Enhanced publications, state of the art

Woutersen-Windhouwer, S.; Brandsma, R.

Publication date
2009

Document Version
Final published version

Published in
Enhanced publications: linking publications and research data in digital repositories

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
Download date:16 Aug 2021
PART 1. Enhanced Publications, State of the Art

Saskia Woutersen-Windhouver & Renze Brandsma

The authors would like to thank Thomas Place (Universiteit van Tilburg), Anneloes Degenaar and Henk Ellerman (Rijksuniversiteit Groningen), Barbara Sierman and Paul Doorenbosch (Koninklijke Bibliotheek), Rosemary Russell (UKOLN), Peter Verhaar (Universiteit Leiden), Marc van den Berg (Universiteit van Amsterdam) and Maurice Vanderfeesten (SURFfoundation) for carefully reviewing our manuscript and for providing us with their comments and suggestions. We especially would like to thank Rosemary Russell (UKOLN) for correcting the English in our manuscript.
1. Introduction

In the digital world of scholarly publishing online access is provided to articles, hyperlinked reference and supplementary data. Connection with social networking, e.g. blogs, relation with other materials, e.g. multimedia, and semantic context, e.g. XML, is not realised widely at present. Publications and related objects are processed separately as single objects and connections between them are not easy to find. As no relation between single objects is provided, it is difficult to find out whether related objects are available.

Meanwhile, the number of scholarly objects on the Internet is growing very quickly. Integration of all this scientific information by linkage is necessary to keep publishing efficient and to maintain control over the process. Therefore, publications should provide those links, resulting in ‘Enhanced Publications’.

Seringhaus and Gerstein (2007) have proposed an information infrastructure for Enhanced Publications, it should:

- capture a broad range of data in digital format and facilitate database deposit alongside manuscript publication;
- index all full-text journal articles, associate keywords and identifiers with database records and link textbooks, laboratory websites and high-level commentary;
- provide multiple levels of peer review, community comment and annotation;
- make articles fully machine-readable providing intelligent markup and structured digital abstracts.

Such a network of information should be accessible through a single seamless portal.

The infrastructure of information is becoming a more prominent feature now and the focus of attention is shifting from the publication as a whole to the structure and pattern of linkage between elements. In an Enhanced Publication this link pattern should support and reflect the relation between the publication and all relevant objects like data, Web sites, commentaries. This pattern can become quite complex and goes beyond what can be captured in the linear or sequential structure of the traditional publication model.
This leads to the question whether the current infrastructure of repositories, based on OAI-PMH, XML, Dublin Core, DIDL, or DDI, is capable of handling the more complex patterns associated with Enhanced Publications. If not, how will this affect the design of a new infrastructure? Will it be necessary to set new rules for Enhanced Publications?

To answer these questions, an overview is given of the current status of Enhanced Publications. Models are reviewed that reflect and support relations between objects in the world of scholarly publishing. Finally, a short checklist for repositories with Enhanced Publications is provided.

More specifically the following items will be discussed:
- Structural elements of an Enhanced Publication
- Publication models
- Characteristic features of objects
- Relations
- Current repository projects
- Conclusions
2. Structural Elements of an Enhanced Publication

Summary
With a view to preservation the publisher of an Enhanced Publication should set up rules about the quality of the file, file types, and links which may be used. Both publishers and researchers should invest in enhancing publications. Software and several tools for publishers and repositories are available for adding comments to online publications. Some publishers enhance articles by tagging the data in the article. For tagging they use many standards and ontologies, which are all discipline-specific. It will be difficult to provide these services on a more general level.

2.1 Elements provided by the Author

According to Van de Sompel and Lagoze (2007) structural components of a publication are:
- Semantic type, e.g. article, simulation, video, data set, software;
- Media type, e.g. text, image, audio, video, mixed;
- Media format, e.g. PDF, XML, MP3;
- Network location, because different components are accessible by different repositories.

Electronic annexes with the article are a common option, e.g. Springer, Elsevier, and Blackwell. Typical examples are: data sets, simulations, software, annotations, movies, computer-enhanced images, audio clips, tools and databases. The content may as well be the outcome of research as part of the research process itself, e.g. surveys. Annexes may be referred to in the article itself, but this is not mandatory.
A few examples of supplementary materials can be viewed at the following urls:

**Excel (data)**
http://dx.doi.org/10.1111/j.1365-2117.2007.00326.x (record)
http://www3.interscience.wiley.com/cgi-bin/fulltext/118541732/sm001.xls (data file)

**PDF (data)**
http://dx.doi.org/10.1111/j.1365-2958.2006.05221.x (record)
http://www3.interscience.wiley.com/cgi-bin/fulltext/118630076/sm001.pdf (data file)

**Movies**
http://dpc.uba.uva.nl/ctz/vol71/nr01/art02

**Screenshot of several results**
http://www.elsevier.com/wps/find/authorsview.authors/multimedia supplementary

---

*Figure 1. Example of supplementary materials*

‘Instructions for authors’ for supplemental materials as produced by some traditional publishers give some idea of what this means in practice. Publishers’ checklists usually require that:

- All supplementary material is produced with a legend stating what it is, what format it is, and where necessary how it was created;
- Colour images are provided in the RGB colour space;
- Physical dimensions of the artwork match the dimensions of the journal;
- Lettering used in the artwork does not vary too much in size;
- Recommended file naming conventions are used;
- If a native data set is supplied, the program and/or equipment used are mentioned; or with specialist software like Latex, the software and version number;
- Extra rules apply for delivery on disk.

Most publishers supply instructions for multimedia files with details on the file types to be used. The instruction pages show how to prepare supplementary materials for electronic submission and include information on common problems, suggestions on how to ensure the best results and image creation guides for popular applications. They also offer special guidelines for Windows and Macintosh platforms.

---

3 We found some supplementary material that could have been published as raw data to be converted to PDF. This makes reuse of the data unnecessarily complicated.


24
Accepted formats are TIFF, EPS or PDF, but publishers are also aware that a number of authors already submit their artwork in MS Office formats for convenience. They accept MS Office files Word, PowerPoint, and Excel, provided that they meet certain conditions. Elsevier guarantees that they will continue to support these submission types, now and in the future.

On media types, Van de Sompel writes that a unit of communication should not discriminate between media types, and that it should recognise the compound nature of what is being communicated. It should be possible to allow multiple heterogeneous data streams as a single communication unit; as well as to recognise references to previously communicated units as formal components of a new unit (Van de Sompel et al., 2004).

Links to all types of objects by an author on the Internet, like movies and data sets, can be easily made in scholarly publications. It is possible to link to all kinds of materials on institutional Web sites or in repositories, and to blogs, independent of the location. Stable URL’s are not required and not checked. At this moment links to an article on a publisher’s site can only be added when the publication is created: not after official publication.

2.2 Elements provided by Peer Review

Researchers have increasingly the opportunity to add information to a published article. The most famous example can be found at the Public Library of Science (Figure 1). PLoS ONE is an international, peer-reviewed, open-access, online journal. The pre-publication peer review of PLoS One concentrates on technical rather than subjective concerns. It may involve discussion with other members of the editorial board and/or the solicitation of formal reports from independent referees. If published, papers will be made available for community-based open peer review involving the addition of online notes, comments, and ratings. This requires that they offer researchers an easy way of adding annotations, starting a discussion on an article, adding trackbacks, e.g. to blogs (Figure 2) and ratings. Other PLoS journals use or will use the same model although they use higher, not only technical, peer review standards. PLoS uses the Topaz software platform for the infrastructure of their journals.
Below some examples of other journals with open peer review are described.
Atmospheric Chemistry and Physics (ACP)
ACP is an interactive scientific journal with open peer review. It facilitates open peer review to show how other researchers have improved or enhanced a publication by their comments. Initial access peer review of ACP assures the basic scientific and technical quality of papers published. Subsequent interactive discussion and public commenting by referees, authors and other members of the scientific community is expected to enhance quality control for papers published in ACP beyond the limits of the traditional closed peer review.

Journal of Interactive Media in Education (JIME)
The open peer review process model of JIME shows three stages: preprint under private open peer review, preprint under public open peer review, and publication. This open model results in a more responsive and dynamic review. Reviewers are named and accountable for their comment. Their contribution is acknowledged. Thus, the research community has the opportunity of shaping a submission before final publication. Beyond this stage, discussions remain online, but cannot easily be referenced lacking a unique identifier or link. However, name of the author, email address, date and time are mentioned.

Figure 4. The names of reviewers are stated with the publication

Economics
Economics adopts “an open source approach to publication, viewing research as a cooperative enterprise between authors, editors, referees and readers”. This approach is considered to be:

- **Quick.** As it uses a public peer review process, publication lag is radically reduced;
- **Democratic.** The quality of an article is decided not just by the editors and referees, but also by the community of registered readers;
• *Up-to-date.* Authors can upload revised versions of their publication in response to public peer review.

Registered readers may rate, recommend and comment and have the option to do so anonymously. Sometimes readers are invited to comment. Comments cannot be easily referenced however. They may have a name and have a date and time, but they are part of the publication.

![Image of a diagram](image.png)

**Figure 5.** The whole process of commenting is registered, but cannot be cited easily

At the moment, none of the large publishers Elsevier, Springer, or Blackwell offer such post publication peer review. Nature only allows comments to news, not to publications. Some repositories provide the same opportunities for open peer review. In the UK, Richtags\(^5\) facilitates cross-repository EPrints browsing and has an option to discuss and tag articles in the repository. The CERN Document Server\(^6\) offers researchers the opportunity to write a review or a comment, and to rate a document. The National Science Digital Library (NSDL) allows insertion of a trackback in the repository when a link is created in a wiki or on a blog to an item in the repository.

\(^5\) [http://www.richtags.org](http://www.richtags.org)

\(^6\) [http://cdsweb.cern.ch/](http://cdsweb.cern.ch/)
2.3 Elements provided by the Publisher

The most important role for a publisher is thought to be in enhancing articles, using XML and, in particular, a markup language for some specific areas of knowledge, e.g. Chemical Markup Language, Mathematics Markup Language and Biology Markup Language. Markup language assists in finding information and linking it to external sources. To enhance information in an article all relevant terms should be tagged. Relevant questions are: which are those terms, what standards are to be applied and who is responsible for the actual markup (Lynch, 2007).

The Royal Society of Chemistry has started an interesting project Prospect in which publications are enhanced.

