than the hype, and have also an archival purpose for Long Term Preservation.

Examples of Quality labels can be: the Data Seal of Approval and the DINI Certificate.

Annex: Use of Persistent Identifiers

Persistent Identifiers for web resources are needed to create a stable and reliable infrastructure. This does not concern technicalities, but mainly agreements on organisational level.

DRIVER Guidelines could make some recommendations on the implementation for repository managers. This is based on the Report on Persistent Identifiers of the PILIN project.

An implementation plan has been provided below.

It should be made clear how this fits in with oai_dc exchngge of metadata

---

In the era of paper the International Standard Book Number (ISBN), a unique, numerical commercial book identifier, was developed. Each edition and variation (except reprinting) of a book is given an ISBN. In the digital age, there is a growing need for such a unique, numerical, identifier for digital publications as well. Moreover, not just for publications, but for all kinds of digital objects.

status: final 2008-11-13

On the Internet, we consider the URL as the identifier of a digital object. However, we are all familiar with broken or dead links that point to web pages that are permanently unavailable.

An URL might change overtime, due to server migrations and other technical reasons. With undesired consequences for links and citations within scholarly communication.

Therefore a 'persistent identifier' is needed with which a digital object is permanently associated. This persistent identification number always refers to the digital object to which it has been assigned, regardless of the underlying locator technology (at the moment these are web addresses; in the future, however, an object's location may be completely different).

In several countries, a system for such a persistent identifier has been developed and 'national resolvers' have been set up. A resolver is a transformation and redirection service, transforms a string of characters to an URL, and is hosted by a national organisation. Common identifiers in the case of scholarly communication are DOI, Handle and URN:NBN. In case of DOI and Handle the resolution mechanism is located in the US at CNRI[23]. In case of URN:NBN resolution mechanisms are hosted by a national organisation, often this is done by the National Library.

Every digital object is assigned a number that represents that object forever. Even if technology moves on, the national organisation will ensure that the documents can be read. But the documents must be traceable as well. The Persistent Identifier ensures that it can be located. A stable information infrastructure makes research citations a lot more reliable.

Currently the URN:NBN and the Handle are popular ways for Persistent Identifiers. Since the URN:NBN namespaces are distributed in a controlled manner, we would expect it will be recognised as authoritative as the DOI has as a reputation.

---

[23] CNRI: http://www.cnri.reston.va.us/

The differences between Persistent Identifiers are described by Hans-Werner Hilse and Jochen Kothe in Implementing Persistent Identifiers[24]. There is also an article Persistent Identifiers: Considering the Options[25] in Ariadne, issue 56 by Emma Tonkin

Using Persistent Identifiers involves an obligation for the repositories to sustain persistence of the Identifier over a long period of time! This persistence can be guaranteed in so called "trusted repositories" with the appropriate certification. See chapter Annex: Use of Quality Labels on page 124.

for more information see http://www.persistent-identifier.de

and https://www.pilin.net.au/

The Scandinavian countries, Germany, the Czech Republic and the Netherlands are using URN:NBN. The main reason for choosing urns is because it is an internet standard that is future proof. The only drawback now is that a urn is not actionable without using an http resolution address as a prefix. Further work is still needed to be done to integrate URN in the DNS system[26] by using NAPTR records[27] that is also used for VOIP phone calls.

Recently Norway, Sweden, Finland, and the Netherlands have come to a promising proposal for a Global Resolver of Persistent Identifiers (URN:NBN). In cooperation with representatives of the Hopkins and Berkeley Universities (US) a working proof of concept[28] of a global resolver (GRRS) has been developed. This GRRS integrates four different national resolvers into one global resolver. The GSRS (n2t.info) receives the Identifier from a browser plug-in and redirects the browser to the appropriate national

---

[24] Hilse, H., Kothe, J., Implementing Persistent Identifiers, KNAW,
http://www.knaw.nl/ecpa/publ/pdf/2732.pdf

[25] Tonkin, E., Persistent Identifers: Considering the Options, Ariadne, issue 56,
http://www.ariadne.ac.uk/issue56/tonkin/

