SEVENTH FRAMEWORK PROGRAMME

CAPACITIES

Research Infrastructures

INFRA-2007-1.2.1 Research Infrastructures

DRIVER II

Grant Agreement 212147

"Digital Repository Infrastructure Vision for European Research II"

Report on Object Models and Functionalities

D4.2.

Document Description

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

Document	
Deliverable number:	D4.2
Deliverable title:	Report on Object Models and Functionalities
Contractual Date of Delivery:	November 2008
Actual Date of Delivery:	December 2008
Editor(s):	Thomas Place (University of Tilburg) Maurice van der Feesten (SURF Foundation) Maarten Hoogerwerf (DANS) Magchiel Bijsterbosch (SURF Foundation) Martin Slabbertje (University of Utrecht) Arjan Hogenaar (KNAW)
Author:	Peter Verhaar (Leiden University)
Reviewer(s):	UvA, KNAW, RUG, SURF
Participant(s):	SURF; UGent
Workpackage:	WP4
Workpackage title:	Discovery
Workpackage leader:	SURF
Workpackage participants:	SURF, DTU, CNR, UniBI, UGOE, UKOLN, UGENT
Distribution:	General Public
Nature:	Report-book

Version/Revision:	Final version
Draft/Final:	
Total number of pages: (including cover)	39
File name:	Deliverable 4.2
Key words:	Enhanced Publication, metadata format, Object Model, OAI-ORE, compound objects

Disclaimer

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

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

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

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

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

DRIVER is a project funded by the European Union

Table of Contents

Table of Figures

Summary

This report identifies the requirements for storing and managing enhanced publications within the DRIVER infrastructure. Enhanced publications are defined as compound digital objects which combine ePrints with one or more data resources, one or more metadata records, or any combination of these. ePrints are understood as a textual resource with original scholarly work which are intended to be read by human beings, and which put forward certain academic claims. It usually contains an interpretation or an analysis of certain primary data. Enhancing a publication involves adding one or more resources to this ePrint. The report has identified ten requirements:

1. It must be possible at any moment to specify the component parts of an enhanced publication.
2. Enhanced publications must be available as web resources that can be referenced via a URI. The same goes for its components.
3. It must be possible to add compound digital objects to the publication.
4. It must be possible to keep track of the different versions of both the enhanced publication as a whole, and of its constituent parts.
5. It must be possible to record basic properties of the resources that are added to the publication.
6. It must be possible to record authorship of the enhanced publication in its entirety and authorship of its component parts.
7. It must be possible to secure the long-term preservation of enhanced publications.
8. It must be possible to record the relations between the web resources that are part of an enhanced publication.
9. Institutions that offer access to enhanced publications must make sure that they can be discovered.
10. Institutions that provide access to enhanced publications must ensure that these are available as documents based on the OAI-ORE model.

These requirements are clarified visually in the form of an entity-relation diagram. This diagram represents the notion that enhanced publications may consist of five types of entities, namely ePrints, data objects, metadata records, compound datasets, and other enhanced Publications. One ePrint must minimally be present. Compound datasets are created by combining a data object with one or more metadata records, or with one or more other data objects. The report also discusses various vocabularies that can be used to describe the properties of enhanced publications and also provides guidelines on how the enhanced publication can be serialised using an RDF/XML syntax.

1 Introduction

One of the most striking properties of current e-Research projects is their unprecedented data intensity. Disciplines such as astrology, chemistry, geology and archaeology increasingly make use of network technologies, automated instruments, image capture techniques and simulation software. Such technologies have had a vast impact on the way in which scientists can conduct and disseminate their research. Hey and Treffenden (2003) speak of a "data deluge" and note that scientists currently "generate several orders of magnitude more data than has been collected in the whole of human history" (p. 3). For many scientific communities, the curation and the continued accessibility of such vast quantities of research data obviously poses a serious challenge. Unfortunately, much of the data that is produced, often at high costs, also gets lost.

Within various disciplines, efforts have been taken to develop repositories geared especially towards the curation of research data. This is also fully in line with the 2003 Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, which states that open access contributions can now also include "original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material".[1] Examples of projects that focus on the stewardship of raw data include the EMBL Nucleotide Sequence Database,[2] ARROW, the Australian Research Repositories Online to the World project,[3] and eCrystals, an archive for Crystal Structures created by the Southampton Chemical Crystallography Group and EPSRC UK National Crystallography Service.[4] The electronic archiving of data is also stimulated actively by funding agencies that demand more and more that research projects secure the submission of research data in trusted repositories.

Open access publishing of scientific data can yield a number of important advantages, especially in combination with the on-line availability of academic manuscripts. When researchers have deposited their raw data, this may enable peers to replicate and, thus, to verify the claims that are made in scientific publications. In addition, it enables other investigators to re-use the data and to compare and combine them with other data so that new research can be generated. An additional benefit is that it will also become possible to trace the lineage of the various products of e-research projects. Research projects normally evolve through various stages such as data capture, processing, modelling and interpretation. It would be very helpful if it were possible to highlight the various

[1] http://oa.mpg.de/openaccess-berlin/berlindeclaration.html
[2] http://www.ebi.ac.uk/embl/
[3] http://www.arrow.edu.au/
[4] http://ecrystals.chem.soton.ac.uk/

connections between the resources that are produced during the different stages in the scientific process.

The current infrastructure for academic communication still focuses strongly on the storage and dissemination of individual resources. Libraries and publishers currently use the web primarily to provide access to single articles or single monographs. Many academic publishers do not accept other products of e-research projects, such as databases, video recordings, and web services. At the same time, data repositories rarely link data to the publications in which these data are discussed. An important next step in enhancing the infrastructure for e-science is thus to devise a system that can interconnect related scientific web resources. Such an architecture should enable authors to provide access to all the results of the full scientific process. On the basis of such an improved infrastructure, new services can be built that should enable researchers to reuse existing results, and to exchange scientific and scholarly resources across institutions, across disciplines and across repositories.

In the last few years, various studies have investigated the possibility to intertwine distributed e-Research products. Hunter (2007) envisions the creation of “Scientific Publication Packages” which are described as “compound digital objects that encapsulate and relate the raw data to its derived products, publications and the associated contextual, provenance and administrative metadata” (p. 33). Similarly, the Object Reuse and Exchange (OAI-ORE) working group has developed “standards for the description and exchange of aggregations of Web resources”. This report is deeply indebted to the findings from these studies. In addition, within the DRIVER-II project, deliverable 4.1.,the *Report on Enhanced Publications: State of the Art* and deliverable 4.3., the *Technology Watch Report*, have also provided many valuable insights.

2 Aims of this report

This report is the second deliverable of work package 4 of the DRIVER-II project, which aims to investigate the ways in which the availability of research data can be used to enhance the traditional academic publication. Such combined packages of text and research data are referred to as 'enhanced publications'. The aim of work package 4 of DRIVER-II is to ensure that the development of innovative services for the management of enhanced publications is driven by effective application domain requirements.

This report will identify the requirements for storing and managing enhanced publications within the DRIVER infrastructure. Chapter 3 firstly gives a description of the term 'enhanced publication'. Chapter 4 provides an overview of the most important technical and functional requirements. It will also give a number of recommendations for the implementation of enhanced publications. On the basis of these requirements and recommendations, a data model has been developed. It is presented in chapter 5. Finally, chapter 7 will provide recommendations for a serialisation of this data model in XML.

This report aims to propose a generic model for the storage and management of enhanced publications. As such, this report prepares some of the ground for D9.3, which describes the specification requirements for a new DRIVER service, Active Information Discovery, whose main task is to demonstrate the concept of enhanced publications. In addition, this report will also ensure that the model that is proposed in this report will be in line with more encompassing vision on compound digital objects that is presented in D8.1.

The model that is discussed in this report may also be adopted by other projects that intend to publish textual documents in combination with related resources. It must be emphasised, however, that this report offers a very broad perspective on this topic. Many details of the implementation will still need to be filled in by individual projects. This qualification of the scope of this report is inevitable, since issues related to terminology, data reuse, methodologies and certification will often be specific to a particular discipline, or perhaps even to a particular research institute or laboratory. The main aim is to formulate guidelines that can enable system developers to set up the general outline of a technical infrastructure for enhanced publications. The model that is proposed is also intended to stimulate further discussions on this relatively new phenomenon in academic communication.