The following standards are used:

- InChI. The International Chemical Identifier;
- CML. Chemical Markup Language;
- GO Consortium. Controlled vocabulary to describe gene and gene product attributes in any organism;
- Sequence Ontology;  
- Open Biomedical Ontologies;  
- IUPAC Gold Book. IUPAC Compendium of Chemical Terminology;  
- SMILES. Simplified Molecular Input Line Entry Specification.

The publisher offers an Ontology Terms feature. If the article contains terms from the Gene Ontology, Sequence Ontology or Cell Ontology, drop-down boxes appear. Selecting a term opens a pop-up window, presenting further details and related articles.

---

7 http://www.rsc.org/Publishing/Journals/ProjectProspect/ The Prospect View availability will be free or subscriber-only depending on the status of the other article formats (PDF and unenhanced HTML). The Examples page contains several examples that are freely available to all users. See also: http://www.rsc.org/Publishing/Journals/ProjectProspect/Features.asp  
http://www.rsc.org/images/prospect11divx_tcm18-84747.avi  
8 http://old.iupac.org/inchi/  
9 http://www.geneontology.org/  
10 http://www.sequenceontology.org/  
11 http://www.obofoundry.org/  
12 http://goldbook.iupac.org/
Another feature is Highlight Terms. It offers three options, each of which will highlight different terms in the paper:

- **Show Gold Book.** Highlights in yellow the terms contained in the IUPAC Gold Book. Clicking on a highlighted term opens a pop-up window that links directly to the online IUPAC Gold Book.

- **Show Ontology Terms.** Highlights in blue terms from the three ontologies that may appear in the paper, the same terms that appear in the drop-down boxes, so the readers can judge the context. Clicking on a highlighted ontology term opens a new pop-up window providing a definition of the term, a link to online ontology and a listing of other related RSC papers that are enhanced and also contain that term.

- **Show Compounds.** Highlights in pink the compounds that are identified in the paper. Clicking on it will open a new pop-up window containing compound name, any synonyms, the InChI (IUPAC's Chemical Identifier), SMILES strings for the compound, a downloadable CML (Chemical Markup Language) file, and a 2-D graphic of the compound. A list of other RSC enhanced articles containing that compound is also available.

This enhanced functionality enables readers to judge the content of the article quickly, bringing up definitions and lists of related articles. The compound information identified allows for viewing and downloading standard information on each compound and provides links to other articles containing this compound. Downloaded CML may be viewed in the JMOL applet.

Some publishers use numbers or IDs from other databases, also external ones. Only a few of them have established active links, e.g. Nature linking to Genebank and the Protein Data Bank. Fink and Bourne are working with the Public Library of Science and a major biological database, the RCSB Protein Data Bank, to integrate the traditional concept of a publication and a separate data repository. They demonstrate that many publishers require the author to deposit the data of the publication in an appropriate public repository. On publishing a manuscript describing the structure of the macromolecule Macromolecular structure, data must be deposited in the Protein Data Bank. A reference to the publication is included as part of the deposition process. Unfortunately this requirement does not apply to further publications on the protein. They will link to the Protein Data Bank, but the bank itself does not contain information on these articles.
3. Publication Models

Summary
After describing current models, we propose as a definition of an Enhanced Publication: a publication that is enhanced with research data, extra materials, post publication data, database records, e.g. the Protein Data Bank, and that has an object-based structure with explicit links between the objects. In this definition an object can be part of an article, a data set, an image, a movie, a comment, a module or a link to information in a database.

In every model it is tried to add a structure by using metadata on very different levels: module (Kircz), classes and instances (Van de Sompel) and concepts (Hunter, Marcondes). The aggregated object of all these models becomes more richly structured and exposes its semantics more explicitly. This structure is not only important for human reading and comprehension, but should also be machine-readable for data mining purposes. One of the difficulties is that adding metadata and/or rich structure is very time consuming. Almost all researchers mention the importance of creating tools to assist the author. Moreover (Lynch, 2007) mentions the importance of tools: “I hope that we will see a new generation of viewing and annotation tools deployed, presumably working on semantically rich XML document representations.”

It appears that most metadata formats, like Dublin Core, MARC, and METS, are too limited for Enhanced Publications. The information that needs to be recorded cannot be mapped easily to a hierarchical XML structure. We are dealing with objects, properties and types of objects, and relationships between objects. To describe this intricate set of properties, we really need a graph structure instead of a tree structure. For this new infrastructure we will need to use the OAI-ORE model.

3.1 Introduction

Van der Poel (2007) has interviewed scientist about Enhanced Publications. He found three ways of enhancement:
- Research data, evidence of the research;
- Extra materials, to illustrate or to clarify;
- Post publication data, e.g. commentaries, ranking.
The practice of traditional publishers corresponds with the researchers’ view of Enhanced Publications: short articles in print, and much additional material as supplements in online form only. These supplementary materials can be a valuable addition or they can make for a disjointed piece of work. In practice most of the supplementary material is ad hoc and cannot be readily queried across all articles (Bourne, 2005).

For the DRIVER II project we have searched for an alternative publication model that might help to structure all these objects into a logical unit, an Enhanced Publication. We have limited ourselves to Enhanced Publications, a “combined package of a textual resource and additional materials such as original data”. It may be noted that publications that have no textual resource, e.g. data sets or movies, or movies demonstrating a protocol, are still rare, but are becoming increasingly important.

In the next two sections we will look at new publication models that have been developed to improve and enhance scientific articles:

- The modular article described by Kircz, and
- The semantic model by Hunter (Scientific Publication Packages) and Marcondes (scientific methodology structure in XML).

### 3.2 Modular Article

According to Kircz (1998; 2002) data sets, images, sounds, simulations and videos are part of the publishing environment, next to text. However, these objects should preferably not be simply added to the traditional article, which is described as linear. The better choice is linking them together into a modular article. This development into an article as an aggregation of independent but interacting objects or modules is made possible by Internet technology.

A modular article consists of modules and links between them into a coherent unit for the purpose of communication. A module is defined as a uniquely characterised, self-contained representation of a conceptual information unit, aimed at communicating that information (Harmsze, 2000). A module shows different forms, e.g. a data set in its basic form as a list of data, or conceived in a histogram. Each type of information unit should be well defined and therefore be endowed with different
sets of metadata, each set describing a different aspect of the information entity (Kircz, 1998; 2002).

A modular structure allows for reuse of relevant modules that do not need to be rewritten for a new purpose. Well-informed readers may skip rereading them, depending on their wish to do so, their expertise or level of understanding. The modular structure brings more efficiency to both reading and publishing. The journal Cell uses a modular structure, be it in a somewhat primitive form. Older issues are freely accessible.

Proposed modules of a modular article are (Kircz, 1998):

1. Module meta information; the central module of an article
   a. Bibliographic information;
   b. Contents, revealing the structure of the article;
   c. Index terms, according to the relevant applicable standards of classification;
   d. Bibliographical references;
   e. Acknowledgments;
   f. Abstract.

2. Goal and setting
   a. The definition of the problem;
   b. The embedding of the research (methods, techniques, goals).

3. Results
   a. Raw data;
   b. Fitted data.

4. Discussion

5. Conclusions

3.3 Semantic Publication

Hunter (2006) introduces the ‘Scientific Publication Package’ (SPP) as a new information format that encapsulates raw data, derived products, algorithms, software, textual publications and associated contextual and provenance metadata. This new information format is fundamentally different from the traditional file-based formats. Hunter describes a
high-level architecture that is currently under development. It enables scientists to capture, index, store, share, exchange, reuse, compare and integrate scientific results through SPPs.

The SPPs are based on an extension of the ABC model. The ABC model and the SPP are based on a number of scientific concept models for publishing scientific data and results, and for documenting the lineage of scientific theories and advances.

Hunter stresses the importance of workflow technologies as a component of the scientific process. SPPs capture the chain of processing steps used to generate scientific data and derived products, like publications. They also enable scientists to describe and carry out their experimental processes in a repeatable, verifiable and distributed way and to track the source of errors, anomalies and faulty processing (Van Horik, 2008).

At the end of the scientific discovery process, the scientist publishes his/her SPP, a complex, composite digital object encapsulating a variety of related heterogeneous components. These must be specified and can either be included as references to a unique identifier or actual bit streams incorporated within the package. Tools are provided to the scientist that allow him/her to specify the precise components, including:

- Data: database values, images, visualisations, graphs;
- Mathematical functions represented in MathML: input variables, output variables, constants, constraints;
- Software specifications: source code, executables, applets or links to Web services;
- Textual documents: EndNote files, notes, reports, documentation, annotations, publications.

The Scientific Publication Package (SPP) is then generated. It is a compound digital object represented as an RDF package. The relationships between the atomic objects within the compound object are either explicitly defined during the metadata capture, inferred from the rules associated with the ontology, or defined by the scientist during the SPP specification. Descriptive metadata for the SPP is entered and validated.
It is envisaged that this metadata set could be based on the extensible CCLRC scientific metadata model:

- Identifier;
- Title;
- Research focus/topic;
- Study;
- Model type, drawn from a hierarchical thesaurus;
- Creator/investigator – name and contact details, organisation;
- Date created;
- Date published.

The creator/author attaches a Science Commons License to the SPP. The SPP object now may be ingested and saved to a DSpace or Fedora digital library/institutional repository (Hunter, 2006).

Marcondes (2005) investigated the potential of ‘Web published scientific articles’, conceived not only as texts, but also as machine-readable knowledge bases, which are explicitly and formally related to a Web-based public ontology representing the assented knowledge of a specific domain. His goal is to enhance Web electronic publishing to embody new facilities provided by the Web environment in the semantic Web initiative context. As a first step towards this objective he has presented a model of an article’s scientific methodology structure that has some elements in common with that of Hunter (2006). The methodology of Marcondes follows six steps:

1. Facts, or more precisely, problematic facts;
2. Formalisation of a research problem or question;
3. Development of a hypothesis which is an answer to the research problem;
4. Empirical testing of the hypothesis;
5. Analysis of the test result(s);
6. Conclusion: hypothesis ratification or refusal.

