eBank UK: linking research data, scholarly communication and learning

Dr Liz Lyon, Rachel Heery, Monica Duke,

UKOLN, University of Bath

Dr Simon Coles, Dr Jeremy Frey, Prof. Michael Hursthouse,

School of Chemistry, University of Southampton

Dr Leslie Carr, Christopher Gutteridge,

School of Electronics and Computer Science, University of Southampton

Abstract

This paper includes an overview of the changing landscape of scholarly communication and describes outcomes from the innovative eBank UK project, which seeks to build links from e-research through to e-learning. As introduction, the scholarly knowledge cycle is described and the role of digital repositories and aggregator services in linking data-sets from Grid-enabled projects to e-prints through to peer-reviewed articles as resources in portals and Learning Management Systems, are assessed. The development outcomes from the eBank UK project are presented including the distributed information architecture, requirements for common ontologies, data models, metadata schema, open linking technologies, provenance and workflows. Some emerging challenges for the future are presented in conclusion.

1. Introduction and context – the scholarly knowledge cycle

The background and context for the eBank project is the changing landscape of scholarly communication afforded by the developing global network infrastructure (Internet, World Wide Web, Semantic Web and Grid) together with the potential for increasing links between the activities of e-research / e-science, digital libraries and e-learning at all levels from schools to post-graduate.

The flow of data and information in the process of research and learning in an academic setting leads ultimately to the creation and acquisition of knowledge. Data and information is continuously used and re-used in new ways as part of a research activity or in the development of course materials in support of learning. These workflows are interlinked and together underpin the scholarly knowledge cycle (figure 1) [SKC].

The creation, management, sharing and interoperation of original data (which may be, for example, numerical data generated by an experiment or a survey, or images captured as part of a clinical study together with the associated metadata) is an intrinsic part of e-Science, keeping data & metadata together. This process consists of one or more subprocesses which might include aggregation of experimental data, selection of a particular data subset, repetition of a laboratory experiment, statistical analysis or modelling of a set of data, manipulation of a molecular structure, annotation of a diagram or editing of a digital image, and which in turn generate modified datasets. This in turn entails managing version control and access to previous versions, as well as auditing information on processes and transformations.

The successive levels of derived data are closely related to the original data (through refinement, transformation, interpretation, generalisation *etc*.) and may themselves be further processed to form successive interpretations or abstractions. If the data at any particular stage of processing is considered to be generically useful, it may be considered for inclusion in a publicly-available database for use beyond the boundaries of the project which funded the data. This has only been undertaken in fairly specialised fields of joint community endeavour (for example in the bioinformatics field, protein and genome sequences databases, and in chemistry, material safety information, and the crystal structures Cambridge Structural