The meaning of modal verbs such as 'must, 'shall', 'should', 'may' in chapters 5 and 6 of this report conform to the description of these words in RFC 2119.[5]

[5] http://www.openarchives.org/ore/1.0/datamodel#RFC2119

3 Definitions and principles

This chapter will firstly provide a working definition of the term 'enhanced publication'. It may be said that the impetus to enhance publications emerged from the realisation that the traditional publication is limited in its capacity to incorporate the results from the entire scientific discovery process. Especially when large data sets have been generated, an academic text can normally present the research data in a condensed form only. Cheung et al. (2008) note that scientific publications "inadequately represent the earlier stages [of the scientific process] that involve the capture, analysis, modelling and interpretation of primary scientific data" (p. 1). This limitation has become more problematic in recent years since many scientific disciplines are currently producing digital data at highly prodigious rates and in ever growing quantities. Borgman (2007) argues that the "predicted data deluge is already a reality in many fields" (p. 113).

Fortunately, the data that are produced are increasingly stored in trusted data repositories. The aim of such data curation is to ensure that scholarly and scientific materials can be preserved and reused. However, a major shortcoming in the current infrastructure for academic communication is that these datasets are usually not connected to the scientific publications in which they are discussed.[6] Enhanced publications are created with the aim of bridging this imminent gap between the contents of institutional repositories and the contents of data repositories. They will ultimately enable agents to access the complete results of academic studies, and, in addition, to trace the workflow that was followed in research projects.

Enhanced publications are envisioned as compound digital objects which can combine various heterogeneous but related web resources. The basis of this compound object is the traditional academic publication. This latter term refers to a textual resource with original work which is intended to be read by human beings, and which puts forward certain academic claims. Following the *Investigative Study of Standards for Digital Repositories and Related Services*, a study carried out as part of the DRIVER project, this report will use the term 'ePrint' in this context. This term is defined by Foulonneau and André as an "electronic version of academic research paper" (p. 109). An ePrint is understood as a scholarly work which contains an interpretation or an analysis of certain primary data, or of derivation from these materials. Examples of ePrints may include dissertations, journal articles, working papers, book chapters or reports.

In the humanities and in the social sciences, it may frequently be the case that textual materials such as electronic transcriptions or e-Books actually form the object of academic research. When such textual resources are used as primary data in a study, they are not

[6] A notable exception is the CDS info hub in astronomy: http://cdsweb.u-strasbg.fr/

considered ePrints, as this latter term is assumed to be a text that discusses the outcome of an investigation of such source materials. In a growing number of disciplines, it can also be seen that non-textual resources will be accepted as academic publications. Activities such as the creation of an authoritative database, or the determination of a crystal structure can also form the basis of academic rewards. Nevertheless, this report will restrict itself to publications in the traditional sense.

Enhancing a publication involves adding one or more resources to this ePrint. These can be the resources that have been produced or consulted during the creation of the text. In general, these additional resources support, justify, illustrate or clarify the scientific claims that are put forward in a publication. The regular situation will be that an academic manuscript is stored in an institutional repository and that other components, potentially from other repositories, are added to this publication as part of the workflow in scientific research projects.

A basic example of an enhanced publication may consist of an ePrint combined with a metadata record. It is assumed that such a metadata record can be made available as an individual resource, for instance, as an XML stream that is returned as the result of an OAI-PMH *getRecord* request. ePrints will usually be described using descriptive or bibliographic metadata. Examples of descriptive metadata standards include the Metadata Object Description Schema (MODS), Qualified Dublin Core (QDC) and Machine Readable Cataloguing (MARC).

A second type of entity that can be connected to the ePrint is data. OECD, in its *Recommendation of the Council concerning Access to Research Data from Public Funding*, describes research data as "factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings". [7]More concretely, the broad term "data" may refer to any of the following types of objects:

- Data collections containing, for instance, the results of experiments, measurements performed by technical instruments or the results of surveys
- Data visualisations, such as graphs, diagrams, tables, or 3D models
- Machine readable chemical structures
- Multimedia files such as images, video files or audio recordings
- Mathematical formulae, possibly expressed in MathXML, or algorithms
- Text documents that form part of a corpus created for research purposes
- Software, which may be provided as source code, or implemented as web services
- Commentaries and annotations made by agents who have consulted digital objects. Notes on why certain components are relevant or valuable for a specific line of research can be very useful for other academics.
- Specifications of instruments or other hardware

[7] http://webdomino1.oecd.org/horizontal/oecdacts.nsf/Display/?OpenDocument

- Digital certificates for research instruments

Such related data objects may also be described in specific metadata records. Research data can be described on the basis of the Data Documentation Initiative (DDI), which is a highly expressive standard that can be used to describe, among other things, the coverage of a study, and the methodologies that were used. Preservation metadata capture information that is needed to ensure the long-term curation of information resources. PREMIS is the most widely accepted standard in this context. Technical metadata standards such as NISO/MIX provide a format for storing technical aspects of resources. A more advanced example of an enhanced publication may thus consist of a combination of an ePrint, metadata for this ePrint, data objects, and metadata for these data objects.

In conclusion, enhanced publications can be defined as compound digital objects which combine ePrints with one or more metadata records, one or more data resources, or any combination of these. Since the DRIVER project focuses strongly on academic publications in the traditional sense, it is assumed in this report that an enhanced publication must minimally include one ePrint.

4 Requirements and recommendations

4.1 Specification of the structure of enhanced publications

In chapter 3, it was explained that enhanced publications are understood as compound digital objects that combine ePrints, research data and metadata. The creation of such scholarly packages should eventually become part of the natural working environment of scientists. Ideally, some simple tools should be developed that enable academics to archive their data and the descriptions of these data in a digital repository shortly after their creation. Seringhaus and Gerstein (2007) express a similar need. They argue that a new information architecture should be developed which must "ensure that every author is able to archive pre-prints, host supplementary data, and make their findings available in digital format".

It is advisable to make sure that the manuscripts and research data that are deposited in repositories also become part of the web architecture. This implies that all these objects become available as web resources that can be referenced via a URI. Although it is certainly not always feasible, a recommended practice would be to associate globally unique and persistent identifiers with each of these resources. Foulonneau and André (2008) provide an overview of the identifiers that are currently in use within various scientific communities. When scientists publish and share multiple heterogeneous resources through the web, it is essential to ensure that there is a framework that enables scientists to specify which resources belong together. Enhanced publications are produced precisely for this purpose. They can be understood as envelopes, or as enumerations that provide an overview of which ePrints, research data and metadata are published in conjunction.

It is advisable to separate the identification of a resource from its localisation. The latter task should be the responsibility of resolvers that are capable of translating the identifier to a certain bitstream that can be accessed at a specific network location. Such a resolver should also be able to adapt in the situation where one digital object is moved from one repository to the other. This provision is needed to ensure that references to the components of the enhanced publication are stable and reliable.

The components of an enhanced publication do not necessarily have to be stored in a single repository. They may be distributed over different network locations. When publications are enhanced with resources that are maintained at various locations, this may give rise to certain legal issues. Authors of enhanced publications whose components span multiple institutions are advised to ensure that they also have permission to aggregate these different resources. Furthermore, the architecture for enhanced publications should also support references not only to resources in their entirety, but, under certain conditions, also to specific locations within these resources. For instance, it should be

possible to point to a specific table or even to a specific record or group of records within a database.

Requirement 1. It must be possible at any moment to specify the component parts of an enhanced publication.

An enhanced publication basically creates a new layer on top of existing resources. It functions as an overlay that clarifies the structure of a coherent collection of resources. Normally, the resources that are aggregated also exist as independent information units outside of the context of the compound object, which means that they can also be used in other environments. This is important, since there is rarely a strict one-to-one relationship between, for instance, an ePrint and an e-research object. An ePrint may make use of various databases, and one database may have inspired various academic texts.

To stimulate the usage of enhanced publications in academic communication, it is essential to ensure that they can be cited. For this reason, the institution that publishes an enhanced publication must make it available as a web resource and associate an identifier with it. This identifier should ideally be globally unique and persistent. In addition, it should be possible to resolve this identifier to a representation of the enhanced publication.

Requirement 2. Both the enhanced publication and its components must be available as web resources that can be referenced via URIs.

4.2 Compound objects

Kahn and Wilensky (2006) distinguish elemental and composite digital objects. Research data are often available as elemental or atomic web resources. Concretely, this means that they can be represented as a single bit stream at a single network location. In addition, it may also be the case that several atomic resources are clustered into a larger compound object. Such compound data sets may consist of multiple data files and of multiple metadata records. It must also be possible to add such compound objects to the publication. In addition, one enhanced publication may also wholly aggregate a second enhanced publication. Such a nested structure may occur in the case of e-theses that consist of parts that are also available as separate resources, such as journal articles or pre-publications. These texts may in turn aggregate other resources such as images, video

files, audio recordings, data sets or metadata records. Enhanced publications can thus be highly complex and multi-tiered objects.