This methodology is a part of the proposed model. The model is an XML structure, hierarchically organised and mapped to XML elements, and has two types of relations, expressed as links: from an article’s ‘deep structure’ to other articles cited in it and an available Web ontology.
All these elements are published as a ‘knowledge base’, using XML, thus outlining a Scientific methodology Markup Language (Sm-ML). Concepts expressed in the different parts of a scientific article are to be linked to public Web ontologies, thus enabling the establishment of a formal relationship between the scientific article specific knowledge base to ontologies like the UMLS – the Unified Medical Language System\(^\text{13}\). In Paragraph 2.3 we have mentioned the Prospect project, where Chemical Markup Language (CML) is used in journals of the Royal Society of Chemistry to link to ontologies.

Marcondes advises development of tools that permit:
1. Electronic publishing of research results both as text and as a knowledge base;
2. Explicit relation of this knowledge base to other scientific articles and to public Web ontologies, which store the established corpus of knowledge of a specific domain;

\(^{13}\) http://www.nlm.nih.gov/pubs/factsheet/umls.html
3. Other researchers to navigate throughout a semantically rich network of enhanced text/ontologies articles, to check their validity and coherence, and to compare, comment and semantically query them.

According to Marcondes, the proposed model will improve the scientific communication process. With the aid of intelligent software agents the rich Web environment will permit browsing and navigation, semantic retrieval, critical enquiry, semantic citation, comparison, coherence verification and validating a scientific article against Web public ontologies. His model is also conceived as the basis for developing enhanced authoring and retrieval tools.

Fink and Bourne (2007) are developing authoring tools to make the use of XML significantly easier. In their project, they use the NLM DTD (National Library of Medicine - Document Type Definition) to store a publication in a standardised and machine-readable format. This DTD also includes some semantic markup of the content, unique identifiers for the article itself and for the objects, e.g. figures and tables, within it. The tool they are developing in the project BioLit is a set of open-source tools that will facilitate the integration of open literature and biological data. Initially, they will develop and test these tools using the entire corpus of the Public Library of Science (PLoS) and the Protein Data Bank (PDB). Although biological objects are being used, the aim is to design tools that will be generally applicable to all literature and other biological data.

### 3.4 Repository Interoperability

The essence of the open archives approach is to enable access to Web-accessible material through interoperable repositories for metadata sharing, publishing and archiving. The OAI-Protocol for Metadata Harvesting (OAI-PMH) defines a mechanism for harvesting records containing metadata from repositories. OAI-PMH offers a simple technical option for data providers to make their metadata available to services, based on the open standards HTTP and XML. The metadata that is harvested may be in any format that is agreed by a community or by any discrete set of data and service providers, although the default format for exchanging metadata between institutional repositories and service providers is simple Dublin Core (DC). For many services based on harvesting repositories, this format has significant
limitations. A well-known problem of simple Dublin Core is its lack of detail or granularity. For example, there are no separate elements for volume, issue and page.

3.4.1 Other descriptive Metadata Formats
Instead of Dublin Core (DC) other descriptive or bibliographic metadata formats can be used to obtain more granularity (Foulonneau & André, 2007).

Qualified Dublin Core 14
This is not a metadata format but rather a list of DC terms that can be used in any application profile. No official schema exists for encoding qualified DC. In practice, a multiplicity of schemas has been created to use QDC. It is maintained by the Dublin Core Metadata Initiative.

MARC 15
A set of metadata formats traditionally used in the library community to create and exchange bibliographic information. Multiple local implementations of MARC exist. MARCXML has been created from MARC21, merging USMARC and CAN/MARC. The MARCXML schema is maintained by the Library of Congress.

MODS 16
Metadata Object Description Schema (MODS) is a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. It is maintained at the Library of Congress.

Scholarly Works Application Profile (SWAP) 17
SWAP (originally the EPrints Application Profile, renamed Scholarly Works Application Profile) is a Joint Information System Committee (JISC) initiative to create an application profile ‘for describing scholarly publications (EPrints) held in institutional repositories’. Instead of defining new metadata elements for simple DC, it reengineers the design of metadata for EPrints.

14 http://dublincore.org/documents/dcmi-terms/
15 http://www.loc.gov/marc/
16 http://www.loc.gov/standards/mods/
17 http://www.ukoln.ac.uk/repositories/digirep/index/SWAP
The data model considers EPrints in the context of research activity. The model does not represent stable documents but a scholarly work that corresponds to one or more copies of that work and one or more agents that were involved in the creation of that work. In the SWAP model, the scholarly work is conceived as the result of a process in which a number of actors (creator of work, funding organisation, identifier) intervene.

### 3.4.2 XML containers

There are also a number of XML containers, instead of descriptive metadata formats, that include different type of metadata and (parts of) digital objects or links to digital objects.

#### METS

METS is a structured container of different types of metadata: descriptive, structural and administrative. It does not determine which metadata format(s) is/are going to be used for the description of the digital object. As a result, it is implemented with application profiles. However, a number of metadata formats have been endorsed as METS extensions, i.e. DC, MODS, PREMIS. It has been extensively used in the library community to represent and exchange information about complex objects, including a hierarchical representation of their structure, especially for digitised books. It is maintained by the Library of Congress.

#### DoMDL

The Document Model for Digital Library (DoMDL) is used in OpenDLib. OpenDLib is a Digital Library Service System that allows publishing or self-publishing, maintenance and dissemination of documents that conform to the DoMDL document model to represent multi-edition, structured, multimedia documents that can be disseminated in multiple manifestation formats.

#### MPEG21-DIDL

In the MPEG-21 Framework, complex digital objects are declared using the Digital Item Declaration Language (DIDL). DIDL is a subset of the MPEG-21 ISO standards. It includes the representation of the object as multiple metadata sets or pointers to multiple parts of the object. DIDL is more suitable as exchange format than METS for repositories, because each object or part of an object can have an identifier and a date stamp in DIDL. When a part of an object is renewed, the date stamp
also changes and, due to a cascading effect, the date stamp of the whole DIDL document also does.

**LOM, IMS CP and SCORM**

LOM IEEE standard is developed for Learning Object Metadata (LOM). LOM is a metadata standard to describe educational resources. It is aimed at the exchange and reuse of learning objects. It contains many types of metadata for the inclusion of learning resources into Learning Management Systems. The IMS Global Learning Consortium has developed specifications for Learning Resource Metadata as well as for content packaging. IMS Content Packaging (IMS CP) is a standard for the assembly of resources, metadata and sequencing information into a learning object. The SCORM 1.2 profile extends IMS Content Packaging with sophisticated sequencing and Contents-to-LMS communication.

**Data Documentation Initiative (DDI)**

Enhanced Publications also need some other metadata formats or XML containers different from those mentioned above, e.g. to structure the associated data sets to publications. The Data Documentation Initiative (DDI) is an important format to structure data sets. The DDI is an XML specification for social science metadata that is being developed by an international group called the DDI Alliance. The bylaws of the Alliance mandate an open Public Review period for the latest draft release (version 3.0). DDI also has the potential to structure data sets from other scientific disciplines than the social sciences. Metadata models or XML containers are also developed to structure data and data sets in specific scientific disciplines. For instance:

- MIDAS Heritage is the UK data standard for information about the historic environment (archaeology).
- NetCDF Climate and Forecast (CF) Metadata Convention defines metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data.

### 3.5 Scholarly Communication

An Enhanced Publication may be seen as a type of compound object. It is not one file or a composition of several files, but consists of digital objects or resources that can be distributed over several locations. In DIDL the entities of an Enhanced Publication are digital items. In METS digital objects are modelled as tree structures, e.g. book with chapters.
with subchapters. Every node in the tree can be associated with
descriptive/administrative metadata and individual/multiple files or
portions thereof. For the description and structuring of an Enhanced
Publication XML containers are essential.

Yet, the use of XML containers and metadata formats for data sets is
just a starting point. In an earlier publication, Herbert van de Sompel
proposed a more advanced notion of a scholarly communication system
that would fully cover the way scholars work. Such a system would
allow flexible composition of information units – text, data sets, images
– and distributed services for the formation of new types of published
results and new methods of collaborating. This results in a loosely
coupled system based on an interoperability fabric where the units of
scholarly communication may follow a variety of scholarly value chains.

A core component of this vision is a new notion of the scholarly
document or publication. Rather than being static and text-based, this
scholarly artefact combines data, text, images, and services flexibly in
multiple ways regardless of their location and genre. This vision
requires an interoperability fabric that is considerably richer than is
provided by OAI-PMH. Rather than just allowing exchange of structured
descriptive metadata, it needs to represent and exchange information
about compound digital objects: their structure, lineage, and persistent
identity. OAI-ORE is the model for such a vision.

Open Archives Initiative Object Reuse and Exchange (OAI-ORE)
OAI-ORE defines standards for the description and exchange of
aggregations of Web resources. The World Wide Web is built upon the
notion of atomic units of information (‘resources’) that are identified
with URIs. In addition to these atomic units, aggregations of resources
are often units of information in their own right.

The OAI-ORE specifications are based around the ORE Model. The ORE
Model introduces the Resource Map enabling association of an identity
with aggregations of resources and making assertions about their
structure and semantics. The primary serialisation format for Resource
Maps is a profile of Atom. Because the ORE Model is expressed in RDF,
Resource Maps may also be serialised in any format capable of
serializing RDF.
A Resource Map describes an Aggregation which is a set of resources, and possibly their types and relationships between the resources. Resources in the Aggregation are called Aggregated Resources. In order to be able to talk about the Aggregation on the Web, it must have a URI (say A-1). According version 0.2 of OAI-ORE this URI is constructed by appending #aggregation to the Resource Map URI, i.e. ReM-1#aggregation. This syntactic device ensures that there is a unique Aggregation resource for every Resource Map.

Figure 6b. Basic of the OAI-ORE model

Figure 6c shows a complete Resource Map with statements indicated as arrows from subject resource to object resource or literal.
Figure 6. Model of OAI-ORE
4. Characteristic Features of Objects

Summary
A full registration requires a timestamp, a unique identifier, a resolver and in the case of a data set a universal numeric fingerprint. Scholarly objects should be provided with a unique citation to be traceable to the cited work.

All objects should clearly have a mark of peer review including quality and integrity standards. Concerning legal issues we can learn a lot from archives, which have dealt with data sets for many years, and have developed contracts for deposit and use.

To optimise the visibility of all objects of an Enhanced Publication, all objects should
• be provided with a unique global identifier so they can be easily linked to and retrieved;
• be cited in such a way that citations can be easily found;
• be stored in a Unique Global Identifier;
• have a Universal Numeric Fingerprint and a Bridge service;
• be conceived in a resolver sustainable way, in which repositories are preferable to publishers’ Web sites;
• be machine-readable, for indexing and using or reusing the materials;
• be freely available, for indexing and easy accessibility.