[26] DNS-URN integration

http://www.persistent-identifier.de/english/335-project-proposal.php#URNscope

[27] NAPTR Record: http://en.wikipedia.org/wiki/NAPTR_record

[28] Global Resolution Proof of Concept:

http://www.surfgroepen/sites/surfshare/public/software/pihandler

resolver where the browser again is redirected to the current location of the web resource. The architecture of this multi-system process is depicted below.
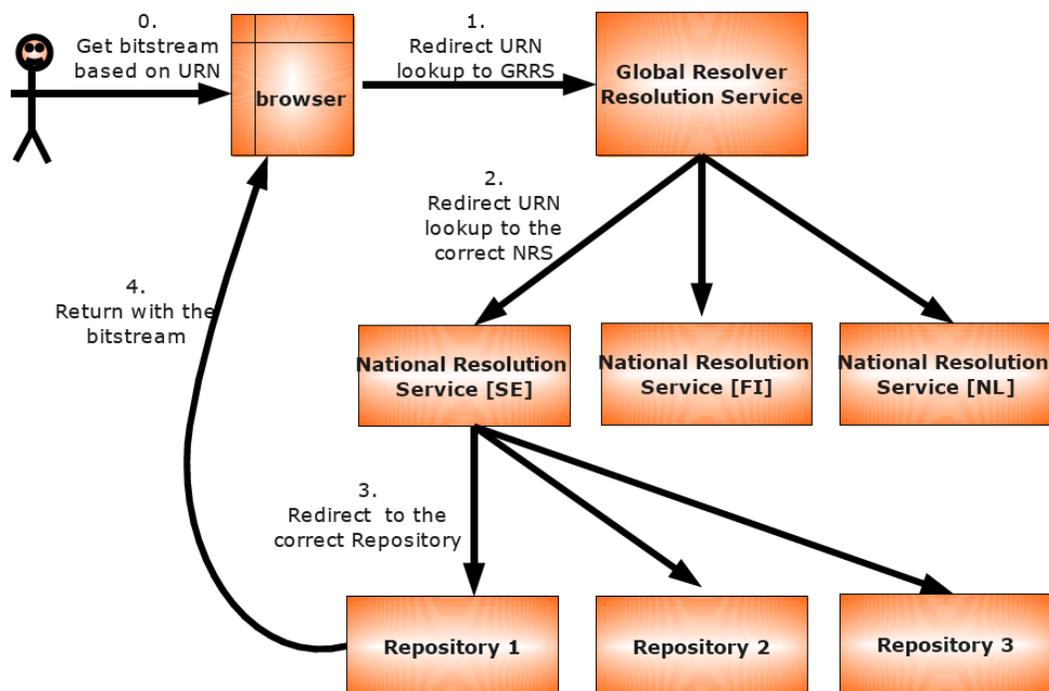
Implementation plan on using URN:NBN Persistent Identifiers

First of all we would like to say that the persistency of Identifiers and web resources is not about the technology one uses, but about organisation and sustainable business models. For more information about Persistent Identifier policies take a look at the successful Persistent Identifier Linking (PILIN) project[29] in Australia that is part of the ARROW[30] project.

To setup a persistent Identifier program based on National Bibliographic Numbers (NBN) URN identifiers and a resolver one needs to take the following steps:

1. **Work group:** Create a work group that manages all the technical and organisational details of such project. Also think about the syntax that is going to be used. For example urn:nbn:{country}:{sub-namespace}:{repositoryid}-{localid}. Country is the

---

[29] Persistent Identifier Linking Infrastructure project: https://www.pilin.net.au/

[30] ARROW project: http://www.arrow.edu.au/

short name of the country, sub-namespace represents web resources that come from the repositories, repositoryid is a two digit representation of the repository and local id is the Identifier generated at the repository. This can for example result in the following Identifier for one publication urn:nbn:ie:ui:21-1234/5678 .

2. **Formalities:** Since the urn:nbn:ie namespace is by default claimed by the National Library, one has to arrange an agreement with the National Library to use a sub-namespace for scientific material. This name should be short and have no semantic meaning. For example urn:nbn:ie:ui, or urn:nbn:ie:oa, or urn:nbn:ie:sp.

3. **Registration Agency:** Create a registry in which repositories are given a short random number of two digits. This will create a sub-namespace in which a repository autonomously can distribute Persistent Identifiers for their publications. For example Trinity College Dublin (TCD) is registered as 21. The namespace for TCD to operate in will be urn:nbn:ie:ui:21.