Requirement 3. It must be possible to add compound digital objects to the publication.

4.3 Versioning

Enhanced publications are potentially very dynamic resources. When they contain data from research projects that are still in progress, it may be the case that resources can be added, updated or even removed on a regular basis. Such alterations potentially invalidate certain applications that were based on this enhanced publication. For this reason, it is important to ensure that agents who make use of an enhanced publication can refer to specific versions of the compound object. The versioning issue is also important on the level of individual components. Research data can be very dynamic since, for instance, data sets can grow, multimedia files can be modified and software specifications may be altered. Similarly, ePrints may also be modified heavily in the course of a project. The question what exactly qualifies as a new version must be answered by individual repository managers. The Version Identification Framework (VIF),[8] which was funded by JISC, provides useful documentation on the issue of versioning. In this framework, a version is defined as “a digital object (in whatever format) that exists in time and place and has a context within a larger body of work”. Versions can be identified by recording the date of the last modification, a version identification, or a textual description of the version.

Requirement 4. It must be possible to keep track of the different versions of both the enhanced publication as a whole, and of its constituent parts.

[8] Website: http://www.lse.ac.uk/library/vif/

4.4 Basic properties

To enable service providers to develop applications on the basis of enhanced publications, it is important to ensure that a number of key properties of the various resources in the enhanced publication can be described. In addition, to make enhanced publications interoperable, these properties should be described using a standardised and controlled vocabulary as much as possible. Chapter 6 will propose a number of vocabularies that can be used in this context.

The following attributes will be relevant in the majority of cases:

- Each component should be typed semantically to make it clear what kind of resource is being referred to.
- ePrints can have a title.
- For atomic or compound datasets, a brief description may be given. In addition, it is advisable to provide a title which makes it explicit that these particular resources are data objects.
- For enhanced publications as a whole, it will be useful to record the date of the last modification. In the case of newly created publications, this date will coincide with the date of creation.
- Since different applications are mostly needed to process or to present the various resources that are aggregated, it will be necessary to describe the technical format of the resource. Media types and media formats can be recorded using the IANA registered list of Internet Media Types. Recording this aspect is optional since the precise media type may not always be known beforehand. This may be the case for web resources which are available in different representations. Such resources can be resolved to a certain media type at the moment of request through the process of content negotiation.
- The MIME type can also be specified for metadata. However, since there are many metadata vocabularies that have not yet been acknowledged as an IANA Media Type, it will be more useful to record the namespace of the metadata schema.

Requirement 5. It must be possible to record basic properties of the publication, and of the resources that are added to it.

E-science projects are increasingly collaborative and interdisciplinary processes. To be able to trace individual contributions, it is important to ensure that it is possible to record authorships on all levels of the enhanced publication. Capturing the provenance may also help clients to establish the trustworthiness of the resource. A clear distinction must be made between the author of the enhanced publication and the authors of its component

parts. Authors of ePrints and data resources are the agents who are responsible for their intellectual contents. The author of the enhanced publication as a whole is the agent who has decided to combine the various resources into a single compound object. Evidently, authorship of the enhanced publication does not automatically imply authorship of associated ePrints or associated data sets.

Requirement 6. It must be possible to record the authorship of the enhanced publication and that of its component parts.

4.5 Long term preservation

A growing number of institutions are developing repositories that aim to preserve digital content for future generations. Notable initiatives are the e-Depot of the Dutch National Library and the NESTOR and KOPAL projects, which were initiated by the German National Library and the University and State Library of Göttingen. Such institutions mostly employ a combination of techniques to guarantee longevity, including migration, which involves transferring bits from one format or medium to another), and emulation. The latter technique involves an imitation of the functionality of a certain obsolete program or operating system to preserve the usability of a digital resource (Lorie 2001). Firstly, it must be possible to harvest the document that serialises the enhanced publication from local repositories and to ingest it into digital archiving systems. In addition, institutions responsible for the long-term accessibility and usability may choose to harvest and preserve the individual parts of the enhanced publication as well. For this reason, it must be possible to harvest representations of the web resources that are referred to in the scientific package. As was explained earlier, the structure of an enhanced publication is not always fixed. This is especially the case for publications that are created for research projects that have not yet been completed. Consequently, it may be difficult to decide for repository managers or for owners of the Long Term Archives when exactly the publication should be archived. In the case of the more dynamic publications, it is especially important to capture versioning information. Long Term Archives can leverage this information by deciding to preserve one specific version of the enhanced publication, instead of having to wait until the entire enhanced publication is complete. The key data that are recommended in section 4.4. above also help managers of Long Term Archives to generate appropriate preservation metadata for the various objects whose long term accessibility need to be ensured. Chapter 2.3. of deliverable 4.3. of Driver-II, the *Technology Watch Report*, contains a more detailed discussion of the issues that are involved in the digital preservation of enhanced publications.

Requirement 7. It must be possible to secure the long-term preservation of enhanced publications.

4.6 Relations

Whereas repositories currently focus mostly on storing individual digital objects, it is important to recognise that there are often strong links between the various resources that can be found in data repositories or institutional repositories. A crystal X-ray structure determination in a data repository, for instance, may have lead to further studies, the results of which are stored elsewhere. Resources can be related both to atomic objects and to compound objects. In the data model for enhanced publications, it is essential that these relations can be stated explicitly. Such descriptions of the links between the resources help to clarify the reasons why these resources were added to the collection. This will enable authors of enhanced publications to present these resources in a coherent framework. Relations between the various component parts need to be described and classified using a standard and generic vocabulary, as much as possible.

Requirement 8. It must be possible to record the relations between the web resources that are part of an enhanced publication.

This section provides an overview of the most common kinds of relations that can occur.

a. Containment relations

Two resources may be said to be connected through a containment relation if one unit is included physically or logically within another unit. This is a very common kind of relation, as it occurs each time a number of resources are grouped or clustered into a larger unit. Examples include e-theses that consist of different chapters, or directory structures that are published in their entirety as enhanced publications. Containment relations can always be represented visually by means of a tree diagram.

b. Sequential relations

In certain situations, it may be necessary to record the order in which resources need to be consulted. This is the case, for instance, if the separate

components of a monograph or an article are included as individual parts. The aim of sequential relations is to establish a reading path within a session.

c. Versioning information

It is often necessary to maintain different versions of a certain resource. Specific relator terms may be used to provide information on the relations between such different versions.

d. Lineage relations

Lineage relations provide information on the order in which research data are produced. Hunter (2006) explains that the lineage of data production refers to the "chain (or pipeline) of processing steps used to generate scientific data and derived products" (p. 37). When such relations are made explicit, this enables peers to trace the various stages of the scientific process.

e. Manifestations

Web resources are often available in different technical formats. For instance, an article may be available both as an HTML file and as a PDF document. Similarly, archive copies of images are often stored in TIFF or JPEG2000 format, in addition to the presentation copies in JPG or GIF format. The various manifestation may be clustered within an aggregation that brings these various formats together.

f. Bibliographic citations

Academic publications usually contain many bibliographic references to other publications. References can be both to resources that are stored in a trusted repository and to resources that are stored at other types of locations. The decision on whether a link to such external and less reliable sources is allowed is left to the discretion of individual repository managers.

Relations can be unidirectional and bidirectional. Many ontologies define only unidirectional relations. This implies that, if resource A has a relation with resource B, the inverse relation cannot be assumed automatically. If that inverse relation can exist, it mostly needs to be expressed explicitly, by using an antonym of the first term. For example, if resource A contains resource B, this can be expressed by using the term "hasPart". The inverse relation can be expressed explicitly by using the term "isPartOf". Defining such unidirectional relations may introduce a degree of redundancy, and it may certainly not always be feasible, especially when resources are distributed over many different repositories. However, when it is applied, it will have the effect that each resource carries explicit information about the relations that it is involved in. It will also ensure that web

resources can be viewed both individually and as part of an enhanced publication. This strongly increases the flexibility of the constituent parts of the enhanced publication.

4.7 Discovery

Enhanced publications must be usable and visible in largely the same systems that are used to store, index and retrieve atomic digital objects. Their contents must be accessible to services that leverage repository content such as web crawlers, citation analysis tools, harvesters and data mining applications. In addition, individual agents who may be interested in using the compound publication must be enabled to discover them. This means, more concretely, that they should have the possibility to learn about their existence. Processes of locating, retrieving and promulgating enhanced publications can be based on a wide range of techniques, including site maps, syndication or OAI-PMH. The question of which technique is the most relevant largely depends on the requirements of the discovering agent. Evidently, a technique such as OAI-PMH will only be relevant if service providers actually have the tools that are needed to perform OAI-PMH based harvests.