Universities and research institutes should assume the responsibility of archiving the digital scholarly output of their organisation in a sustainable way. Repositories have a suitable harvesting infrastructure for archiving scholarly output in national libraries or archives; disciplinary repositories and the Dataverse Network should also be considered.

Repositories should enable provision of a complete reference to all deposited objects. Researchers who use the scholarly object can then properly cite the material. It is also very important to show the number of citations with the materials. This will reveal the importance of the research(ér).
4.1 Introduction

Van de Sompel and Lagoze (2007) have suggested that the crucial functions of scholarly communication, being registration, certification, awareness, archiving and rewarding, should be re-implemented in a new context. A ‘service hub’ manages each function of the publication system. The service(s) needed determine which ‘service hubs’ are necessary and in which order (Herbert van de Sompel et al., 2004). In this section we look at each function of the publication system to see what is of importance for that particular function in composing an Enhanced Publication.

4.2 Registration

Registration is the most important function in the system of scholarly communication. It is necessary that discoveries, inventions and innovations are publicly registered and claimed. With proper registration, you can prove that a particular idea was yours, that you were the first. Altman and King (2007) propose that, besides the three traditional components author(s), date of publication and title, citations to numerical data should include three other components, a timestamp, a universal Numeric Fingerprint and a unique identifier. These components are necessary to identify a particular object unambiguously and to guarantee a stable location, easy retrieval and verification of the study.

4.2.1 Timestamp

The registration date is recorded by the journal publisher as the date the manuscript is received. It is important to look at the timing of registration and how this relates to the quality of what is to be registered (Herbert van de Sompel et al., 2004). Time-stamping is not only important for a publication but also for all other objects of scholarly communication, like data sets, images, movies, models, database queries (Sayeed Choudhury, 2008), and also for each original aggregation of objects.

4.2.2 Universal Numeric Fingerprint (UNF)

Normally it is easy to check whether two objects are the same or not, but with data sets this task may be quite complicated. Data correction, updating and maintenance follow separate curatorial mechanisms, if so at all. These mechanisms are not synchronised to those that manage the article literature (Lynch, 2007). Besides this we also need to guar-
antee and verify independently that the object has not changed in any meaningful way when the data storage format has changed. A Universal Numeric Fingerprint (UNF) solves these problems.

The UNF is a short, fixed-length string of numbers and characters that summarises all the content in the data set in a way that a change in any part of the data would produce a completely different UNF.

A UNF works by translating the data firstly into a canonical form with fixed degrees of numerical precision. Then it applies a cryptographic hash function to produce the short string. The advantage of canonicalisation is that UNFs, but not raw hash functions, are format-independent. They keep the same value even if the data set is moved between software programs, file storage systems, compression schemes, operating systems, or hardware platforms (Micah Altman & King, 2007).

A UNF differs from an ordinary file checksum in several important ways:

- UNFs are format independent. The UNF for the data will be the same regardless of whether the data is saved as R binary format, SAS formatted file, Stata formatted file, but file checksums will differ;
- UNFs are robust to insignificant rounding error. A UNF will also be the same if the data differs in non-significant digits; a file checksum will not;
- UNFs detect misinterpretation of the data by the statistical software. If the statistical software misreads the file, the resulting UNF will not match the original, but file checksums may match.

UNFs are strongly tamper resistant. Any accidental or intentional change to the data values will change the resulting UNF. Most file checksums and descriptive statistics detect only certain types of changes. UNF libraries are available for standalone use, for use in C++, and for use with other packages (Micah Altman, 2006). More information about the UNF can be found in the work of Altman et al. (2003) platforms (Micah Altman & King, 2007).

### 4.2.3 Unique (Global) Identifier (UGI)

Each unit of scholarly communication or aggregate thereof needs a unique identifier. This is important for all levels of scholarly communication: articles, modules, objects, queries and also for database objects (Seringhaus & Gerstein, 2007).
The Unique Global Identifier (UGI) is a short name or character string guaranteed to be unique that permanently identifies the data set, independently of its location. Altman and King (2007) allow any naming scheme as long as it identifies the object unambiguously, is globally unique, and is associated with a naming resolution service that takes the name as input and shows how to find one or more copies of the identical object.

They recommend the unique global identifier not to resolve to the object itself, but to a page containing the descriptive and structural metadata describing the object, presented in human-readable form to Web browsers. This metadata description page should include a link to the actual object, as well as a textual description of the object, the full citation in the format we will discuss below, complete documentation, and any other pertinent information. The advantage of this general approach is that identifiers in citations can always be resolved, even if the data are proprietary; require licensing agreements to be signed prior to access; are confidential; demand security clearance; are under temporary embargo until the authors execute their right of first publication, or for other reasons. With metadata description pages like these search engines can find data more easily. Metadata may follow emerging standards, or any other scheme (Micah Altman & King, 2007).

A list with identifiers and identifier resolution services from the e-bank Web site[^18] is included in the Appendix of Part 1.

**Resolver**

Most Web browsers do not recognise global unique identifiers. According to Altman and King (2007) we need a ‘bridge service’. This service is referred to as a Resolver of the identifier.

4.3 Certification

Certification establishes the validity of a registered scholarly claim. Mostly it takes the form of a peer-review process, conducted under the auspices of the journal publisher who certifies the claims made in the manuscript.

King (2007) thinks that not many journals with supplementary

[^18]: http://www.ukoln.ac.uk/projects/ebank-uk/data-citation/
materials have engaged in legal counsel to consider the potential liability surrounding the way they accept and distribute data. He discusses the fact that most journals merely post on their own websites whatever material authors submit, with no check by journal staff, no internal review board (IRB) approval, and not even any signed testimony by the author that distributing the data would not violate any laws, as an author is obliged to do on the copyright-transfer agreement for the publication. King remarks: “Publishers have well honed procedures for dealing with copyright and liability issues for printed matter, but these standard copyright assignment forms do not cover acquiring and distributing data off the printed page.”. His advice is to learn from or work with the archives that have already dealt with these issues.

Some journals explicitly mention that supplementary materials are not peer reviewed and thus are a pure responsibility of the author. Lynch (2007) questions the relationships between traditional peer review and the material that underlies an article under review. He says that it is often unclear if the peer review of an article extends to peer review of the underlying data, even when journal policies are explicit on this. When linking objects we have to know something about quality and integrity standards. We need to be sure that within the law of proper scientific discourse, all knowledge presentations are equal (Joost G. Kircz, 2002).

Some claim that it is impossible to peer review materials other than publications. In some fields presently there is an effort to build peer-review systems around data, so data can be judged formally on qualities of coherence, design, consistency, reliability of access, and so on. In the UK, scientists and professional associations in the field of meteorology have joined forces to establish a new kind of electronic publication called a data journal, where practitioners can submit data sets for peer review and dissemination: the Overlay Journal Infrastructure for Meteorological Sciences (OJIMS). NCAS (Leeds) will provide feedback on how this will work from an editor’s point of view and help write the reviewers guidelines (Arms & Larsen, 2007). Another example is the Journal of Visualised Experiments (JoVE)\(^\text{19}\), a peer-reviewed journal for video-protocols.

\(^{19}\) http://www.jove.com/
4.4 Awareness

Awareness allows actors in the scholarly system to be informed of new claims and findings. Authors want the maximum amount of visibility, authority and prestige to ensure the validity of their claim. This is unnecessarily restrained by the way publishers and repositories have organised the publishing environment. Supplementary materials are mostly handled as individual objects, even though they are sometimes closely related. As we will see, in most cases there is no link between related objects.

4.4.1 Current Practice

Supplementary materials are mostly contained within the publisher’s domain. Publishers started enhancing publications with digital objects on the Web around 2000. An article from 2001 lists movies with the abstract. In 2008 the same approach is used, for example at Elsevier and Blackwell. This method of presenting supplementary materials, called publishing with the article, will help other researchers to gain easy access to them once the article itself is found. Some publishers (BioMedCentral and Nature) are now also using separate channels for supplementary materials. Nature has a special Web site Nature Multimedia with streaming videos featuring discussion and analysis with scientists as they share their discoveries, pod casts with a free weekly audio show highlighting the best of the week in science, blog on which you can have your say on the news of the day, tools to find out how the novel use of programs like Google Earth can help scientists discover and share information and specials like interactive graphics, quizzes, video galleries, pictures and more.

BioMed Central publishes movies about some of their articles on SciVee. But SciVee presents more than movies. It features figures, supplementary materials, references, related pubcasts, tags, submitted by, rating, uploaded, views, comments, and add to favourites like digg.

---

22 http://www.nature.com/nature/multimedia/
23 e.g. http://www.biomedcentral.com/1471-2105/7/242
24 e.g. http://www.scivee.tv/node/2727
IT, del.icio.us, Slashdot, Furl, Facebook, StumbleUpon, Newsvine, Technorati, Reddit, Magnolia, Blinklist, and Netscape. It also contains URL, comments, viewers’ notes and copyrights concerning the movie. As well as with Scivee, the movie can also be found on YouTube²⁵.

Amazingly, BMC and Nature do not have an active link between articles²⁶ and movies²⁷ or podcasts²⁸. On the SciVee Web site information about articles is not linked, and on the Nature Web site the reference only appears for a few seconds at the beginning of the movie; on the article ‘homepage’, in both cases no information whatsoever can be found about movies on the Scivee or Nature video Web sites. Applying our definition of an Enhanced Publication it can be said that these publishers own the building blocks, but do not use the opportunity to create an Enhanced Publication.

Very few journals were found that link to authors’ institutional Web sites to show movies and tools, like the Journal of Interactive Media in Education²⁹. It is surprising that after more than six years many links are still working, even though researchers and researcher groups within institutions come and go. Most objects could still be found, albeit sometimes via a roundabout route. Linking to authors’ institutional Web sites should never be supported in our view, since every other solution, e.g. repositories, publisher’s Web site and commercial databases, is much safer. Although a publisher like IOP (Wray, 2007) has declared that they would welcome to collaborate with repositories, especially for grey literature, we have found no connections to a repository on the publisher’s Web site. Repositories on the other hand often refer to the official publication on the publisher’s Web site.