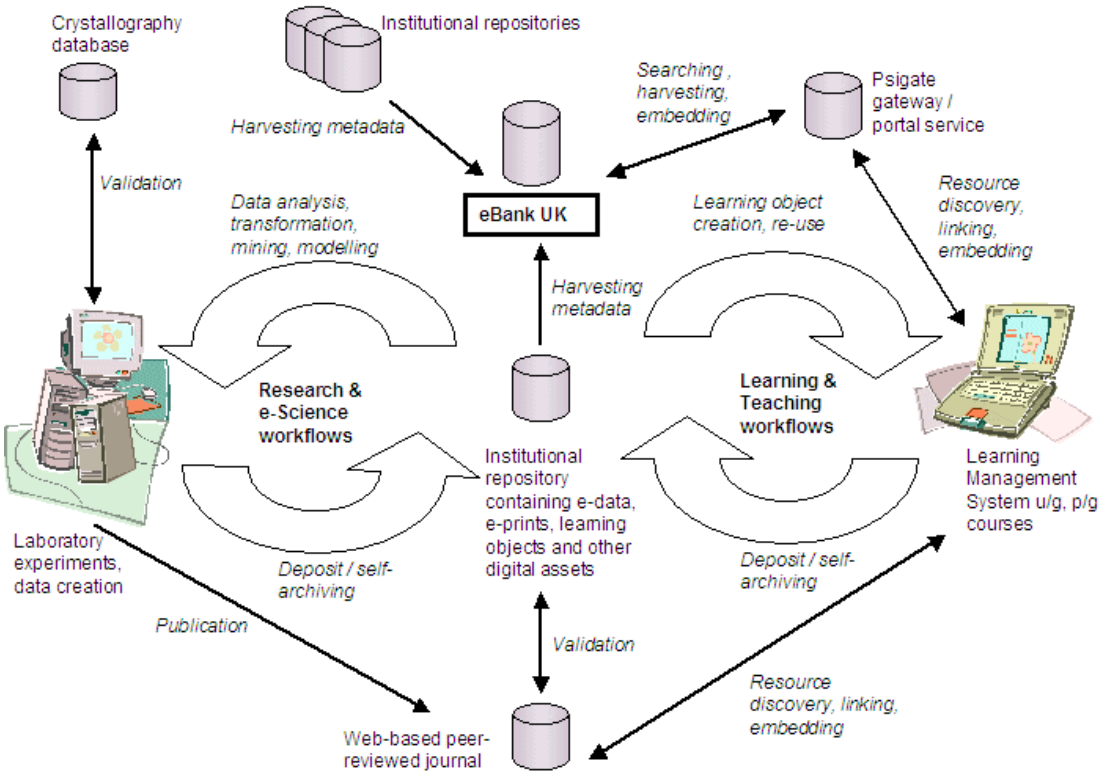

Figure 1: The Scholarly Knowledge Cycle

Database), as such databases require some level of service guarantee from a service provider. Provision in other chemistry areas *i.e.* on spectroscopic data is patchy due to the cost of maintaining the data as well as attitudes towards sharing.

Eventually, useful and interesting knowledge can be interpreted from the processed data; this knowledge is interpolated into a scientific article for communication with the scientific community through the publication process. Sufficient data must be included to support the arguments and claims advanced in the article, however the practical limits of the publication medium mean that the data is shown in a very restricted form (such as a graph or summary table). As well as submitting the paper for peer review and publication, it may be deposited in an eprint archive for immediate dissemination and discussion by the community.

These secondary items (the articles and database tables) can themselves re-enter the research cycle through a citation in a related paper or inclusion in a meta-study, or inclusion in postgraduate training or undergraduate learning activities by a reference in a reading list or in modular materials as part of an online course in a Learning Management System.

The eBank UK project is addressing this challenge of whole-lifecycle reuse by investigating the role of aggregator services in linking data-sets from Grid-enabled projects to e-prints contained in digital repositories through to peer-reviewed articles as resources in e-learning portals.

2. The eBank UK Project – addressing the data publication bottleneck

This innovative JISC-funded project which is led by UKOLN in partnership with the Universities of Southampton and Manchester, is seeking to build the links between e-research data, scholarly communication and other on-line sources. It is working in the chemistry domain with the EPSRC funded eScience testbed CombeChem [combechem], which is a pilot project that seeks to apply the Grid philosophy to integrate existing structure and property data sources into an information and knowledge environment. The specific examplar chosen from this subject area is that of crystallography as it has a strict workflow and produces data

that is rigidly formatted to an internationally accepted standard, that demonstrates the usefulness of an end-to-end philosophy, tracking from laboratory to literature and back again (figure 2).

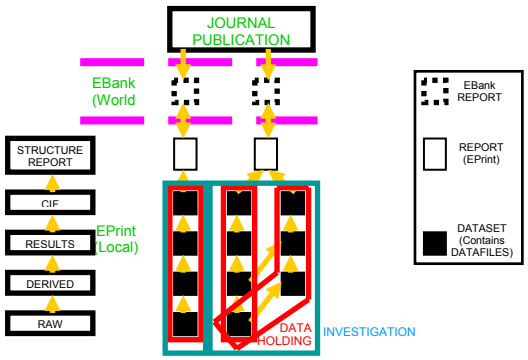

Figure 2. A generalised workflow for crystallography experiments.