4. **Implementation at local level:** Each repository must generate Persistent Identifiers for each publication within their namespace that is provided and store this identifier in the database record. For example TCD can use existing identifiers to add after their namespace followed by a dash. In case TCD uses handle, the Identifier for one publication could look like the following urn:nbn:ie:ui:21-1234/5678. In case TCD uses database numbers urn:nbn:ie:ui:21-15874. (Make sure to store the identifier and not generate them on-the-fly. In case of database migrations these numbers might change and persistency is lost.)

5. **Transport of identifiers and URL's:** Each repository must generate a DIDL package in which the URN and URL are included. See the MPEG-21 DIDL section in the main report.

6. **National Resolution Service:** A national resolver can be made by harvesting the DIDL packages from each repository where the URL and URL bindings are extracted and stored. A web location must be created where the user or machine can go to for resolution of the identifier. For example http://resolver.ie where the user can insert an identifier and receive the current location of the web resource.

For   example   http://resolver.ie/urn:nbn:ie:ui:21-1234/5678   resolved   to
http://repository.tcd.ie/1234/5678

Annex: Use of Usage Statistics Exchange

This section will not appear in the DRIVER Guidelines 2.0 Final release. The input for this section will be make from the experiences and best practices that comes from the two European projects who harvest COUNTER reports from repositories to present statistics on an aggregated level.

PIRUS: Publisher and Institutional Repository Usage Statistics

"The aim of this project is to develop COUNTER-compliant usage reports at the individual article level that can be implemented by any entity (publisher, aggregator, IR, etc.,) that hosts online journal articles and will enable the usage of research outputs to be recorded, reported and consolidated at a global level in a standard way."

Cited from http://www.jisc.ac.uk/whatwedo/programmes/pals3/pirus.aspx

Project contact: Peter Sheperd at pshepherd@projectcounter.org

OA-Statistik

"The ease of access experienced with Open Access publications lacking any need for authentification, financial transactions or personal identification makes it much easier to achieve a satisfying level of reception in a scientific community. This and similar hypotheses can be investigated by empirical analysis.

1. What data needs to be gathered?
2. How can it be transferred to the statistics provider?

Open-Access-Statistics (OA-S) is a joint project addressing these questions. Starting in July 2008 an infrastructure for the standardised accumulation of heterogeneous web log data with an emphasis on institutional repositories will be built. In tight cooperation with the Network of Open Access Repositories (OA-N) various added value services will be made available to users."

Cited from http://www.dini.de/projekte/oa-statistik/

Project contact: Nils K. Windisch at windisch@sub.uni-goettingen.de

Preliminary results of the project OA-Statistik

Goals of OA-Statistics

We aim to produce valid and reliable document usage statistics based solely on information gathered from the HTTP layer.

There are two main issues addressed by all existing standards which generate the bulk of the necessary corrections:

- Identification of non-human access
- Multi-Click correction

Besides this, we investigate the amount of data and effort necessary to produce complex statistics, for example, click-streams, without violating privacy laws. At the

bottom of this page there is a comparison table including links to all standards mentioned. A detailed description of OA-S can be found at http://www.dini.de/projekte/oa-statistik/#c1203

Usage statistics - and even more important raw usage data - have to be described on an abstract level. It is not sufficient to define a derivative of the Apache Access Log as there is a multitude of different software solutions in use to operate a full text repository. Many do not even produce a log file let alone utilise an Apache Server.

Information needed to generate COUNTER, LogEc and IFABC

Note: The field names might still be subject to change as the project goes on.

OA-S-Fieldname	Description	COUNTER	LogEc	IFABC|-
Document-Identifier	non-ambigious label identifying the full text	needed	needed	needed|-
File Format	File format of server reply (e.g. HTML orPDF)	needed	needed	needed|-
Service Type	nature of server reply (e.g. full text,ab-stract)	needed	needed	-|-
Time of Request	Time of request processing to the second	needed	needed	needed|-
IP	IP-Adress of user (Client)	needed	needed	IF Session-Identifier is not available: needed|-
Session-Identifier	server generated non-ambiguous session/visit label	optional	-	IF IP is not available: needed|-
User Agent	User-Agent-String of the requesting client	needed	needed	IF Session-ID is not available: needed|-
HTTP Status Code	Server-Status-Code of the HTTP-Requests	needed	needed	needed|-
Bytes sent	server reply size	-	-	IF File Format is not

				HTML: needed

Additional pieces of information which comply with OpenURL Context Objects

The following fields are important to our advanced research interests and thus implemented from the beginning.