We have seen that supplementary materials need not necessarily be linked to the publication. They can be found on several locations:

- Repository, institutional or disciplinary;
- Web site of the author’s institution;
- Publisher’s Web site;

²⁵ e.g. http://www.youtube.com/watch?v=V72stkORlgo
²⁶ article: http://dx.doi.org/10.1038/nature06669 (subscription needed).
²⁷ video: http://www.nature.com/nature/videoarchive/selfhealingrubber/ (free access).
²⁹ http://www-jime.open.ac.uk/2002/8/
• Blogs and wikis, like Open Notebook Science\textsuperscript{30};
• Database, like the Protein Data Bank.

The fact that the publication and supplementary materials are not linked reduces the visibility of the supplementary materials.

\subsection*{4.4.2 Model for the Future}

We have seen that supplementary materials are hosted across several locations. They are managed in many different ways, and sometimes are not linked to the publication. What can be done to improve the visibility of the publication and the supplementary material?

• The publication and the related objects should both have a unique identifier to ensure they can be referenced;
• The links between the publication and the supplementary materials should be set up on both sides, from the publication to the supplementary materials and vice versa;
• The objects must be held at a trustworthy location and published on an open access basis, which will result in more downloads and more citations\textsuperscript{31}.

What locations are most suited to store supplementary materials? According to Lynch (2007) disciplinary repositories, national or international, are preferred. Journals are less suited:

• Not every journal accepts supplementary materials;
• Journals have no clear policy about preserving data or the tools to work with them;
• Some journals put constraints on the amount of data they will accept;
• Access to the data is for subscribers only;
• It is often unclear to what extent supplementary materials are part of the peer review.

The most important advantage of a disciplinary, as well as institutional, repository is the use of persistent identifiers that can be used to reference objects uniquely. None of the journals use persistent identifiers for supplementary materials. The benefits of international disciplinary repositories are that most of these communities have established norms, enforced by editorial policies. Journals can use

\footnotesize\textsuperscript{30} The projects Useful chemistry/Open Notebook Science are using wikis and blogs to record research http://usefulchem.wikispaces.com/
\footnotesize\textsuperscript{31} http://opcit.eprints.org/oacitation-biblio.html
repository identifiers to reference objects. However, the problem is that not all communities have a repository available. As yet, there are international repositories for only a few disciplines, like crystallography, astronomy, economics and molecular biology. Another problem is that these repositories do not accept all types of objects. In cases where objects cannot be stored in a disciplinary repository, an institutional repository may be a solution. The disadvantages of institutional repositories, set up as data providers, are that they lack easy centralised searching of material on a disciplinary basis, and that it may be very difficult to develop and maintain specialised discipline-specific software tools, as they are not intended as service providers (Lynch, 2007).

Other alternative locations for storing data sets are the Dataverse Network and data grid. In contrast to Lynch (who advises disciplinary repositories) King (2007) pleads for archiving data in the international Dataverse Network (DVN)32. This is a World Wide Virtual Network for data archiving, housed at the Institute for Quantitative Social Science (IQSS) at the Harvard-MIT Data Center. Coding of the DVN software began in 2006, but it is a continuation of the earlier Virtual Data Center (VDC) project (1999-2006) which was collaboration between the Harvard-MIT Data Center and Harvard University Library. The current DVN offers many features, uses many standards and is not intended for social science data only. A data grid is a grid-computing system that controls sharing and management of large volumes of distributed data. These systems are often, but not always, combined with computational grid-computing systems. Many scientific and engineering applications require access to large volumes of distributed data, terabytes or petabytes. Current large-scale data grid projects include the Biomedical Informatics Research Network (BIRN), the Southern California Earthquake Center (SCEC), and the Real-time Observatories, Applications, and Data management Network (ROADNet), all of which use of the SDSC Storage Resource Broker as the underlying data grid technology. These applications require widely distributed access to data by many people in many places. The data grid creates virtual collaborative environments that support distributed but coordinated scientific and engineering research33.

Another question is whether data should be stored at multiple locations.

32 http://thedata.org/
33 source: http://en.wikipedia.org/wiki/Data_grid
LOCKSS\textsuperscript{34} (Lots of Copies Keep Stuff Safe) is open-source software that provides librarians with an easy and inexpensive way to collect, store, preserve, and provide access to their own, local copy of authorised content. According to Lynch (2007) repeated publication of the same data is clearly undesirable. The right approach would be citation or similar forms of reference. Therefore, it should be very easy to link to data, including data in databases like the Human Genome Project\textsuperscript{35}. Using the Van de Sompel and Lagoze model (2007), duplication of objects will not be necessary. The unique persistent link is sufficient to reuse the object. It even offers the possibility of ‘virtual publication’ in which combining objects can express a new idea without duplicating the objects\textsuperscript{36}. This may be considered as a ‘virtual publication’ as all parts are found at different locations.

Use or reuse of publications or parts of them improves visibility and is also part of the reward for publication. For this, one should know what access rights are applicable to an object, article, aggregates. Who is the owner, and what rights apply. Creative Commons, copyright, transferred copyright? Although free and open access to all materials is normally preferable because of more visibility, it is neither always feasible nor necessary for the purpose of guaranteeing public distribution of quantitative information. Those who wish to access the objects can reasonably be asked to fulfil whatever authorisation requirements the original author needed to meet in order to acquire, distribute, and archive it in the first place. This may include signing a licensing agreement such as agreeing to respect confidentiality pledges made to research participants, signing the equivalent of a guest book or belonging to an institution with a membership to the archive. Different requirements may apply to different objects (King, 2007).

Rights for articles are clearly stated on all publishers’ Web sites. For most supplementary materials this is not the case. For most materials no copyright owner is stated. Most toll-access publishers give no access to the supplementary materials, which is surprising because the

\begin{itemize}
\item \textsuperscript{34} http://www.lockss.org
\item \textsuperscript{35} http://www.ornl.gov/hgmis/home.html
\item \textsuperscript{36} As has been done in the NSDL Science Literacy Maps: Concept maps for science and math. education (http://NDSL.org). The service is helping teachers to connect concepts, standards, and NSDL resources but also uses (parts of) objects from other repositories, for example: http://strandmaps.nsdl.org/?id=SMS-MAP-1594
\end{itemize}
materials are not part of the peer-review process. Blackwell states: “supplementary material will be published as submitted and will not be corrected or checked for scientific content, typographical errors or functionality. Although hosted on Blackwell Synergy, the responsibility for scientific accuracy and file functionality remains entirely with the authors. A disclaimer will be displayed to this effect with any supplementary material published.”. Contrary to this statement the disclaimer cannot always be found on the Web site or on the supplementary materials. Only a few publishers (e.g. Nature) give access to the supplementary materials as well.

In addition to the access barriers, which will make it difficult to link between the worldwide results of scientific research, formats create another problem. In most cases it would be much more interesting having also the data on which a graph was based. Another problem is that not all PDF files can be searched as they are sometimes saved as images. These image files will never be machine-readable even when we have the right to access the file. To make data mining possible, repositories should require data being as ‘raw’ as possible.

4.5 Archiving

Archiving aims to preserve scholarly materials over time. In the past, libraries took care of publishers’ paper materials. Should libraries also take care of publishers’ digital supplementary materials? We believe that universities and research institutions should take responsibility for archiving their own digital scholarly materials in which they have invested. This holds in particular for supplementary materials, as publishers are not clear about archiving these. Only a few publishers guarantee that they will continue to support the accepted submission types of supplementary materials, now and in the future.

We know that researchers centuries from now need to be able to find supplementary material, access it, ascertain that it is the supplementary material associated with the article in question, and verify that it contains the same information that the author originally provided. Of course updates and new versions are desirable for some purposes, but the original version associated with the article must always remain available. The Version Identification Framework37 (VIF)

37 http://www.lse.ac.uk/library/vif/index.html
provides practical advice and recommendations to authors and content creators, repository managers and those involved with repository software on how to identify versions better.

We will need to follow procedures that ensure that the results of research will remain valid in the indefinite future, regardless of changes in methods of data distribution and network access, data storage formats, statistical and database software, operating systems, and computer hardware (King, 2007).

As a result of issues identified in UKOLN’s ‘Dealing with Data’ report, the Digital Curation Centre in the UK has recently set up a Research Data Management Forum to support the exchange of experience and expertise in the management and preservation of research data output.

Are libraries prepared to do the job? To keep the information accessible by migration or emulation libraries should cooperate on a national, European or international level. In the Netherlands for instance, libraries and research institutions can archive publications at the National Library of the Netherlands and data sets or images at Data Archiving and Networked Services (DANS) or the 3TU Data Centre of the Technical Universities in the Netherlands. The Dataverse Network (DVN) should also be considered. Libraries should help their organisations to develop policies on archiving.

4.6 Rewarding

Researchers reward each other by citation or when their material is published in a high impact journal. As is already pointed out, it is obvious that visibility enhances the impact of a publication, since a more visible publication generally will be downloaded and cited more often.

Metrics have been developed on the basis of citations to measure the performance of the research or the researcher. These metrics only count publications. That is why Seeber (2008) criticises citations in supplementary materials, which cannot be found in Web of Science. He

http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_dealing_with_data.aspx
http://www.dcc.ac.uk/data-forum/
suggests adding the references to the article instead of to the supplementary materials. However, publications and data sets should be seen as two separate objects of scholarly communication with their own references. The solution should be to make the references in the supplementary materials visible.

Separating publications and supplementary materials is important, as citation credit should be given to the right objects. That object can be the original publication but also another unit of scholarly communication. Researchers should be rewarded for the measurements, images, tools, peer review, comments and database annotations they have published (Fink & Bourne, 2007; Seringhaus & Gerstein, 2007). Using clear and unique citations to all objects would give credit to them. This applies also to scholarly output.

The data citation model of King (2007) offers proper recognition to authors as well as permanent identification through the use of global, persistent identifiers instead of URLs, which may change frequently. The Dataverse Network provides a unique reference to the data after depositing. After uploading, the author is issued a formal citation to the data, including a unique global identifier, a universal numeric fingerprint, and the other components, to be used in the author's article and to be cited in other publications (King, 2007). When data can be stored with this proper recognition to authors as well as permanent identification it will be possible to publish raw data and leave the analyses to other researchers. In this way credit can be given for the data that has been measured and/or analyzed.