The EPSRC National Crystallography Service (NCS) is housed in the School of Chemistry at the University of Southampton, and is an ideal case study due to its high sample throughput, state of the art instrumentation, expert personnel and profile in the academic chemistry community. Moreover, recent advances in crystallographic technology and computational resources have caused an explosion of crystallographic data, as shown by the recent exponential growth of the Crystal Structure Database (CSD). However, despite this rise it is commonly recognised that, via traditional publication routes, approximately only 20% of the data generated globally is reaching the public domain. This situation is even worse in the high throughput NCS scenario where approximately 15% of its output is disseminated, despite producing more than 60 peer reviewed journal articles per annum. With the advent of the eScience environment imminent, this problem can only get more severe.

The key issue is that the accepted unit of dissemination is the peer-reviewed journal article; the relationship between the article and the collection of data that underpins it (as made available in the published article itself) is weak, being expressed in reduced graphical or tabular form. Consequently not only is a small minority of scientific research being published in recognised outlets, but the amount of useful information being disseminated by those articles is a tiny fraction of the available (original) information.

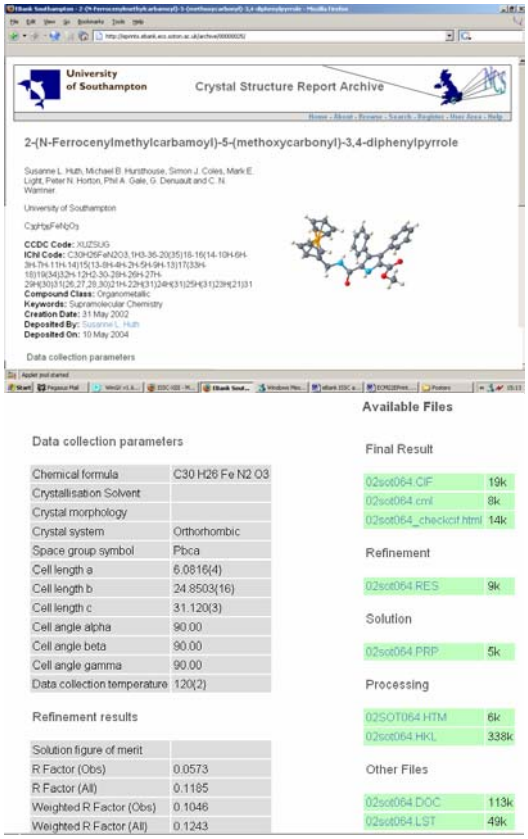

Figure 3. A crystallographic EPrint record, containing bibliographic information, an interactive visualisation of the derived structure, metadata and links to all the data generated during the course of the experiment.

To improve dissemination of published articles, the Open Access Initiative allows researchers to share metadata describing papers that they make available in 'institutional' or 'subject-based' repositories. Building on the OAI concept we present an institutional repository that also makes available all the raw, derived and results data from a crystallographic experiment that cannot be included in the peer-reviewed published article (figure 3). A schema for the crystallographic experiment has been devised that details crystallographic metadata items and is built on a generic schema for scientific experiments.