Requirement 9. Institutions that offer access to enhanced publications must make sure that they can be discovered.

4.8 OAI-ORE

Since it is assumed in this report that OAI-ORE will be established as the de facto standard for capturing and exchanging information on generic compound objects, institutions should take measures to ensure that enhanced publications can be available as documents based on the OAI-ORE model. In short, OAI-ORE provides a way to combine disparate web resources into a single unit. As is explained in Van de Sompel (2007), a central aim of OAI-ORE is “to develop standardized, interoperable, and machine-readable mechanisms to express compound object information on the web”.[9] It provides a framework for capturing machine-readable descriptions of compound objects through a mechanism that is known as named graphs. More specific information on OAI-ORE is provided in chapter 7. The requirement that enhanced publications must be available as documents based on the OAI-ORE model does not imply any subsequent prescriptions for the internal storage of the resource. An enhanced publication is essentially information about a collection of resources,

[9] http://www.openarchives.org/ore/documents/CompoundObjects-200705.html

and such data may be stored, for instance, in relational databases, or in other XML formats such as MPEG-DIDL or METS.

Requirement 10. Institutions that provide access to enhanced publications must ensure that these are available as documents based on the OAI-ORE model.

5 Data Model

The previous chapter has identified a number of functional and technical requirements for the storage and management of enhanced publications. This chapter will present an abstract model that can represent these requirements. This data model may form the basis for the further development of the infrastructure for compound objects within the DRIVER infrastructure.

A starting point for the content model is the observation that enhanced publications are compound digital objects which combine ePrints with one or more data resources, one or more metadata records, or any combination of these. This report will assume that an enhanced publication must minimally include one ePrint. The latter term has been defined in chapter 2. When a data object is combined with one or more metadata records, or with one or more other data objects, the resulting package is referred to as a compound dataset. Enhanced publications can thus include five types of entities:

- ePrints
- data objects
- metadata
- compound datasets
- enhanced publications

The first two types of entities are atomic objects and the latter two entities are compound in nature, which means that they aggregate other atomic objects.

Advanced discovery services may require that some key metadata is present for the various resources that are aggregated. Therefore, it must be possible to record a number of basic properties, not only for all digital objects that are included but also for the enhanced publication in its entirety. In accordance with the Kahn and Wilensky architecture for distributed digital object services (2006), a digital object is understood as “an instance of an abstract type that has two components, data and key-metadata” whereby the “key-metadata includes a handle, i.e., an identifier globally unique to the digital object” (p. 117).

The properties that are likely to be relevant in the most common situations have been listed in chapter 4. It must be noted that it is not mandatory to implement all the attributes that have been mentioned. They must be considered optional, with the possible exception of the identifier. Moreover, repository managers may also choose to include other types and properties if they are considered necessary. Figure 1 offers a first overview of the five entities and their key properties.

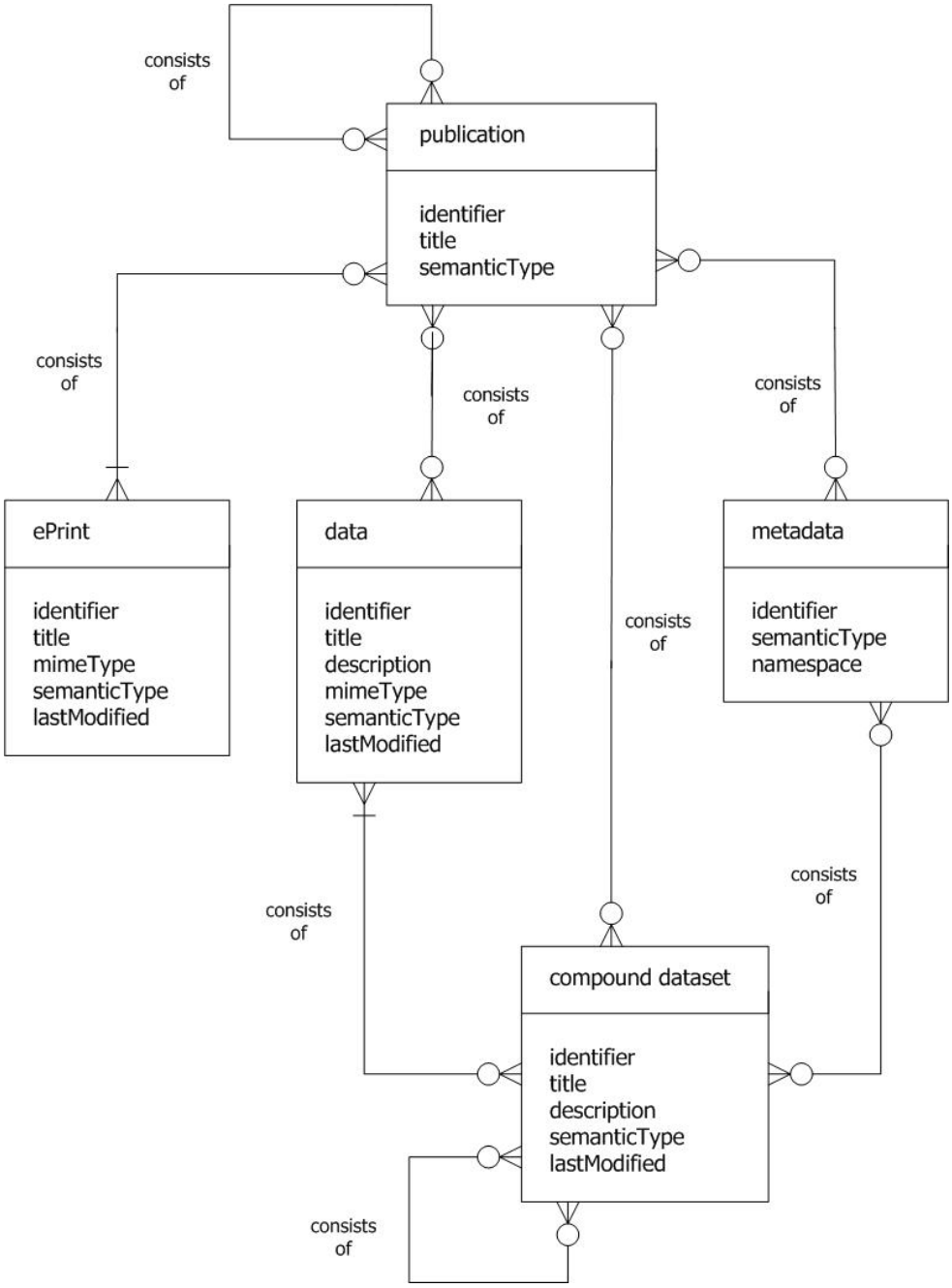

Figure 1: Entity-relation diagram for basic enties and key properties

The model that is presented in figure 1 is still incomplete, as there are three additional requirements that need to be represented. Firstly, an important requirement is that it must be possible to keep track of the different versions of both the enhanced publication as a whole, and of its constituent parts. Secondly, it must be possible to capture the provenance of the enhanced publication and of the various resources that it combines. All the entities that are listed in figure 1 can be associated with an agent who is responsible for its

existence. This agent can be an individual, an institution or perhaps a fully automated application. A third requirement that must be represented is that it must be possible to describe relations between resources. These three additional requirements apply to all the entities that are given in figure 1. To be able to visualise these additional requirements in an orderly fashion, it has been decided to extend the model with the generic notion of a digital object. This abstract object may be either atomic or compound in nature. Figure 2 visualises the observations that digital objects can be produced by zero or more authors, that they can be related to zero or more other objects, and that they can appear in one or more versions. In addition, the diagram also indicates that these digital objects can be described by metadata records, and that these objects may have been produced by three types of agents, namely individual authors, institutions, or automated applications. For the sake of clarity, all attributes have been omitted from this diagram.

