Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 19-26

Metadata in the Context of The European Library Project

Theo van Veen
Koninklijke Bibliotheek, The Netherlands
Theo.vanVeen@kb.nl

Robina Clayphan
British Library, United Kingdom
Robina.Clayphan@bl.uk

Abstract

The European Library Project (TEL), sponsored by the European Commission, brings together 10 major European national libraries and library organisations to investigate the technical and policy issues involved in sharing digital resources. The objective of TEL is to set up a co-operative framework which will lead to a system for access to the major national and deposit collections in European national libraries. The scope of the project encompasses publisher relations and business models but this paper focuses on aspects of the more technical work in metadata development and the interoperability testbeds. The use of distributed Z39.50 searching in conjunction with HTTP/XML search functionality based on OAI protocol harvesting is outlined. The metadata development activity, which will result in a TEL application profile based on the Dublin Core Library Application Profile together with collection level description, is discussed. The concept of a metadata registry to allow the controlled evolution of the application profile to be inclusive of other cultural heritage institutions is also introduced.
Keywords: *European Digital Library, Interoperability, Dublin Core Metadata, Application Profiles, Collection Level Description, Search and Retrieve via URLs, SRU.*

1. Introduction

The European Library Project (TEL) [7] is partly funded by the European Commission as an accompanying measure under the cultural heritage applications area of *Key Action 3* of the *Information Society Technologies* (IST) research programme.

Co-ordinated by the British Library the project partners are:
Biblioteca Nacional, Portugal (BN)
Biblioteca Nazionale Centrale Firenze, Italy (BNCF)
Conference of European National Librarians (CENL)
Die Deutsche Bibliothek, Germany (DDB)
Helsingin Yliopiston Kirjasto, Finland (HUL)
Istituto Centrale per il Catalogo Unico, Italy (ICCU)
Koninklijke Bibliotheek, The Netherlands (KB)
Narodna in Universzitetna Knjiznica v Ljubljani, Slovenia (NUK)
Swiss National Library, Switzerland (SNL)

The objective of The European Library project is to set up a co-operative framework which will lead to a system for access to major European national and deposit collections. TEL will lay down the policy and develop the technical groundwork for the development of a pan-European digital library that is sustainable over time. The operational system will be implemented once the results of the project are known. Although the focus of the project will be on digital material as provided by the TEL-partners and publishers of digital material, traditional materials are not excluded.

This paper will discuss the development of a metadata model and the development of an interoperability testbed. This testbed will offer distributed searching in the national collections via Z39.50 alongside searching a central index of metadata harvested from other collections via the Open Archives Initiative protocol (OAI) [8]. This central index will be accessible directly via http. The design of the metadata model must enable current functionality and be open to future requirements with regard to the access of collections, digital objects and services.

The combination of distributed searching and central indexing and the use of two major search and retrieve protocols, Z39.50 and http/XML(SRU) - explained later in this paper, make the TEL project unique as similar projects usually use only one or the other access method.

2. The workpackages

The TEL-project consists of six workpackages:
1) Relation with publishers
2) Business plans and models
3) Metadata development
4) Interoperability testbeds
5) Dissemination and use
6) Management

This paper will focus on the more technical workpackages: workpackage 3, concerning the metadata development and workpackage 4, concerning the development of the interoperability testbeds.

These workpackages are interdependent: testbeds cannot work without the appropriate metadata and the metadata development needs an operational system for testing and developing the metadata models. It was therefore decided to work on both workpackages in parallel and for each to make use of the other's results in an iterative and incremental way. This meant that at the start of the project, for the http/XML testbed, any metadata format available in XML record syntax could be chosen. The results of the metadata development will be directed towards the operational TEL service and therefore do not have to be available until a later stage in the project. During the course of the project the work on metadata can use the http/XML testbed for the development of ideas and the data model can be brought in line with these ideas.

3. Metadata development

The various national libraries and publishers have different descriptive metadata formats. To access these different distributed sets of metadata a common datamodel will be developed. The data model will also support the functionality required within TEL thereby enabling data sharing amongst the TEL partners.

The TEL project aims at consensus building rather than delivering an operational service. Metadata arises from a functional analysis and an operational TEL service will probably reveal more functional requirements than we are currently aware of. The approach being followed is therefore directed towards identifying the functionality we can foresee and defining the metadata needed to support it. The metadata world is becoming more and more complex with an increasing number of standards (such as EAD, MARC, METS, MODS, RDF, DC, ONIX, CIMI, XML), so it will be a big challenge to develop a common datamodel that enables us to find, identify, select and access services from the individual members.