The deposition process generates metadata to be presented to the OAI interface by two different mechanisms. Firstly the depositor must enter metadata manually, which is mainly common bibliographic information but also includes chemical information items that have been added to the conventional Dublin Core

Name	Description of the stage	Files associated with this stage			Metadata associated with this stage	
		File	Type	Description	Name	Data Type
Initialisation	Mount new sample on diffractometer Parameterisation to set up data collection	datcol.non i*.kcd .drx	ASCII BINARY ASCII	Parameters for producing i*.kcd files Unit cell determination images Unit cell	Morphology	STRING
Collection	Collect Data	collect.rmat datcol.non .py s*.kcd *scan*.jpg	ASCII ASCII ASCII BINARY JPG	Orientation of the crystal Command file Proprietary configuration file Diffraction images Visual version of .kcd file	Instrument_Type Temperature Software_Name (x *n*) Software_Version (x *n*) Software_URL (x *n*)	STRING INTEGER STRING (x *n*) INTEGER (x *n*) URL (x *n*)
Processing	Process and correct images	scale_all.in scale_all.out .hkl .htm	ASCII ASCII ASCII HTML	Result of processing Result of correction on processed data Derived data set Report file	Cell_a Cell_b Cell_c Cell_alpha Cell_beta Cell_gamma Crystal_system Completeness Software_Name (x *n*) Software_Version (x *n*) Software_URL (x *n*)	INTEGER INTEGER INTEGER INTEGER INTEGER INTEGER STRING INTEGER STRING (x *n*) INTEGER (x *n*) URL (x *n*)
Solution	Solve Structures	.prp xs.lst	ASCII ASCII	Symmetry file, log of process Solution log file	Space_group Figure_of_merit Software_Name (x *n*) Software_Version (x *n*) Software_URL (x *n*)	STRING INTEGER STRING (x *n*) INTEGER (x *n*) URL (x *n*)
Refinement	Refine Structure	xl.lst .res	ASCII ASCII	Final refinement listing Output coordinates	R1_obs wR2_obs R1_all wR2_all Software_Name (x *n*) Software_Version (x *n*) Software_URL (x *n*)	INTEGER INTEGER INTEGER INTEGER STRING (x *n*) INTEGER (x *n*) URL (x *n*)
CIF	Produce CIF	.cif	ASCII	Final results		
Report	Generate e-Print report	.html	.HTML	Publication format (HTML/XHTML)	Authors Affiliations Formula Compound_name 2D_diagram	STRING STRING STRING STRING STRING

Figure 4. A representation of the crystallography experiment schema, indicating all the files generated during the course of the experiment.

schema. The chemical information includes, for example, an InChI code, which is an international identifier that uniquely describes the two dimensional structure.

During the deposition of data in a Crystallographic EPrint metadata items are seamlessly extracted and indexed for searching at the local archive level. The top level document includes 'Dublin Core bibliographic' and 'chemical identifier' metadata elements in an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [OAIPMH] compliant form, which allow access to a secondary level of searchable crystallographic metadata items (figure 4), which are in turn directly linked to the associated archived data. In this manner the output from a crystallographic experiment may be disseminated as 'data' in such a way that aggregator services and researchers may add value to it and transform it into knowledge and the publication bottleneck problem can be overcome.

3. Architecture and Data Flow

The metadata about the datasets available in the repository will be harvested by a central service using the OAI-PMH. This metadata will then be indexed together with any other available metadata about research publications. A searchable interface will enable users to discover datasets as well as related literature. The metadata contains links back to the datasets which the user will be able to follow in order to obtain access to the original data, when this is available. Harvested metadata from different repositories not only provides a common entry point to potentially disparate resources (such as datasets in dataset repositories and published literature which may reside elsewhere) but also offers the potential of enhancement of the metadata such as the addition of subject keywords to research datasets based on the knowledge of subject classification terms assigned to related publications. A further area of work investigates the embedding of the search interface within web sites, adopting their look-and-feel. The PSIgate[1] portal will be used to pilot these embedding techniques based on CGI-mechanisms and portal-related standards, particularly for resource discovery in a teaching and research training context.