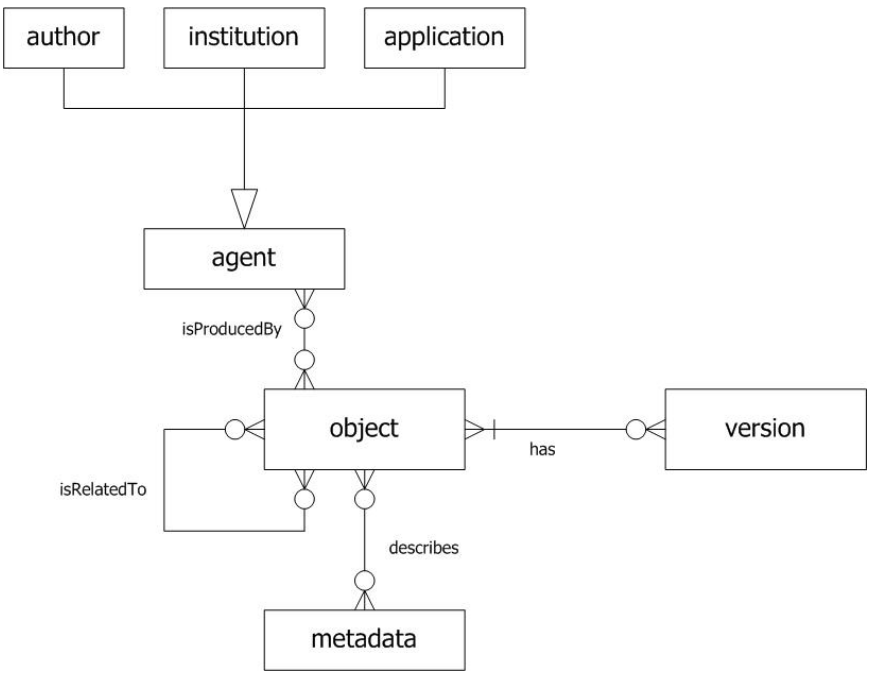

Figure 2: Entity-relation diagram for a generic digital object

To finalise the data model, it must be indicated that the abstract data object may occur in two basic types. It may be either an atomic object or a compound object. The atomic object may be included fully as a bitstream, or it may be included in the form of a URI reference. Compound objects are created by combining one or more atomic objects. This diagram also states that publications and compound datasets are specific instances of compound objects.

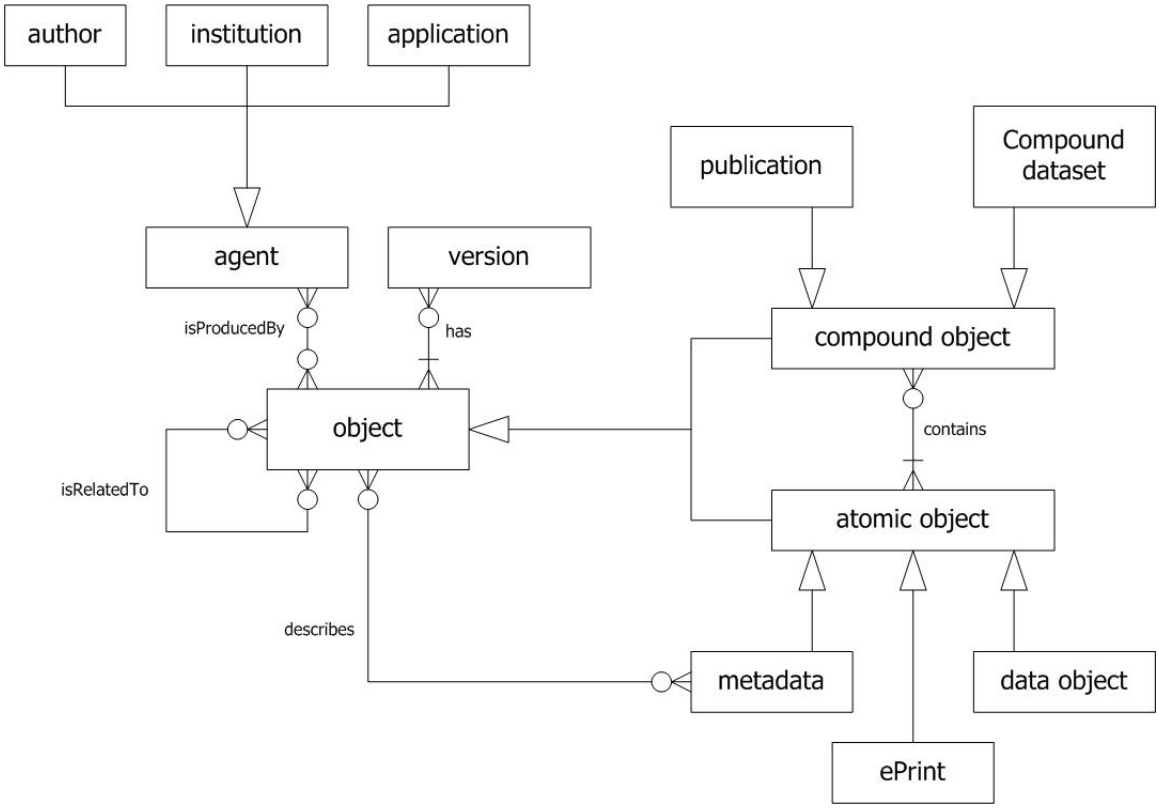

Figure 3: Full entity-relation diagram for enhanced publications

In conclusion, the model that is presented in figure 3 captures the most important requirements for the storage and management of enhanced publications. It also indicates the global manner in which enhanced publications must ultimately manifest themselves within the Driver infrastructure. Since an enhanced publication is essentially a specific instance of the more general notion 'compound object', it is important to ensure that they can actually be supported as special instances of the Compound Object data model that is proposed in D8.1.[10] The *Compound Object Model Specification* discusses the modeling abstractions that are required to describe document models of any Digital Library application domain. The Compound Object data model that has been developed also

[10] *D8.1. Compound Object Model Specification*. The model is also discussed in *Typed Compound Object Models for Digital Library Repository Systems*, Leonardo Candela, Donatella Castelli, Paolo Manghi, Marko Mikulicic, Pasquale Pagano. ISTI Technical Report 2008-TR-023.

inspires the implementation of Content Services which are capable of supporting efficient storage and search of Driver Compound Objects.

The model that is proposed in D8.1 can be applied very effectively to express the basic requirements for enhanced publications that were identified in this report. Although a full translation of the requirements of the enhanced publication model is to be provided in WP9 of Driver-II, the remainder of this chapter will give a first impression of how the two models can collaborate. D 8.1. essentially describes compound objects as sets of digital objects that are associated on the basis of relationships. The low-level model provides the minimal set of primitives required to design efficient compound object digital libraries. It supports the notions of Type, Set and Object. Types define the abstract structure and the operators of the Object entities, and they can be instantiated as Sets, which are the concrete types of the objects they contain. Sets can be of three main types: Atom Type, Structure Type and Relation Type. To construct instances of enhanced publications, it is firstly necessary to define the resources that, in this report, were referred to as 'atomic' resources. More concretely, this refers to the entities which in figure 3 are labelled 'ePrint', 'metadata object' and 'data object'. This must be followed by an instantiation of relative Sets. Once these atomic objects are defined, they can used to define Relationship types, which are "dependent types". This basically means that their creation depends on two existing, i.e. created, Sets. A second mechanism to construct dependent types are the so-called Union types, of the form "Union(A1,…Ak)". They can be used to create Sets whose objects belong to any of the Sets A1, …,Ak. The TDL expression below exemplify how a compound object can be created by combining ePrints with data objects:

```
Set compoundObjects = create Union(ePrints, dataObjects)
```

Again, these dependent types must be followed by an instantiation of the relative sets. In short, enhanced publications can be constructed by firstly defining atomic types, and by subsequently defining dependent types that combine these atomic types into larger units. The enhanced publication model thus clearly complies with the basic vision on compound digital objects that is expressed in D8.1.

6 Vocabularies

One of the requirements that were identified in chapter 4 is that it must be possible to record key metadata of digital objects. To make enhanced publications semantically interoperable, the properties that were mentioned must be described using a standardised and controlled vocabulary as much as possible. This section will propose a number of vocabularies that can be used in this context.

The DCMI Type Vocabulary provides a number of terms that may be used to describe the semantic type.

Value URI	Label
http://purl.org/dc/dcmitype/Dataset	Dataset
http://purl.org/dc/dcmitype/Event	Event
http://purl.org/dc/dcmitype/Image	Image
http://purl.org/dc/dcmitype/InteractiveResource	InteractiveResource
http://purl.org/dc/dcmitype/MovingImage	MovingImage
http://purl.org/dc/dcmitype/Software	Software
http://purl.org/dc/dcmitype/Sound	Sound
http://purl.org/dc/dcmitype/Text	Text

ePrints may be classified as such using the term 'Text'. A more specific classification of the semantic type of ePrints can be provided on the basis of the vocabulary set that has been compiled for the info:eu-repo namespace. A full overview can be found at the following address: http://info-uri.info/registry/OAIHandler?verb=GetRecord&metadataPrefix=reg&identifier=info:eu-repo/

The ePrints Application Profile[11] defines a simple vocabulary to describe access rights. It prescribes terms such as 'Open Access', 'Restricted Access' and 'Closed Access'.