There appear to be two options. One is to convert all the partner's metadata into a single format. An alternative is to develop a metadata model that is generic and that can incorporate multiple metadata standards - the solution to this may be to introduce a TEL metadata-registry system.

At the outset it was agreed to use XML as the record syntax and unqualified Dublin Core as the temporary record schema to enable the test-bed development to proceed. The TEL metadata working group has since concluded that the DC-Library Application Profile (DC-Lib) [3], which is a combination of qualified Dublin Core and other namespaces, would be the best choice as a starting point for the datamodel for the operational TEL service.

4. The interoperability testbeds

The work on the interoperability testbeds will initially be focussed on the development of separate testbeds for http/XML and Z39.50, later in the project both will be brought together into one interoperability testbed. For Z39.50 it was agreed to conform to the Bath profile and this conformance will be the subject of testing. For http/XML there is not such a profile. A mapping is needed from user queries to Bath-conformant queries on one hand and the same user queries to http/XML queries on the other hand. A big challenge will eventually be the semantic interoperability between both testbeds.

There are two aspects to the http/XML testbed. First is the development of a mechanism to harvest records from contributing partners and, secondly, the specification and implementation of a protocol to make the data accessible by an external portal. For harvesting it was decided to use the OAI protocol.

At the same time as the specification of a protocol for search and retrieve was underway in TEL, the Z39.50 implementers group were working on the Z39.50 International Next Generation (ZiNG) [10] initiative. Under this umbrella two protocols for Search and Retrieve were initiated: Search and Retrieve via URLs (SRU) and Search and Retrieve via the Web (SRW). SRU uses the URL-GET method for its input; SRW makes use of the Simple Object Access Protocol (SOAP). Both protocols return XML as output and are similar with respect to request parameters, query language and the XML-output. As the original TEL specifications for the http/XML testbed were very close to the SRU specifications, it was decided to follow the SRU-standard for this testbed. It is likely that this will also be used in the final operational TEL-service.

Being one the earliest implementations of the SRU while it is still under development is quite an exciting aspect of the TEL-project.

5. Overview of infrastructure

An overview of the infrastructure is shown below. The TEL operational service will be a central portal and/or local portals. Separate portals will be used for

the two testbeds during development. Later in the project these testbeds will be combined to a central portal for the interoperability testing. In the overview this is illustrated by the ellipse around both portals. For the operational TEL service, when the integration of national services is sufficiently stable, the TEL-portal may be mirrored to local portals.

Five partners in the project will provide metadata in XML via the OAI-protocol. These records will be indexed in a central index. The portal will search and retrieve them via the SRU-protocol. The databases of four other partners will be accessible via Z39.50. The metadata will offer links to the digital objects and services. These links will be either direct or indirect via link services using OpenURL's or URN-resolvers. Additional services might be offered, for example multi-linguality (translation of specific subject headings) or thesaurus services allowing the use of search results from a thesaurus database as input for subsequent searches in TEL.

6. The approach to metadata development

The first stage in this workpackage consisted of a review of the partner's use of metadata. This "state of the art" report was based on a survey of current practice and desk research. Following that a metadata working group was installed comprising members from each participating library. Analysis of the state of art review resulted in the decision to define the metadata requirements by analysing the functionality required for TEL, and then determining what metadata elements were needed to fulfil those requirements.

6.1. State of the art review

Analysis of the responses to the metadata questionnaire produced five main conclusions:

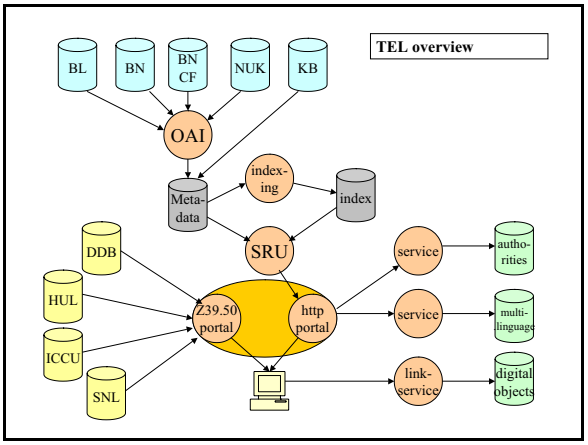

Figure 1. Overview of TEL infrastructure

1. There is no consensus between the partners about categories of metadata. Partners have interpreted the categories differently according to the scope and purpose of their implementations. This is illustrated in the following diagram, which shows how the partners defined different categories of metadata.