Marcondes (2005) raises the question of why we do not use qualified citation. This is a citation in which the reasons to cite and the relationship between this specific scientific article and its citations are made explicit.
5. Relations

Summary
Instead of inserting an URL, as almost all publishers do nowadays, it is also possible to provide links with a ‘meaning’ in which the logical connection of terms establishes interoperability between parts of an Enhanced Publication. However, for the parts of an Enhanced Publication to interact in a semantic way, we need servers that expose existing data systems using the RDF standards, documents ‘marked up’ and ontologies, and authors must be asked for extra effort.

Three models can help to shape an environment for Enhanced Publications.

- The DCMI Abstract Model can give a better understanding of the kinds of descriptions that are encoded, and how to facilitate better mappings and cross-syntax translations;
- The CIDOC CRM has shown how a successful ontology has been set up; that it can be used to support integration, not only in cultural heritage but also in a large range of different domains including e-science and biodiversity; that it can help to reduce development time; that it improves system quality, and provides basic semantic interoperability for free;
- ORE Vocabulary and Fedora Relationship ontology have defined a default set of common relationships that can be refined or extended.

The conclusion is that ontologies should be used in the environment of Enhanced Publications and that OWL helps to make ontologies compatible.

5.1 Introduction

In the text of a publication or an Enhanced Publication authors can refer to research data and extra materials. Referring may be done manually but also through automated or semi automated semantic linking. Enhanced Publications may also import data or metadata from other systems, in order to complement a publication.
5.2 State of the Art

If authors wish to refer in the text to an annex containing research data and/or extra materials, they are required by the publisher to use a standard format for this reference, for example the phrase "see Electronic Annex 1 in the online version of this article.". The production department of the publisher will insert the relevant URL at the typesetting stage, by means of a hyperlink after this statement. Publishers currently only provide a one-way link from the article to the supplementary material, but never vice versa.

We have found that only one publisher (the American Astronomical Society, AAS) provides an extensive system of links between the literature and other external online information. The basis for this system is the identification of data sets. All astronomical data centres assign unique identifiers to each set of data. It is up to the data centre to decide what they consider a data set: one spectrum, one exposure or a set of exposures of the same object at different wavelengths. In some cases, data sets are defined by the query parameters to a database query. Some data centres also provide the means for authors to define a collection of data sets that they used in an article and to give this collection a unique identifier. The main requirement for data set identifiers is that they have to be unique and permanent. This means that the data centres have to agree to recognise published identifiers in perpetuity (Eichhorn et al., 2007).

Publishers mostly include the links manually. One exception is the addition of hyperlinks in lists of literature references. For instance, Crossref\textsuperscript{40} automatically provides the DOIs for research content in the publication. The Crossref database covers millions of articles and other content items from several hundred scholarly and professional publishers.

If we want to create Enhanced Publications in the repository environment using OAI-ORE, we can build on a part of the current infrastructure and some of the current standards. All objects with a URI, for example a publication and a data set, can already be connected; and some parts of the current infrastructure and standards can be converted to the new infrastructure and standards: DIDL containers can be converted to Resource Maps to which we can add

\textsuperscript{40} http://www.crossref.org/
URIs. Another example is the Digital Author Identifier, which is used in the Netherlands. Aggregates can be made on DAI\textsuperscript{41}. An author's new publications will be automatically added to the publication list (Place, 2008).

### 5.3 Semantic Publishing

As we have seen, Enhanced Publications consist of heterogeneous data, heterogeneous information types, information from different disciplines, different languages. The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

Professionals and researchers in the field require that the descriptive information is sufficiently detailed and precise. For this purpose, interoperability of the parts of an Enhanced Publication is required not only at the syntactic and system level but also at the more semantic level. Semantic integration is the process of using a conceptual representation of data and of their relationships to eliminate possible ambiguities. The problem is that the same elements may express different meanings for different cases and types. Ontologies offer the solutions to the semantic heterogeneity problems and can be used in integration architectures as a global schema to which metadata from different sources can be mapped (Kakali et al., 2007).

Machine-readable descriptions should enable content managers to add meaning to the content, i.e. to describe the structure of the knowledge we have about that content. The semantic web makes it possible, instead of only inserting plain URLs in the text, to set up a logical connection of terms that establishes interoperability between the different parts of an Enhanced Publication.

The semantic web has several components.

- *Uniform Resource Identifiers (URI).* Identifier, used to identify resources on Internet and so they are central to the Semantic Web. Computer scientists may classify a URI as a locator (URL), or a name (URN), or both.

\textsuperscript{41} http://<author_service>/rem/atom/<dai>
• **Extensible Markup Language (XML).** Syntax, provides an elemental syntax for content structure within documents, yet associates no semantics with the meaning of the content contained within.

• **XML Schema.** Language, for providing and restricting the structure and content of elements contained within XML documents.

• **Resource Description Framework (RDF).** Data interchange format, a formal description of concepts, terms, and relationships within a given knowledge domain. It is a representation language for Universal Resource Identifiers (URIs). The semantic web is based on the Resource Description Framework (RDF).

• **RDF Schema (RDFS).** Taxonomy, a structured vocabulary for describing properties and classes of RDF-based resources, with semantics for generalised-hierarchies of such properties and classes.

• **Web Ontology Language (OWL).** Ontology, adds more vocabulary for describing properties and classes, and describes the function and relationship of each of the components above. Ontologies can refer by virtue or URIs.

URIs have a global scope and are interpreted consistently across contexts. Associating a URI with a resource means that anyone can link to it, refer to it or retrieve a representation of it. They provide the grounding for both objects and relations. All scientific objects should be mapped for reuse into the system of URIs. Anything that can be identified with a Uniform Resource Identifier (URI) can be described. So the semantic web can reason about animals, people, places, and ideas.

HTML describes documents and the links between them. RDF, OWL, and XML, by contrast, can describe arbitrary things such as people, meetings, or airplane parts. Tim Berners-Lee calls the resulting network of Linked Data the Giant Global Graph, in contrast to the HTML-based World Wide Web ("Semantic Web", 2009).
5.4 Tools

Tools for publishing papers on the web can automatically help users to include more machine-readable markup in the papers they produce. With current XML-tools\(^{42}\) it is already possible to assert that some part of a document is about an experiment. A new set of languages and tools\(^{43}\) is now being developed to create machine-readable content to make the publications widely available. According to Berners-Lee & Hendler (2001) authors will need to be asked to make some extra effort, in repayment for major new functionality:

- Experimental results on the web can be shared with (trusted) colleagues, outside the context of a research paper;
- More detailed information on the research subjects (chemicals, reactions, location, species);
- The semantic web can break down the walls between disciplines, since it can bring together those concepts that different disciplines have in common.

To enhance the usability and usefulness of the Web and its interconnected resources we need servers which expose existing data systems using the RDF standards, documents 'marked up' and ontologies.

5.4.1 Markup Language

Markup Language is one of the languages that give more detailed information on the research subject and the structure of the object, being embedded formal metadata in documents. Examples are the 'deep structure' (XML) of an article and links with a defined relation to other online objects cited in the article. Marcondes recommends use of qualified citations, in which the reasons to cite and the relationship between this specific scientific article and its citations are made explicit (\(<\text{reason\_to\_cite}>\)). He claims that the use of these XML-tags can enhance the scientific communication process, permitting semantic retrieval, critical enquiry, semantic citation, comparison, coherence verification and validation of a scientific article (Marcondes, 2005).

\(^{42}\) http://www.nature.com/nature/webmatters/xml/xml.html
\(^{43}\) List of Semantic Publishing Tools:
http://esw.w3.org/topic/HCLS/ScientificPublishingTaskForce;
http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/PublishingTools
5.4.2 Ontology

An ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts. It provides a shared vocabulary, which can be used to model a domain. A domain is to be understood as the type of objects and/or concepts that exist, and their properties and relations. Ontologies are attempts to define parts of the data world more carefully, and to allow mapping and interactions between data held in different formats as can be the case in Enhanced Publications. Ontologies are always developed, managed, and endorsed by practice communities, and defined through a careful, explicit process that attempts to remove ambiguity (Shadbolt, Berners-Lee, & Hall, 2006).

Although ontologies are usually developed for specific information domains, and are used to formally represent data in such domains, they share many structural similarities, regardless of the language in which they are expressed. Most ontologies describe individuals or instances, classes or concepts, attributes, and relations.

The Web Ontology Language\(^4\) (OWL) is a semantic markup language specially developed for publishing and sharing ontologies on the World Wide Web. OWL is intended to provide a language that can be used to describe the concepts definitions and the relations between them that are inherent in digital documents and applications. Ontologies can become distributed, as OWL allows ontologies to refer to terms in other ontologies. OWL is developed as a vocabulary extension of RDF. By providing additional vocabulary along with formal semantics, OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) (McGuinness & van Harmelen, 2004).

Since we are interested in models that can be of use for Enhanced Publications, we will look in more detail at four models. We will consider the DCMI Abstract Model, the CIDOC Conceptual Reference Model (CRM), ORE Vocabulary, and Fedora Relationship Ontology.

**DCMI Abstract Model**

The DCMI Abstract Model specifies the components and constructs used in Dublin Core metadata. It describes an information structure called a DC ‘description set’ and specifies how those description sets are to be

\(^4\) [http://www.w3.org/TR/owl-features/](http://www.w3.org/TR/owl-features/)
interpreted. The DCMI Abstract Model is based on the Resource Description Framework (RDF) and builds on work by the World Wide Web Consortium (W3C). The model consists of a resource model, a description set model, a vocabulary model, descriptions, description sets, records, values, semantics and guidelines. It is an information model that is independent of any particular encoding syntax. This model has shown that it can help understanding of the kinds of descriptions that are encoded, and that it can facilitate the development of better mappings and cross-syntax translations (Powell, Nilsson, Naeve, Johnston, & Baker, 2007).

**CIDOC conceptual Reference Model**

The aim of the CIDOC Conceptual Reference Model (CRM)\(^{45}\) is to provide a reference model and information standard for museums and other cultural heritage institutions to describe their collections and related business entities, to improve information sharing. For that the CIDOC CRM provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation “to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information, allowing data to be combined from heterogeneous data sources in a meaningful way and without loss of detail”. The model is not prescriptive, but provides a controlled language to describe common high-level semantics that allow for information integration at the schema level. The names of classes and properties of a CRM-compatible form may be translated into any local language, but the identifying codes must be preserved. By virtue of this classification, data can be understood as propositions of a kind declared by the CRM about real world facts, such as “Object x forms part of: Object y”. By adopting formal semantics the pre-conditions for machine-to-machine interoperability and integration have been established.