Containments relations can be stated explicitly by making use of 'isPartOf' and 'hasPart' from the Dublin Core Metadata Initiative

Value URI	**Label**
http://purl.org/dc/terms/isPartOf	isPartOf
http://purl.org/dc/terms/hasPart	hasPart

The Dublin Core Metadata Initiative also provides a vocabulary that can be used to record the relation between different versions.

Value URI	**Label**
http://purl.org/dc/terms/isVersionOf	isVersionOf
http://purl.org/dc/terms/hasVersion	hasVersion
http://purl.org/dc/terms/isReplacedBy	isReplacedBy
http://purl.org/dc/terms/Replaces	Replaces

Separate digital manifestations can be connected by making use of one of the following two terms.

Value URI	**Label**

[11] Website: http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

http://purl.org/dc/terms/isFormatOf	isFormatOf
http://purl.org/dc/terms/hasFormat	hasFormat

Bibliographic references in the publication can be made explicit by using 'References' from the Dublin Core Terms Namespace, and 'Is Referenced By' for the inverse relation.

Value URI	**Label**
http://purl.org/dc/terms/references	References
http://purl.org/dc/terms/isReferencedBy	Is Referenced By

To describe lineage relations, the ABC Model that is developed by Hunter and Lagoze (2001) may be used. The model contains terms such as 'precedes', 'follows', 'contains', 'isSubEventOf', 'phaseOf', 'involves', 'usesTool', 'hasResult', 'has Action' and 'hasPresence'.

Academic work is increasingly made public under a Creative Commons licence.[12] Usage rights may be described using the dc:rights property.

Value URI	**Label**
http://purl.org/dc/elements/1.1/rights	Rights

[12] http://creativecommons.org/licenses/by-nc-sa/2.0/

7 Recommendation for the serialisation

This chapter will explain how the OAI-ORE data model can be applied to exchange information about enhanced publications. Such guidelines are needed, since the OAI-ORE effort does not limit itself to connecting publications with data sets. It offers a broad framework for compound digital objects in general. The OAI-ORE vocabulary can be used in RDF statements to specify that a collection of URI-identified resources together form a compound object. The great advantage of OAI-ORE is that it can be adopted to encapsulate distributed resources. OAI-ORE focuses on resources, and not so much on repositories.

The OAI-ORE data model distinguishes three entities. The 'Aggregation', firstly, is a collection of web resources. Individually, these are referred to as 'Aggregated Resources'. A 'Resource Map' is an entity that contains a description of an 'Aggregation'. In addition, there are five properties that can relate these entities. The connection between the Resource Map and the Aggregation can be established using 'describes' and its inverse relation, 'isDescribedBy'. An aggregated resource becomes part of an enumeration of resources if the resource map asserts the 'aggregates' relation between this resource and the aggregation. The link from an Aggregated Resource to the Aggregation can be expressed using 'isAggregatedBy'. Lastly, there is also the property 'similarTo' to denote the fact that two resources are identical. Figure 4 visualises the kernel of the OAI-ORE abstract data model.

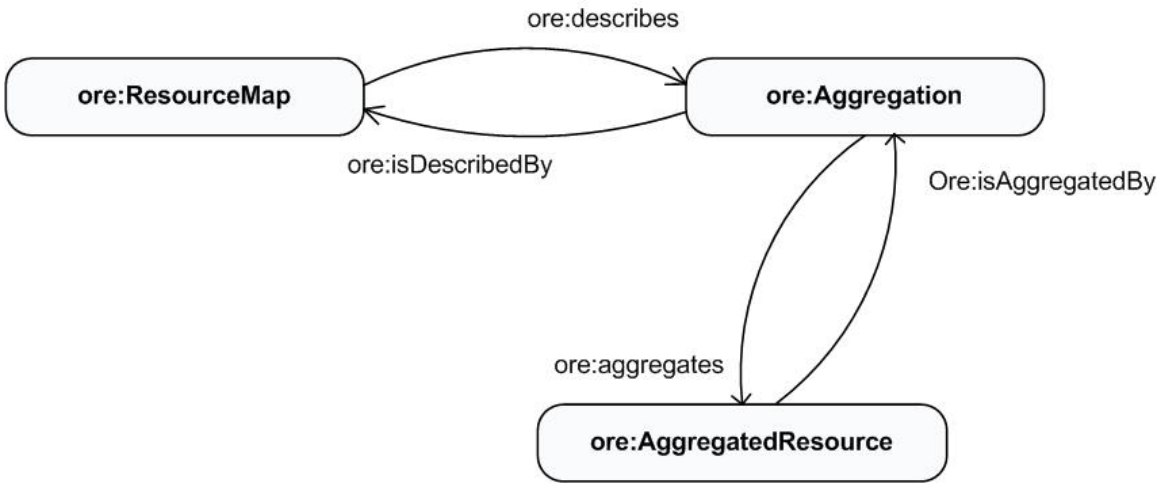

Figure 4: Basics of the OAI-ORE model

In chapter 3 of this report, it was explained that an enhanced publication is headed by an ePrint, and that enhancements can be added to this text in the form of either research data, metadata records, compound datasets, or other enhanced publications. The data model for enhanced publications in the report thus bears a strong similarity to the OAI-ORE abstract data model. The Publication entity may be mapped to the Aggregation and the enhancements correspond to the Aggregated Resources. The Resource Map is the document through which the enhanced publication may be accessed. Lagoze and Van de Sompel (2007) explain that Resource Maps are used "to expose to harvesting clients the compound objects that they provide access to", amongst other things. The Resource Map thus references the full enhanced publication. The Aggregation has an identifier that is derived from the URI of the Resource Map. A Resource Map always describes only one Aggregation, but an Aggregation may be described by more than one Resource Map. In addition, one Aggregated Resource can also be part of more than one Aggregation.

To be able to express the key metadata for the various resources, the attributes that were mentioned in Chapter 4 must be mapped to terms that can be used as properties in RDF statements. Following the OAI-ORE documentation, this report shall make use of vocabulary from the Dublin Core Metadata Element Set, the DCMI Metadata Terms and the Resource Description Framework. The tables below contain a recommended concordance.

Attributes	**Property**
identifier	dc:identifier
title	dc:title
description	dc:description
author	dc:creator
semanticType	rdf:type
versionIdentification	dc:hasVersion
versionDate	dcterms:modified

versionDescription	dc:description
mimeType	dc:format
namespace	dcterms:conformsTo
lastModified	dcterms:modified

The guidelines that have been presented so far shall be illustrated using two examples. The first sample enhanced publication is taken from the Network of European Economists Online (NEEO) project, which is coordinated at the University of Tilburg. The bibliographic details of this publication have been adapted slightly for the sake of clarity. It is a compound object that consists of an ePrint, a metadata record in MODS, and a dataset. Figure 5 indicates how information about such a constellation can be described using the OAI-ORE vocabulary.