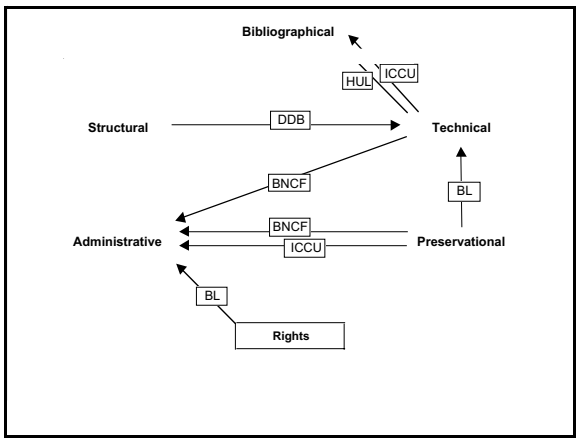

Figure 2. Overview of the differences in terminology of metadata

2. Libraries need to share knowledge on the creation of metadata, especially for collections that will be described in the future.
3. The absence of a common bibliographic format makes simultaneous access to metadata from different partners difficult. Formats in use are:
   - MARC21
   - Finmarc [1]
   - Dublin Core
   - UNIMARC
   - Pica3
   - PicaPlus
   - COMARC
   - Custom built datamodels

   The custom built datamodels in particular cause a problem: there are many of them and there are no generally available mappings for them. It is expected that the use of Dublin Core (or DC-Lib) will make it easier to develop consistent resource discovery services.
4. There is not yet one linking mechanism or resolution service used by all partners. Research is needed on the use of metadata for linking to external resources by means of URNs, PURLs or OpenURL.
5. There is uncertainty about the eventual contents of TEL and a need to be aware of the danger of an unbalanced service that may render small collec-

[1] Finmarc has in the meantime been replaced by MARC21.

tions invisible amongst complete union catalogues.

6.2. Developing the TEL datamodel

It was considered very desirable to adopt existing metadata standards where possible. The DC-Library Application Profile was identified as the most obvious candidate to form the basis of the TEL datamodel. The general approach has been to define the functionality needed to underpin the services and publication types as currently envisaged on the one hand and identify the metadata required to enable that functionality on the other. Special attention will be paid to digital collections and collection level descriptions (CLD).

The functional requirements could then be analysed against DC-Lib to see what gaps existed in the element set. From these results we can determine whether DC-Lib is sufficient for TEL or whether it will be necessary to define a TEL Application Profile which will incorporate DC-Lib with additional elements. Another possibility is to request the DC-Lib drafting committee to incorporate the additional metadata elements into DC-Lib, but the timescales of the project may not allow this.

Finally we need to determine the best mapping between the TEL Application Profile and the partners various metadata schemas and bibliographic formats. Desk research and experimentation with actual data from the different partners will determine how to implement the application profile in XML. The results will become part of the TEL metadata handbook, which will be made available on the web to facilitate the introduction of new collections to TEL.

It is not envisaged that TEL will stop at the end of the project but will continue to evolve afterwards. The impact of this is that the development effort will not be solely focussed on the TEL test-bed, but a more generic approach will be followed. This will allow future TEL functionality to be taken into account. It also raises the possibility of another approach to metadata specification. This option is to create a TEL registry of metadata that would allow the addition of new metadata elements that are in use by the partners. This possibility is discussed later in this paper but at the time of writing this approach has not been fully discussed within the project.

7. Functionality and services

The analysis of functionality and services was the result of desk research and included mapping functions to metadata elements. Functions considered relevant for TEL were put on the horizontal axis of a matrix and the metadata elements were put on the vertical axis. Functions refer to TEL as a whole and not solely to the TEL-portal. The elements were those from draft Library Application Profile of 2001-10-12. The complete overview is contained in a project report which is not publicly available at the time of writing. The mapping was intended to highlight any gaps where a function could not be supported by the available metadata.