During deposition of the crystallographic experiment data sets into the e-data repository, a number of descriptive elements are either entered by the depositor (figure 6), or generated automatically, as described previously. The repository is OAI-PMH compliant, such that it

[1] http://www.psigate.ac.uk/

Data Name	Data Description	Data Type	XML wrapped content
EPrint_type	'Crystal Structure'	String	Phrase 'Crystal Structure'
Creators	ePrint author(s)	String	ePrint authors 'Surname, Christian name, initial'
Affiliations	Institution(s) of creator(s)	String	Various authors addresses
Formula_empirical	Total atom count	String	Atom symbols with their total count (can be real
Title	IUPAC Chemical name	String	number) subscript
CCDC_Code	Cambridge Structural Database identifier	String	Chemical name with text & integers
Compound_class	Chemical category	String (set)	6 digit code
Available_data	Actual data available for various eData Report stages		1 word descriptor of chemical category
Related_publications	Other output relating to this compound/structure	String	Presence of data associated with all stages
Publication_date	Date of releasing eData REport to eBank/world	String	Literature reference link
			Date of public release of ePrint
Keywords	Undefined descriptors	String	Phrase describing chemical relevance
Structure_diagram	3D structure diagram	CML	Three dimensional structural diagram in CML format
		CML	Unique compound identifier (contains some structural information)

Figure 5. The schema elements presented to the OAI interface.

can respond to a number of (HTTP) requests made by 'harvesters'. Harvesters collect metadata from the e-data repositories and can pool metadata from distributed sources, thus performing the function of aggregators. Once an aggregator has harvested collections of metadata, it can take on the role of the 'service provider' (in OAI-PMH terms). eBank UK has developed a demonstrator of one such aggregator, or service provider.

One immediate benefit of the harvesting approach is that the service provider can now act as a single entry point for the discovery of resources that are distributed in disparate repositories. Once the metadata has been harvested by the eBank UK aggregator, an SGML indexing and search engine is used to provide the searching functionality over the metadata. The search results can be presented to the user through an HTML interface with links back to the data repositories where the original data resides. Any links made between different data sets or data sets and published literature by searching across the aggregated metadata are revealed to the user through the display of the search results.

Moreover, the Cheshire search engine [Cheshire] used in the eBank UK service provider makes it simple to make the data available through additional machine interfaces, such as the SOAP-based Search and Retrieve Web Service (SRW) [SRW]. This provides a basis for embedding the search service into external frameworks, such as portals; other mechanisms that could be employed include simple CGI-based mechanisms and portal protocols such as WSRP.

The harvesting approach overcomes the limitations of cross searching protocols, which are well-documented in the digital library literature [Powell, Lossau]. However, the knowledge that the aggregator possesses about the resources in the repositories is embodied in the metadata that has been harvested, and thus the services that it can offer are directly related to the quality, quantity and content of the metadata that is available (see section 4).

4. Meeting the interoperability challenge: design of an XML schema for harvesting.

The OAI-PMH supports selective harvesting of repositories, based on the partitioning of resources into 'sets' or on last updated date, so that cumulative harvesting of repositories can be performed. Data Exchange in the OAI-PMH is carried out in XML. Whilst the protocol specifies provision of the simple Dublin Core set of elements as a minimum interoperability requirement, the exchange of more specialised metadata sets is encouraged (see figure 5 for significant chemistry information used in eBank). The metadata, encoded in XML, is simply wrapped up in the OAI-PMH response wrappers (also encoded in XML). The alternative metadata formats can be specified in the requests by stating the 'metadata format' parameter required.

Experience within the ePrints community has shown that there are significant challenges to be met when building services on aggregated metadata [Guy]. Although a simple set of metadata elements like the Dublin Core goes some way to providing interoperability, guidelines are required to help reach some consensus on the specific interpretation of the use of those elements, for example by specifying an acceptable minimum of metadata to be completed, agreeing on the format for author names and encoding the identifiers of the resources described [Powell2].