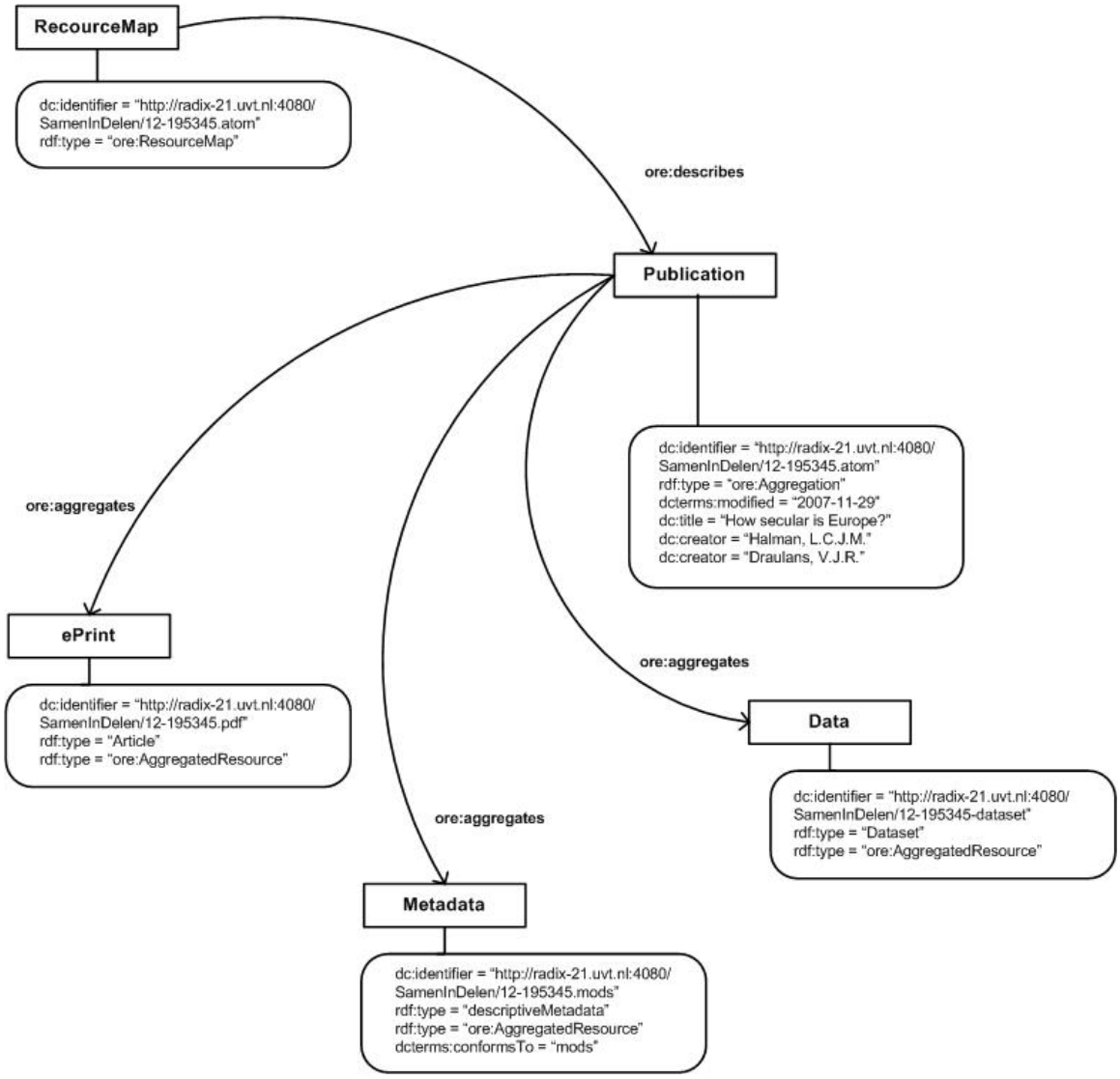

Figure 5: Serialisation of an enhanced publication (Example 1)

Listing 1 below illustrates how this information can be expressed using RDF/XML.