The main functions are:

- *Search and resource discovery*
  This is fundamental functionality. Most metadata elements contribute to this function.
- *Record retrieval*
  Record retrieval follows from a search and also plays a role in harvesting and indexing metadata. Metadata elements for the identification of the original record is required as is specification of the record format.
- *Identification of resources*
  Needed to find and access resources. All elements used for referencing may play a role.
- *Description*
  Many metadata elements help the user in a decision to obtain the object.
- *Linking services*
  Linking services help locating objects or services. In many case these are resolution services. All metadata elements that play a role in dynamic linking are relevant.
- *Multilinguality service*
  Envisaged as a service translating user input or returned metadata into different languages to create new queries or search terms. Most textual metadata may play a role in this.
- *Thesaurus service*
  Envisaged as a service to find main entries for subjects and classification from user input or returned metadata to create new queries. Textual metadata or classification codes may play a role.
- *Collection level services*
  The functionality that helps to find and identify collections, link to those collections or broadcast queries to distributed collections.
- *Authorisation*
  Access may depend on the service, type of publication, year of publication, publisher, the user etc. Restrictions are indicated by terms and conditions or access rights.
- *Administration*
  Functions that keep track of usage, based on, for example, subject or publisher.
- *Hard and software requirements*
  Specific metadata to inform users of the requirements on their workstations or detect whether the users workstation is capable of accessing a publication type. Especially when preservation activities play a role.
- *Navigation*
  This functionality concerns linking related metadata records by dynamic linking, for example tracking hierarchical relationships like journal-issue-article or expression-work-manifestation etc.

Tabel 1. Mapping of metadata elements to functions by their usage

Element	Qualifier/ Scheme/ Role	Search/resource discovery	Retrieval of metadata	Identification	Description	Link service	Multilinguality	Thesaurus service	Collection level	Authorisation	Administration	Hard and software.	Navigation	Copy cataloguing	Miscellaneous	Comment
Record Identifie	Any		X	X		X				X					X	To identify the metadata record
Identifier, Source, Relation	Base-URL		X	?		X			X				X		X	Encoding scheme for a UR with variable query
Identifier, Source Relation	URN			X		X			X				X		X	Encoding scheme foURNs
Identifier	PURL			X		X			X				X		X	Encoding schema for persistent URLs
Identifier, Source, Relation	OpenURL			X		X			X				X		X	Encoding scheme for queri with a variable baseURL
Collection level description Schema	All	X			X				X	X	X		X	?	?	Possibly RSLP or the DC schema.

- *Copy cataloguing*
  Metadata may be re-used by other libraries for cataloguing.
- *Miscellaneous*
  Off-line ordering, ILL and other services. Mostly accessed directly via URLs. Link services are anticipated for TEL. Metadata regarding holding or item information and identification of the original metadata record are most important for TEL.

Mapping the functions and services listed above to the draft DC-Lib of 2001-10-12 some functions can be seen to need additional metadata elements or encoding schemes. This is shown in the table below.

The above table shows the metadata elements or qualifiers that are not present in DC-Lib, but will be needed for some of the required functionality. This will be discussed below. It should be noted that there will be many more metadata elements that will be useful or even necessary to search and access the digital objects from specific collections. To identify these metadata elements the specific collections will have to be examined. How these can be handled is discussed in the Registry section of this paper.

One aspect that still needs special attention, but which is not yet covered in this paper, is the sophistication of search functionality based on the semantic relations between metadata – as described in "The ABC Ontology and Model" [5] for example. The complexity and the human effort needed to create records that support queries based on these more complex semantics are expected to be rather high. This aspect of the functionality will therefore be addressed separately from the basic questions regarding which metadata are needed.

A further aspect of the relationship between functionality and metadata concerns which fields (access points or indexes) that can be searched. All metadata elements are – as long is it is reasonable – implicitly considered to be searchable and a one to one relation between search field and metadata element is assumed.

7.1. Metadata for linking

Most of the metadata elements, that need special attention, have to do with linking. Identification of the metadata record is relevant for TEL in order to maintain the reference to the original records for harvesting purposes and when record identifications are used in dynamic URLs (linking by reference).

TEL has to deal with metadata that should, as far as possible, be independent of the publication or service that is described in the metadata record. In other words, the portal should not have interpret the content of elements but simply act on the rules governing the type of metadata. For the identifier element these rules will generally be different for OpenURL, URLs, URNs and PURLs. The dynamic and context sensitive creation of links in which special link services or resolution services will be involved, will be different for these types of identifier. Using only URI as identifier encoding scheme for linking in DC-Lib will therefore not be sufficient. This also concerns the source and relation elements.

A special type of encoding scheme is the base-URL. This base-URL identifies a collection and will be used in generating URLs representing queries into such a collection (deep linking). This is different from the conventional URL for accessing the website of a collection.

7.2. Collection level descriptions

Collection level descriptions have a place in the TEL metadata set as TEL is essentially a collection of collections. To describe collections, metadata elements are needed that are not used in the description of conventional publication types. Any aggregation of objects could be considered as a collection and a collection can simply be considered as a top level container of records. An important aspect of collections is the way they are accessed: some collections can be searched individually and others are simply a static website.