It looks like CIDOC CRM will become an important information standard and reference model for Semantic Web initiatives, and will serve as a guide for data modelling more generally ("CIDOC Conceptual Reference Model," 2007). CIDOC CRM can be used by software applications that use RDF, XML, DAML+OIL, OWL and others\(^{46}\). Formal mappings have been established for some data structures, including Dublin Core.

\(^{45}\) http://cidoc.mediahost.org

\(^{46}\) CIDOC CRM Tools and RDF mappings http://cidoc.ics.forth.gr/tools.html
Doerr et al. (2007) claim that the CIDOC CRM is well suited as a new standard for knowledge sharing. They have investigated to what extent and in which form global schema integration is feasible, and they have demonstrated the ability of the CRM to support integration in a large range of different domains including cultural heritage, e-science and biodiversity. Conceptual modelling on the basis of such well-tested core ontology reduces development time drastically. It improves system quality, and it also provides basic semantic interoperability, more or less for free. They have written a tutorial that addresses part of the technology needed for information aggregation and integration in the global information environment. The tutorial first addresses requirements and semantic problems to integrate digital information into large scale, meaningful networks of knowledge that support not only access to source documents but also use and reuse of integrated information. The core ontologies of relationships in particular are fundamental to schema integration. They play a completely different role compared to the specialist terminologies that are used within a discipline.

**ORE Vocabulary**
The Open Archives Initiative (OAI) develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. OAI has its roots in the open access and institutional repository movements. Open Archives Initiative Object Reuse and Exchange (OAI-ORE) has defined standards for the description and exchange of aggregations of Web resources. They have described a glossary of terms and the vocabularies needed to describe items of interest and express the relationships between them within the OAI-ORE context. A guiding principle is to reuse existing vocabularies when possible for terms that are not specific and fundamental to the ORE model, for example Dublin Core Metadata Initiative (DCMI) and the Resource Description Framework (RDF). For vocabularies not described by OAI-ORE or other existing vocabularies, domain specific vocabularies should be created and maintained by their respective communities (Lagoze et al., 2008).

The ORE Vocabulary Definition defines and describes the ORE specific terms. These terms are within the ORE namespace and are used to construct ORE Resource Maps.

**ORE Vocabulary, ORE Classes**
A semantic class should be assigned to resources that are described using ORE. This helps applications to understand what the aggregation contains and represents. For example, an aggregation of journal articles
could be typed as a journal, a journal issue, a journal volume, an overlay journal, a special issue of a journal, a reading list, a citations list, and so on. The core objects or entities of interest within the OAI-ORE context are Aggregation, AggregatedResource, Proxy and ResourceMap.

**ORE Vocabulary, ORE Relationships**
Relationships may refer to another object, but the object of the relation may also be a literal value rather than another resource. Some abstract concepts are both, for example the rights statement could be inline as a string or a reference to an external resource. The discussion is shaped by vocabulary, rather than any distinction as to the subject or object of the relationship. Relationships that exist between entities or from an entity to a literal value: aggregates; isAggregatedBy; describes; isDescribedBy; lineage; proxyFor; proxyIn; and similarTo (Lagoze et al., 2008).

**Fedora Relationship**
The Fedora repository system has defined a default set of common relationships in the Fedora relationship ontology\(^\text{47}\) (actually, a simple RDF schema), which defines a set of common generic relationships useful in creating digital object networks:

```
<isPartOf> <hasPart>
<isConstituentOf> <hasConstituent>
<isMemberOf> <hasMember>
<isSubsetOf> <hasSubset>
<hasCollectionMember>
<isDerivationOf> <hasDerivation>
<isDependentOf> <hasDependent>
<isDescriptionOf> <hasDescription>
<isMetadataFor> <hasMetadata>
<isAnnotationOf> <hasAnnotation>
<hasEquivalent>
```

These relationships can be refined or extended. Also, communities can define their own ontologies to encode relationships among Fedora digital objects ("Fedora Digital Object Relationships - Fedora Repository Release 3.0," 2008).

\(^\text{47}\) http://www.fedora.info/definitions/1/0/fedora-relext-ontology.rdfs
6. Current Repository Projects

Summary
If we look at the current products of traditional publishers like Elsevier and Springer, no integrated metadata models, XML containers or models for scholarly communications are used. Enhanced Publications are mostly just articles with additional files that contain data sets and multimedia material.

In the eCrystals Federation Project metadata about the crystal, like chemical formula of the crystallised material, molecule name, and authors, are stored in simple Dublin Core (DC). Additional chemical information is stored as Qualified Dublin Core. The derived data sets are stored as files in the repository together with the files for representations or images of the crystal and the molecule. The eCrystals Federation Project is based on the eBank project which exports metadata using the OAI-PMH protocol in two different metadata formats, simple DC and METS. The eCrystals Federation Project will be a test bed for the OAI-ORE model (Atom Publishing Model), as DC and METS are too limited.

The ARROW/DART/ARCHER strategy does not require the use of a single metadata schema to describe all digital objects stored in the repository. Multiple formats to suit individual content models can be supported. OCLC and ARROW are working together to test a mapping tool developed by OCLC called the Interoperability Core that is based on mappings and crosswalks between different metadata formats. Metadata can be stored and searched in the native format generated by the community of practice. Using this strategy ARROW can be populated with metadata from a variety of formats and through various mappings converted to an interoperable core that can then be converted to DC for harvesting via OAI-PMH by resource discovery services.

If we look at all the above described projects, repositories are used not only to store and ensure permanent access to publications of multiple types, such as articles, conference-papers, reports and books, but also to store and offer access to data sets, research data, images and multimedia. Metadata varies from DC to METS or individual content models for community specific applications. Links between objects in
Repositories are now mostly bidirectional. Repositories may be discipline-specific or more generally oriented like institutional repositories.

In future, repositories will more and more be used for all kind of data: different publication types, data sets, research data and extra materials, e.g. images, video. Enhanced Publications can be created on the basis of the objects held in repositories. The internal format and repository infrastructure must be flexible enough to deliver common metadata formats such as DC, MODS, DIDL or METS, or more community specific metadata formats. Above all repository infrastructures must support the OAI-ORE model. The development of systems to manage the complete cycle of e-research and scientific collaboration will be based on repository infrastructures.

6.1 The eCrystals Federation

The eCrystals Federation Project\(^\text{48}\) aims to establish a worldwide and sustainable federation of open data repositories supporting Open Science in crystallography. This new project is a continuation of the eBank UK project\(^\text{49}\). The earlier project developed the eCrystals archive\(^\text{50}\), an OAI-PMH compliant repository populated with a number of crystallography data sets. The repository is based on modified EPrints 3 software. It also developed a demonstrator for an aggregation service, using ARC for harvesting and Cheshire for indexing and searching, showing the potential for searching across data set and publication metadata. Phase 3 included a scoping study for an international eCrystals federation.

The sub-discipline of crystallography was originally chosen by eBank because it has a well-defined data creation workflow and a tradition of sharing results data in an internationally accepted standard, the Crystallographic Information File (CIF) adopted by the International Union of Crystallography (IUCr). The projects are funded by JISC\(^\text{51}\) and can be seen as demonstrators for applications in other areas.

\(^{48}\) http://wiki.ecrystals.chem.soton.ac.uk/index.php/Main_Page

\(^{49}\) http://www.ukoln.ac.uk/projects/ebank-uk/

\(^{50}\) http://ecrystals.chem.soton.ac.uk/

\(^{51}\) http://www.jisc.ac.uk/
The eCrystals archive stores data from crystallographic analyses as metadata records and files. When a crystal is analysed, terabytes of data are generated and held in proprietary formats. From this, data sets and data presentations can be derived which are substantially smaller and more interoperable. These derived data sets are stored as files in the repository together with the files for representations or images of the crystal and the molecule. The metadata for the crystal, e.g. chemical formula for the crystallised material, molecule name, and authors, is stored as simple Dublin Core. Additional chemical information like International Chemical Identifier, is stored as Qualified Dublin Core. In the eBank project, links were made from the metadata record to scientific articles and vice versa.

As part of the eBank UK project an Application Profile\(^5^2\) was developed, configured according to the Dublin Core Metadata Initiative Abstract Model and the Dublin Core Application Profile Guidelines. The eCrystals project aims to harmonise the application profiles from participating repositories operating on different platforms like EPrints, DSpace, Fedora & ReciprocalNet, and establish a core Federation Application Profile and mappings. In the first phase of the project, only institutions using the EPrints platform will be included; the second phase roll out will include other platforms. Aggregation services will be examined and implemented at a national and international level.

The eCrystals Federation project aims to make a substantial contribution to a repository-based e-infrastructure for research data. Its broad outline will be as follows:

Metadata and data sets will be entered into institutional data repositories. Aggregators will harvest the metadata and provide services, and links will be made to scientific publications. So articles by the scientists depositing data could be held either in institutional or subject repositories. Bidirectional linking to publishers’ ‘repositories’ with scientific articles will be examined and implemented. Within eBank DOIs were used to link data; other options may be considered for the eCrystals Federation project.

The project also aims to enable the Federation to interoperate with international subject archives IUCr and CCDC (Cambridge Crystallographic Data Centre) and other third party harvesters.

\(^5^2\) http://www.ukoln.ac.uk/projects/ebank-uk/schemas/profile/
The eCrystals Federation will be a test bed for the OAI-ORE model (Atom Publishing Model), as DC and METS are too limited. This is likely to be a key role for the project.

6.2 ARROW, DART and Archer at Monash University

Monash University is a large university with 55,000 students in various countries, among which are Australia (its main location), Malaysia, and South-Africa. University policy and the institutional context of an e-research infrastructure for institutional repositories are impressive.

Two years ago a start was made with the development of an Information Management Strategy in close cooperation with researchers and with the full support of higher management and the whole university. This led to the establishment of an E-research Centre to support researchers with e-research. Data size can vary greatly, but in some cases it is extremely large, e.g. Synchrotron 1 terabyte per day, Astronomy 12 terabyte per day.
Three national projects play an important part. The ARROW project aims at developing the management of scientific output in repositories. The DART project focuses on the development of solutions for a complete cycle of e-research: doing experiments, data analysis, publishing, learning. As a result data sets, storing data sets, preprints, and reports are generated. ARCHER is a continuation of DART aiming to become a production service for the national e-research environment (Treloar & Groenewegen, 2007).