Referrer	non-ambigious identifier of the server which created the ContextObject|-
Referring Entitiy	non-ambigious label of the object of origin (e.g. the Abstract Page which links to the full text file)

Additional suggestions

States and properties of the repository software have to be delivered from the available data.

Examples:

- Focus Page in Search Result Paging View
- ID of the current document
- Search arguments and result presentation
- Abstract Page vs. Fulltext Page
- Administrative actions
- Document upload
- Metadata allocation

There should be reliable information about the origin of the client (i.e. the referrer). For example, it should be possible to tell whether a client accessed the file via the frontpage or via a link in the repository's RSS-Feed.

In case of multiple server logs it is mandatory to synchronize the system time on all associated repository servers.

Table of Web Usage Standards

Provider URL	Counting Clause	Multi-Click Time Span	User Identification	Crawler Clause	Crawler Identification	Crawler Count Report
Counter Code of Practice Draft 3	HTTP Status Code is 200 or 304.	for HTML 10s; for PDF 30s	at least IP, preferably Session	robots, prefetches, caching, federated searches(n.a.)	Black-List, client HTTP header	separate report
About LogEc	HTTP Status Code is 200, 206, 301, 302 or 304.	one calendar month	IP	robots, automated downloads (wget)	Access of robots.txt; # of requests 10,000 items/month; C-Class access 10% of stock; known robot-Domain/IP	separate column in report
Interoperable Repository Statistics	HTTP Status code is 200 on abstract or full-text page	24 hours	IP	search engine crawlers + automated|AWStats' black list|discarded		
AWStats	Default: HTTP Status codes {200;304}	Default: 1 hour	IP	search engine crawlers	Black-List	separate column in report
IFABC	HTML: Tracking Pixel; Other: bytes transferred 95% of file size	Each Pageview is counted only once per visit. Visit means series of clicks coming from one IP-Number/Session-ID less than 30 minutes apart.	IP+User-Agent; Cookie-Session, Login-Session	search engine crawlers; automated downloads (optional)	proprietary Blacklist	discarded

Use of Intellectual Property Rights (IPR)

This section addresses an important issue on Usage Rights and Deposit Rights. In practice this must be implemented. The DRIVER Guidelines should say something on how Usage Rights should be exposed and formatted in metadata.

The basis of this section will be the Copyright Toolbox developed by SURFfoundation and JISC that reflect the Zwolle principles.

See: http://copyrighttoolbox.surf.nl/copyrighttoolbox/ for more information.

For more information about copyright and the licences to deposit, to use and reuse, see http://www.surffoundation.nl/smartsite.dws?ch=AHO&id=13591

With Open Access, the Intellectual Property Rights must be managed in a correct way. Even if the document is Open Access available, copyright can limit the use of the material that has been found. Creative Commons provides free tools that let authors, scientists, artists, and educators easily mark their creative work with the freedoms

they want it to carry. You can use CC to change your copyright terms from "All Rights Reserved" to "Some Rights Reserved."

For science, in order to spread the knowledge as freely as possible, without losing the notion of ownership, one could use the Creative Commons license BY-SA in your jurisdiction area.

This means

- *SA - Share Alike*: everyone is allowed to use your material, even commercial use is allowed
    - o Remark 1: every party, commercial or not, have to use the same license for their derived work. As a result: knowledge will not be locked in.
    - o Remark 2: however, innovation speed could be slowed down, because some parties do not want to use the same license model when making derivative work.
- *BY*: everyone always have to refer to your name as the original creator (so you also will get credits for contributing).

If you use copyright, we recommend using copy rights with a good usage description. For example http://creativecommons.org/licenses/by-sa/3.0/nl/

In Unqualified Dublin Core the licenses become machine readable by using the following:

```
<dc:rights>http://creativecommons.org/licenses/by-
sa/2.0/uk/</dc:rights>
<dc:rights>cc-by-sa, Andrew Smith</dc:rights>
```

For a complete technical overview see section Rights on page 79.

For more information see also

- http://copyrighttoolbox.surf.nl/copyrighttoolbox/
- http://sciencecommons.org/projects/publishing/
- http://creativecommons.org
- http://www.surffoundation.nl/smartsite.dws?ch=AHO&id=13591

status: final 2008-11-13