In the eBank UK project we have made use of the extension mechanisms of the so-called qualified Dublin Core. We reuse some of the

basic 15 Dublin Core elements, such as creator and type [SimpleDC]. The identifier field contains a link to the HTML page in the edata repository that contains links to all the data sets related to an experiment. The metadata also contains references to the identifiers of individual sets of data within an experiment, to enable future links to be made should that data be referenced elsewhere. Vocabularies specific to the crystallography data are being used to describe the datasets types as well as the names and identifiers of the chemicals which are the

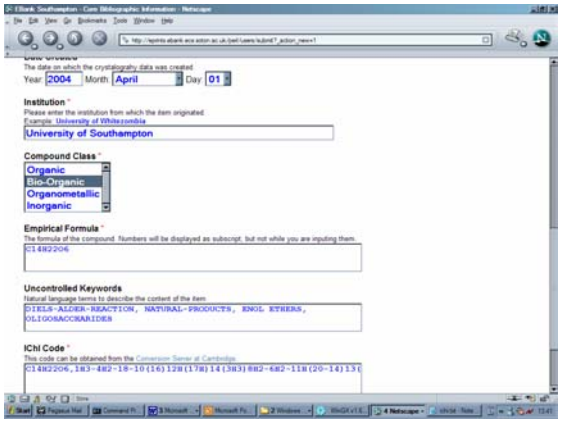

Figure 6. Entering metadata during the deposition process.

subject of the experimental data. The encoding of these vocabularies in the XML schemas has been developed in line with the recommendations of the Dublin Core Metadata Initiative [DCguidelines], and they can be easily substituted when standard vocabularies emerge within the Crystallography community. Further, the Dublin Core guidelines for encoding in XML recommend that "*Encoding schemes* should be implemented using the 'xsi:type' attribute of the XML element for the *property*." There are a number of guidelines for emergent standards in the chemistry community [Chemistry xml]; an example from our schema is the representation of the Chemical empirical formula in the following manner:

<dc:subject
xsi:type="ebankterms:empiricalFormula">C288 H200 Cl24 F48 O48 P16 Pd16</dc:subject>

There are a number of reasons for using the combination of Simple and Qualified Dublin Core. Firstly, the use of qualified DC should facilitate the mapping of the elements to simple Dublin Core, which as mentioned earlier, is a requisite for OAI-PMH compliant repositories. Secondly, by re-using and adhering to a well-documented standard it is hoped that the applicability of the approach will extend to other communities that the eBank UK project hopes to interact with in the future.

In principle, the Dublin Core fields used for 'bibliographic' information in the data set descriptions should work well with descriptions of published literature available through the OAI-PMH and which could be harvested and cross-searched in the eBank UK service. In practice the Open Access approach is not currently as widespread in the Chemistry community as in some other fields, and metadata available via the OAI-PMH that describes the published literature is rather scarce. However, some publishers have recently become more willing to expose bibliographic information on articles in journals that they publish more openly *i.e.* through specified open standards (if not the articles themselves) [announcement]. The eBank UK project has been pro-active in engaging the leading organisations involved in the quality control, publication and dissemination of crystallography data and it is hoped that the relevant publishers can be encouraged to release their metadata for interaction with the eBank UK service demo.

One further point that should be noted is that the current metadata exchanged to an extent hides a degree of complexity in the underlying experimental data being deposited by the crystallographers. Such data is often manifested as a series of related files that may have some underlying ordering, based, for example, on the stage of the experiment from which the data was produced. In this context, the eBank UK project is investigating the use of so-called 'packaging standards' which attempt to address the need to describe more complex objects. This is an issue which is increasingly becoming of interest to the OAI-PMH community [Herbert]. One other initiative which is currently seeking to increase access to scientific (climatic) data, has opted for a simplified approach of offering data (often in the region of terabytes in size) through one 'access' point identified by a single identifier [German]. On the other hand, the CCLRC [cclrc] is developing a model which attempts to make explicit the inherent complexity of scientific datasets, and the eBank UK project could contribute the experience gained to date to this effort.

5. Conclusions – issues and challenges for the future

This paper recounts the lessons learnt so far in the provision of experimental data together with the results of analysis linked to an e-print article and has assessed the requirements for future work and identified the key challenges that still need to be addressed. Finally, the immediate benefits to the learning & teaching and research communities in the scholarly environment are indicated and considered together with the potential for very significant impact in the longer term.