As an example, the whole cycle of the analysis of and publication about a protein structure can be captured. The data sets generated by the analysis are eventually stored in the repository and linked to the scientific article on this protein structure.

It is thought necessary to store data for the following reasons:
- Data is holy;
- Data validation and validation of the scientific article;
- Some data elements could not be analyzed. Perhaps others are able to do this;
- Examples of data sets for those who develop analysis methods.

The institutional and centralised approach resulted in a central data storage facility, called Large Research Data Storage (LaRDS). Researchers are given 100 gigabyte, and if they want more than this, a one-off fee is payable.

For individual research, shared research and public domain, different storage systems are created. For the institutional repository in the public domain FEDORA is used with VITAL from VTLS. The major challenge in an IR is how to deal with large files, since for a protein analysis a sub file may already contain 36 gigabyte (Treloar, Groenewegen, & Harboe-Ree, 2007).
6.3 eSciDoc Project

eSciDoc\footnote{http://www.escidoc-project.de} is a shared project of the Max Planck Society and FIZ Karlsruhe, funded by the Federal Ministry of Education and Research.
(BMBF) in Germany, with the aim of developing a platform for communication and publication in scientific research organisations. The eSciDoc project is intended to:

- Ensure permanent access to the research results and research materials of the Max-Planck Society and seamless integration within eSciDoc as well as integration into an emerging, global, electronic knowledge space;
- Provide effective opportunities for access to information for scientists of the Max-Planck Society and their work groups;
- Support scientific collaboration in future eScience scenarios.

**eSciDoc Solutions**

The first eSciDoc solution, publication management, will provide basic functionalities and user interfaces for the submission of publication data of multiple types like article, conference-paper, poster, report, and book, along with the metadata needed for efficient retrieval and long-term archiving. The second solution developed in the eSciDoc project, the scholarly workbench, is aimed at providing a generic solution for communities in the arts and humanities, to store their digital artefacts and make them ‘processable’ and reusable within a collaborative environment.

**Infrastructure**

The eSciDoc infrastructure is composed of middleware that encapsulates the repository and implements services in all layers of the service-oriented architecture relevant to the eSciDoc system. The core infrastructure is mainly built from existing open-source software packages. The eSciDoc content repository is based on Fedora. Fedora comes with a semantic store, which enables efficient administration of statements about objects and their relations, expressed in RDF (Resource Description Framework). Related objects form a graph, which can then be queried or used to infer new facts, based on existing RDF.

**Metadata**

The eSciDoc system supports various metadata schema or profiles. In fact, the basic services take a metadata-agnostic approach. They accept every well-formed XML tree as metadata record.
6.4 Imag(in)ing the Shuilu Temple in China

In this two-year project a very special and rare Shuilu temple in China was digitised. Recordings were made in 2D (TIFF) and 3D (Flash VR). Data totals circa 1.5 TERABYTE. Digital recordings were annotated by scientists. The aim is to enter this data in FEDORA and make them available using Fedora’s image content model. However, this is not sufficient. For Fedora it is first necessary to:
- Adapt and develop various content models;
- Adapt and develop rich object to object relations;
- Develop standard and specific access mechanisms, in Fedora terms called disseminators, to show the right dynamic views, possibly using tools like Panorama viewer;
- Provide support to link to literature and annotations.

The aim is, among other things, to develop a number of standard content models and disseminators for Fedora.

6.5 SPECTRa: Chemistry Teaching and Research Data

The principal aim of the project Submission, Preservation and Exposure of Chemistry Teaching and Research Data (SPECTRa) was to facilitate the high-volume ingest and subsequent reuse of experimental data via institutional repositories, using the DSpace platform, by developing Open Source software tools which could easily be incorporated within chemists’ workflows. It focused on three distinct areas of chemistry research: synthetic organic chemistry, crystallography and computational chemistry. One of SPECTRa’s technology requirements was to find a suitable way of packaging data, i.e. to associate a number of data files together with some technical and descriptive metadata. The main alternatives were to use RDF, METS or MPEG21-DIDL. Of the three, METS was chosen because:
- It was the simplest technology that met the requirements;
- It had already been adopted by the eBank project;
- DSpace supports a METS profile as its primary package format.
6.6 StORE project: Source-to-Output Repositories

The StORE project seeks to address the area of interactions between output repositories of research publications and source repositories of primary research data. The functionality required by researchers in both types of repository is determined via user surveys identifying options for increasing the value of using primary data in source repositories as well as at the point where researchers submit papers to output repositories or download papers from them. This two-year project is multidisciplinary in scope, embracing seven scientific domains: archaeology, astronomy, biochemistry, biosciences, chemistry, physics and social sciences. Dublin Core is used as the metadata format and linking is achieved via bidirectional URLs between the publications and data sets.
7. Conclusion

An Enhanced Publication is a publication that is enhanced with research data as evidence of the research, extra materials to illustrate or to clarify or post-publication data like commentaries and ranking. Some publishers offer one or two of these categories with their official publications. Elsevier and Blackwell publications can have research data with the official publication, and Nature publications can have research data and also extra material, such as movies. It is surprising that Nature and BMC publish extra materials like movies separately from the publication, without even a link between these related objects. The reader of the article misses out on the movie, and the viewer of the movie remains unaware of the article. Only one publisher, PLoS, provides all these three enhancement services for their publications. In particular, the option to add post-publication data is rarely supported by other publishers. As a contrast, some repositories offer the option of adding commentaries, ranking, trackbacks, or linking publication and research data.

Digital scholarly objects on the Internet such as multimedia materials, data sets and blogs are difficult to find, but their number and importance is steadily growing. Universities and research institutes should therefore assume the responsibility of archiving the digital scholarly output of their organisation in a sustainable way. Repositories have an infrastructure that is very well suited for archiving scholarly output in national libraries or archives. In addition, the disciplinary repositories and the Dataverse Network should be considered.

The main conclusion is that publishers and repositories have the building blocks and the tools, but in general do not use them to create an Enhanced Publication for all three information categories. Publisher and repositories should offer the service and tools to add research data, extra materials and post-publication data to the publications. Researchers should be responsible for the content.

To make it easy for researchers to trace publications and related objects like research data, extra materials or post-publication data, an appropriate infrastructure should be developed. An Enhanced Publication model should support and reflect the relations between a publication and all relevant objects, e.g. data, Web sites,
commentaries. In this respect, an Enhanced Publication is developing permanently. The model must therefore be able to add related objects continually, also at a later stage. To handle this complex process, only the OAI-ORE model is sufficient.

The parts of an Enhanced Publication should be carefully selected. Publishers and repositories should set up a checklist for objects of an Enhanced Publication.

Mandatory:
- The object should have a unique global persistent identifier;
- The link must be resolved;
- All objects should have a timestamp;
- The file type should be common, for future use;
- Data sets should have a universal numeric fingerprint;
- ‘Cite as’ information should be provided;
- Is the quality good enough for preservation?
- Is it legal to publish the object? This question does not only concern copyright.

If known, the following information should be listed with the object:
- Is the object sustainable?
- Does the work have a mark of peer review?
- Is the object machine-readable?
- Is the object freely available?
- Who is the owner?
- Who is responsible?
- Is ranking information available?
- Are there comments available?
- Are there trackbacks available?
- Are there citations to the work? How many times has the material been cited?
- How many times has the material been downloaded?

Finally, the objects should be linked in a meaningful way. It would be helpful for everyone who wants to work with Enhanced Publications to set up general linking models for all publication types like book, dissertation, and article. The links must show what the relation of the object within the Enhanced Publication is. Is it a ‘part of’, like a chapter of a book, or a ‘comment on’ the publication? To provide basic semantic interoperability and also to support integration in a large range of different disciplines, ontologies should be used. It is up to the publisher
and repository holder to decide how complex the link pattern will be, and if the Enhanced Publication can change over time by adding commentaries after publication.

To keep scholarly publishing efficient and to maintain control over published materials, we need integration of all scientific information, with links between objects. Enhanced Publications can help to integrate scientific information, since they provide explicit links between related objects in order to directly show and support the relations between the objects. This link pattern of Enhanced Publications will help to structure the environment of scholarly publishing, and should therefore make scholarly publishing much more efficient.
Appendix: Identifiers and Identifier Resolution Services

A list with identifiers and identifier resolution services from the eBank Web site\(^{54}\), accessed March, 2008:

**Identifiers**

URI, URL, URN\(^{55}\)

HANDLE. The Handle System is a comprehensive system for assigning, managing, and resolving persistent identifiers, known as ‘handles’, for digital objects and other resources on the Internet. Handles can be used as Uniform Resource Names (URNs).

DOI. The Digital Object Identifier (DOI) is a system for identifying and exchanging intellectual property in the digital environment.

PURL. A PURL is a Persistent Uniform Resource Locator. Functionally, a PURL is a URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service.

POI. The PURL-based Object Identifier (POI) is a simple specification for resource identifiers based on the PURL system. The use of the POI is closely related to the use of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and with the OAI identifier format (OAI-identifiers) used within that protocol. The PURL-based Object Identifier (POI) Andy Powell, Jeff Young, Thom Hickey. http://www.ukoln.ac.uk/distributed-systems/poi/


\(^{54}\) http://www.ukoln.ac.uk/projects/ebank-uk/data-citation/

IVOA Identifiers Version 0.2 (IVOA Working Draft 30 September 2003)
http://www.ivoa.net/Documents/WD/Identifiers/WD-Identifiers-20030930.html

LSID (I3C) URN Namespace for Life Science Identifiers 4/03/03
http://www.i3c.org/wgr/ta/resources/lsid/docs/LSIDSyntax9-20-02.htm

http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm

Identifiers for learning objects - a discussion paper Andy Powell.
http://www.ukoln.ac.uk/distributed-systems/lo-identifiers/

http://www.doi.org/topics/020210_CSTI.pdf

NASA Astrophysics Data System Bibliographic Code (bibcodes).
http://adsdoc.harvard.edu/abs_doc/help_pages/data.html

**Identifier Resolution Services**

http://library.caltech.edu/openurl/Public_Comments.htm

Proposal for a Life Science Identifier Resolution scheme using Web Services
http://www.i3c.org/wgr/ta/resources/lsid/docs/LSIDResolution.htm

http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html

56 Missing on the eBank Web site: DANS resolver (National Dutch Resolver)
http://persistent-identifier.nl

http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html

http://www.dlib.org/dlib/july01/vandesompel/07vandesompel.html
References