Listing 1.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
```

```
xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:ore="http://www.openarchives.org/ore/terms/"
    xmlns:xhtml="http://www.w3.org/1999/xhtml">
    <rdf:Description rdf:about="http://radix-
21.uvt.nl:4080/SamenInDelen/listOfItems.atom">
        <rdf:type
rdf:resource="http://www.openarchives.org/ore/terms/ResourceMap"/>
        <ore:describes rdf:resource="http://radix-
21.uvt.nl:4080/SamenInDelen/12-195345"/>
    </rdf:Description>
    <rdf:Description rdf:about="http://radix-21.uvt.nl:4080/SamenInDelen/12-
195345">
        <rdf:type
rdf:resource="http://www.openarchives.org/ore/terms/Aggregation"/>
        <dcterms:modified>2007-11-29T09:40:01Z</dcterms:modified>
        <ore:aggregates rdf:resource="http://radix-
21.uvt.nl:4080/SamenInDelen/12-195345.mods"/>
        <ore:aggregates rdf:resource="http://radix-
21.uvt.nl:4080/SamenInDelen/12-195345-dataset"/>
    </rdf:Description>
    <rdf:Description rdf:about="http://drcwww.uvt.nl/~place/SamenInDelen/12-
195345-eft.doc">
        <rdf:type
rdf:resource="http://www.openarchives.org/ore/terms/AggregatedResource"/>
        <rdf:type rdf:resource="info:eu-repo/semantics/article"/>
        <dc:title>How secular is Europe?</dc:title>
        <dc:creator>Halman, L.C.J.M.</dc:creator>
        <dc:creator>Draulans, V.J.R.</dc:creator>
    </rdf:Description>
    <rdf:Description rdf:about="http://radix-21.uvt.nl:4080/SamenInDelen/12-
195345.mods">
        <rdf:type
rdf:resource="http://www.openarchives.org/ore/terms/AggregatedResource"/>
        <rdf:type rdf:resource="info:eu-repo/semantics/descriptiveMetadata"/>
        <dcterms:conformsTo>mods</dcterms:conformsTo>
    </rdf:Description>
    <rdf:Description rdf:about="http://radix-21.uvt.nl:4080/SamenInDelen/12-
195345-dataset">
        <rdf:type
rdf:resource="http://www.openarchives.org/ore/terms/AggregatedResource"/>
        <rdf:type rdf:resource="http://purl.org/dc/dcmitype/Dataset"/>
    </rdf:Description>
</rdf:RDF>
```

The OAI-ORE documentation contains guidelines for serialisations of the model in RDF/XML and in ATOM. A serialisation in the latter schema has the advantage that, in the case of research projects that are still in progress, clients can be notified of new additions through a feed reader. Details of the serialisations in ATOM and RDF/XML can be found in Van de Sompel et al. (2008).

In more advanced compound objects, it may be the case that one of the aggregated resources is another enhanced publication which is published separately. Such a situation can occur when a publication is a collection of texts written by various authors. A layered or nested structure may also be necessary in the case of large e-theses. When sizeable quantities of digital data are accumulated for different chapters, it can be effective to disaggregate the complete e-theses into smaller units. Figure 6 below depicts an example in which one chapter of an e-thesis is available as a separate compound object. In that situation, aggregated chapters can be enhanced with specific metadata records. The e-thesis aggregates its first chapter by pointing to the URI of the chapter 1 Aggregation. When this URI is dereferenced, it should produce the Resource Map of the first chapter.

A question of debate is whether or not the Resource Map should explicitly specify all the manifestations that are available for digital objects. An ePrint, for instance, may be available as a PDF file, a DOC file, a simple text file or an HTML page. If such manifestations all need to be mentioned in the resource map, it will be useful to cluster them in an aggregation. In FRBR terminology, this aggregation may be said to function as an “Expression” which can be embodied in different “Manifestations”, which may each appear as Aggregated Resources within this aggregation. This solution has the advantage that it is made very clear which digital objects (bitstreams) actually exist for the resource. This way, users can intentionally select one particular manifestation. An alternative solution would be to leverage the Web Architecture, and to simply incorporate the URI of the expression as an Aggregated Resource. The various manifestations can then be reached through the process of Content Negotiation, at the moment when the URI is dereferenced. Such a solution would make the Resource Map more economic. A disadvantage, however, is that it may not always be predictable for clients which kind of manifestation they will actually receive, since this latter question will be determined by comparing the objects that are available in the repository with the objects that the client can accept. Figure 6 illustrates the first method of dealing with manifestations.

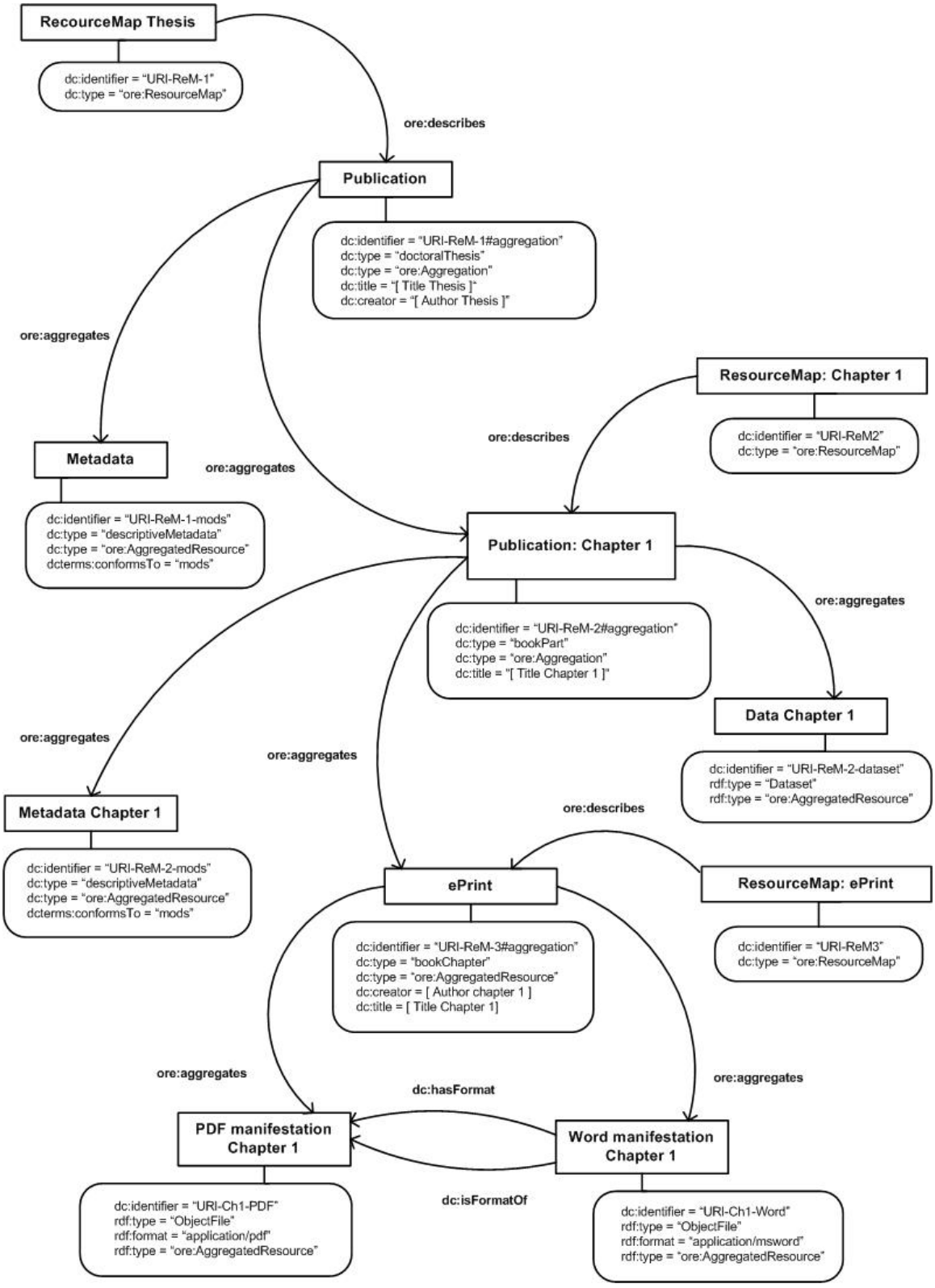

Figure 6: Serialisation of an enhanced publication (Example 2)

8 Conclusion

The amount of digitally available research data is growing continuously. Unfortunately, when scientific resources are made available on-line, they are not always published with reliable and consistent metadata. This complicates the retrieval of these research data, which in turn poses a serious threat to an effective reuse of digital resources. A promising approach to improving the access to research data is through the enhancement of the traditional publication. This entails, more concretely, that publications are enriched with references to the primary data that were used to produce the insights that are put forward in this text. Scientific publications and research data are currently shared and disseminated through largely the same channels, and publishing these resources in conjunction is now more and more a realistic option. Through enhanced publications, researchers can be forwarded effectively to relevant sources, such as underlying data collections, models and algorithms. Such resources can normally not be incorporated fully in the actual publication, but the presence of pointers to the primary data, which are stored separately, should contribute greatly to an improved accessibility of this information. In a sense, the scientific publication will function as metadata for the research data.

An improved visibility of relevant research data benefits the efficiency of scientific processes in a number of ways. Through the use of enhanced publications, scientists can leverage current scientific results to generate future discoveries more speedily and more efficiently. Enhanced publications can also improve the quality of peer review methods, since access to the primary sources enable peers to replicate experiments and to verify the claims that are made in the publication. Importantly, enhancing the traditional publication must improve the visibility and, consequently, the impact of academic studies. With its emphasis on object reuse, the combination of distributed resources and a focus on the lineage of research data, enhanced publications must help repository managers to accommodate the demands of academic communication in the 21st century.

Literature

- *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (Oct. 2003)*. Accessed April 2008. <http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html>
- Borgman, Christine (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Michigan : MIT Press.2007
- *Budapest Open Access Initiative (2001-2004).* Accessed April 2008. http://www.soros.org/openaccess/
- Candela, Leonardo et al. (2008). *Typed Compound Objects Models for Digital Library Repository Systems*. ISTI Technical Report 2008-TR-023
- Cheung, Kwok et al., "SCOPE – A Scientific Compoun Object Publishing and Editing System". T*he International Journal of Digital Curation*, 2 (3), pp. 1-12.
- Fink, J. L., & Bourne, P. E. (2007). "Reinventing Scholarly Communication for the Electronic Age". *CTWatch Quarterly, 3* (3), pp. 26-31 <http://www.ctwatch.org/quarterly/articles/2007/08/reinventing-scholarly-communication-for-the-electronic-age/>
- Foulonneau, Muriel and Francis André (2008). *Investigative Study of Standards for Digital Repositories and Related Services*, Amsterdam: Amsterdam University Press.
- Hey, Tony and and Anne Trefethen (2003). "The Data Deluge: An e-Science Perspective". In *Grid Computing - Making the Global Infrastructure a Reality* (pp. 809-824). Wiley and Sons.
- Hunter, Jane (2006). "Scientific Publication Packages. A Selective Approach to the Communication and Archival of Scientific Output". *The International Journal of Digital Curation*, 1 (1), pp. 33-52.
- Jeremy, J.C., et al. (2005), "Named graphs, provenance and trust". In *Proceedings of the 14th international conference on World Wide Web*. ACM Press: Chiba, Japan.
- Kahn, Robert and Robert Wilensky (2006). "A framework for distributed digital object services". *International Journal on Digital Libraries*, 6 (2), pp. 115-123.
- Kircsz Joost G. (1998), "Modularity: the next Form of Scientific Information Presentation?", *Journal of Documentation*, 54 (2), pp. 210-235. <http://dx.doi.org/10.1108/EUM0000000007185>
- Lagoze, Carl and Jane Hunter (2001), "The ABC Ontology and Model", *Journal of Digital Information*, 2 (2), pp. 160-176.
- Lagoze, Carl and Herbert van de Sompel, "Compound Information Objects: The OAI-ORE Perspective". May 28, 2007. Accessed July 2008. http://www.openarchives.org/ore/documents/CompoundObjects-200705.html
- Lorie, Raymond A. (2001), "Long Term Preservation of Digital Information". In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '01), pp. 346-352.
- Lynch, Clifford A. (2003), "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age". *ARL Bimonthly Report*, Vol. 226

- Murray-Rust, P. and H.S. Rzepa (2004), “The Next Big Thing: From Hypermedia to Datuments”. *Journal of Digital Information*, 5 (1).
- *ORE User Guide – Primer*. 11 July 2008 . Accessed July 2008. <http://www.openarchives.org/ore/0.9/primer>
- *ORE Specification - Abstract Data Model*. 2 June 2008. Accessed July 2008. <http://www.openarchives.org/ore/0.9/datamodel>
- *ORE Specification – Vocabulary*. 2 June 2008. Accessed July 2008. <http://www.openarchives.org/ore/0.9/vocabulary>
- *ORE Specification - Representing Resource Maps Using the Atom Syndication Format*. 2 June 2008. Accessed July 2008. <http://www.openarchives.org/ore/0.9/atom>
- *ORE User Guide - Resource Map Implementation in RDF/XML*. 2 June 2008. <http://www.openarchives.org/ore/0.9/rdfxml>
- Rumsey, Sally and Frances Shipsey (2006). “Scoping Study on Repository Version Identification (RIVER), Final Report”. London: Rightscom Ltd.
- Van de Sompel, Herbert et al. (2004), “Resource Harvesting within the OAI-PMH Framework”, *D-Lib Magazine*, 10 (12). <http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html>
- Van de Sompel, Herbert et al. (2004). “Rething Scholarly Communication: Building the Systems that Scholars Deserve”, *D-Lib Magazine*, 10 (9). <http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html>
- Van der Poel, K. G. (2007). *Verkenning van de interesse van wetenschappelijke onderzoekers in WP1-Verrijkte Publicaties en WP2-Collaboratories*. Utrecht: Surf. <http://www.surffoundation.nl/download/20070524_Rapport_VerkenningWP12_vdPoel_def.pdf>
- Van Horik, R. (2008). “Data curation”. In K. Weenink, L. Waaijers & K. van Godtsenhoven (eds.), *A DRIVER’s Guide to European Repositories* (pp. 137-138). Amsterdam: Amsterdam University Press.
- Seringhaus, M. R. and Gerstein, M.B. (2007). “Publishing Perishing? Towards tomorrow’s information architecture”. *BMC Bioinformatics*, 8 (17).
- Sure, York et al. (2005), “The SWRC Ontology – Semantic Web for Research Communities”, In: Carlos Bento et al. [eds.], *Proceedings of the 12th Portuguese Conference on Artificial Intelligence - Progress in Artificial Intelligence* (pp. 218 – 231). Covilha: Springer.
- Willinsky, John (2005). *The Access Principle*. Michigan: MIT.