An immediate challenge for the project is to work with the crystallographic community to develop an internationally standard vocabulary for the exchange of structural chemistry data via OAI-PMH procedures. The eBank UK approach must then be promoted to crystallography experimentalists and sold to the publishing community as a whole in order to embed it into the scholarly learning and research cycles. At the same time the issues of maintenance of the software and archive along with preservation of the data may also be addressed.

The primary challenge for the future will be to generalise this approach to archiving and disseminating crystallographic data so that it may be employed in other areas of chemistry and also further to all the experimental sciences.

References

[announcement] Institute of Physics http://www.iop.org/news/0467j

[Chemistry xml] XML Data Dictionaries in Chemistry http://www.iupac.org/projects/2002/2002-022-1-024.html

[cclrc] Sufi, S., Matthews, B. & Kleese van Dam, K. (2003). An interdisciplinary model for the representation of scientific studies and associated data holdings. UK e-Science All Hands Meeting, Nottingham, 2-4 September 2003. http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/020.pdf

[Cheshire] Cheshire Web Page http://cheshire.berkeley.edu/

[combechem] Frey, J. G., Bradley, M., Essex, J. W., Hursthouse, M. B., Lewis, S. M., Luck, M. M., Moreau, L., De Roure, D. C., Surridge, M. and Welsh, A. (2003) Grid Computing – Making the Global Infrastructure a Reality, in Berman, F., Fox, G. and Hey, T., Eds. Wiley Series in Communications Networking and Distributed Systems, pages pp. 945-962. John Wiley and Sons.

[dcguidelines] Andy Powell and Pete Johnston (2003) Guidelines for implementing Dublin Core in XML http://www.dublincore.org/documents/2003/04/02/dc-xml-guidelines/

[eBank UK Project] eBank UK Project JISC/EPSRC E-Science http://www.ukoln.ac.uk/projects/ebank-uk/

[German] Publication and Citation of Scientific Primary Data DFG-Support Program: "Information-infrastructure of network-based scientific-cooperation and digital publication"
Project Page: http://www.std-doi.de/

[Guy] Marieke Guy and Andy Powell (2004) Improving the Quality of Metadata in Eprint Archives, Ariadne Issue 38 January-2004 http://www.ariadne.ac.uk/issue38/guy/intro.html

[Herbert] Jeroen Bekaert, Patrick Hochstenbach and Herbert Van de Sompel (2003) Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library. D-Lib Magazine. Vol 9 No 11. http://www.dlib.org/dlib/november03/bekaert/11bekaert.html

[Lossau] Norbert Lossau (2004) Search Engine Technology and Digital Libraries: Libraries Need to Discover the Academic Internet D-Lib Magazine. Volume 10 Number 6. June 2004 ISSN 1082-9873 http://www.dlib.org/dlib/june04/lossau/06lossau.html

[oai-PMH] The Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.html

[powell] Andy Powell (2001) An OAI Approach to Sharing Subject Gateway Content. Poster paper, 10th WWW

conference, Hong Kong. May 2001 http://www.rdn.ac.uk/publications/www10/oaiposter/

[powell2] Andy Powell, Michael Day and Peter Cliff (2003). Using simple Dublin Core to describe eprints. March 2003 http://www.rdn.ac.uk/projects/eprints-uk/docs/simpledc-guidelines/

[SimpleDC] Expressing Simple Dublin Core in XML http://www.dublincore.org/documents/2002/07/31/dcmes-xml/

[SKC] Liz Lyon. (2003) eBank UK – building the links between research data, scholarly communication and learning. Ariadne, Issue 36 http://www.ariadne.ac.uk/issue36/lyon/

[SRW] SRW Search/Retrieve Web Service http://www.loc.gov/z3950/agency/zing/srw/