In TEL there are two different ways to look at collections: 1) they can be considered to be a publication like any other, but being of a specific type, 2) they can be considered as an aggregation of publications. In the latter case a collection may be a target for distributed queries. These two aspects of collections give rise to a potentially very powerful future functionality: they allow the user to find collections as the result of a search and then select these collections as the list of targets for a next – more precise – distributed search.

The importance of collection level descriptions is such that it justifies a complete new set of metadata elements. The resource implications of discussing each individual metadata element for a collection level description within TEL would be onerous. In line with the principle adopting existing (or developing) standards, TEL will utilise an existing CLD

schema such as that developed by RSLP [6]. The DC Collections working group [2] is also considering the RSLP schema. After further work it is anticipated that TEL will include the complete schema for collection level descriptions in its own metadata set, therefore in the functionality matrix no individual metadata elements for collection level elements are shown.

8. TEL metadata registry and metadata formats

Although DC-Lib was identified as a valuable starting point for the TEL application profile it does not contain all the elements that TEL will need. Even if it would suffice for now it will not be sufficient in the future when new functionality is introduced. We therefore need to create a TEL application profile which may contain elements that are not part of a DCMI namespace at the moment.

As seen now, The European Library project is a system for access to all types of collections and materials owned by the European Libraries. In the future it may be opened to other types of memory institutions and, if so, the issue of semantic interoperability will become an important aspect of the development. The flexible structure of Dublin Core and the different sectoral interpretations of how to describe a digital object could be an obstacle to interoperability. In Dublin Core there are no rules of the kind we employ in libraries for how the values in metadata are constructed and a unitary search does therefore not guarantee the localisation of all types of digital objects.

In the creation of a TEL profile based on DC-Lib it is also important to define a model and an ontology as a starting point for the development of vocabularies relating to different applications in the Cultural Heritage sector. Libraries own and catalogue materials that are also owned and catalogued in different types of institution (Archives and Museums for example). It is important to define the correspondence between terms, functions and concepts in the systems describing that material.

There are may be several ways of dealing with this:

1. Promote a common standard schema, independent of what different data providers are using internally. This would entail the definition of one comprehensive mapping table using which all information providers could convert their metadata to a single TEL-schema.
2. Introduce a TEL metadata registry that contains metadata from existing metadata standards, but which can be extended with local metadata. The Library application profile will be the "main" entry, but as soon as new metadata are introduced for which there is no existing element in DC-Lib application profile, or other profiles accepted by TEL, then the introduction of new elements would be allowed. In this context, the TEL Registry is seen as a system that facilitates the procedures involved in allowing the TEL application profile to evolve in line with increased functionality and extension to different types of institution. In this it is slightly different from the concept of a registry in the sense it is currently used in DCMI [9].
3. Use the TEL indexes mainly for the first FRBR [4] objective i. e. to find all resources sharing the same index entry and – for the other objectives – the user will be redirected towards the real catalogs. In this context, Dublin Core presents itself as a metadata pidgin for digital tourists who must find their way in this linguistically diverse landscape [1].

The first option is preferred but we may need the second option to realise the first one: the registry will define a common standard schema, but building the schema in a decentralised and incremental way is enabled. Data providers would be allowed to add new metadata elements but at the same time all providers can monitor the developing schema and raise objections to inappropriate metadata elements. The third option is a last resort option for very specific cases to provide metadata elements from the original record for which there are no corresponding metadata elements in the registry.

The TEL portal would use the metadata elements from this schema/registry for such actions as display, translate, generate a link or generate a new search. This TEL registry would therefore contain information additional to the basic ontology on how TEL will handle these metadata.

The registry will contain at least:

- element name
- name space
- originating metadata standard
- labels for presentation in different languages
- flag to indicate that it should be presented in the full presentation
- flag to indicate whether the element should be used as clickable link for searching
- element name that it maps to (in case it comes unconverted from another metadata standard)

This list may grow in the future as the usage of different metadata elements in practical situations is extended. When new collections from national libraries enter the TEL-system the flexibility of such a registry will facilitate compliance to the TEL system.

9. Implementation

In the figure below an overview is shown of the steps involved in exchanging metadata.

There are several formats involved here, that can also differ per project partner. That is:

1) XML, this can conform to any element set but the

Library application profile is preferred
2) TEL-XML, which is defined by the TEL-project (DC-Lib)
3) MARC
4) HTML meant for the user

For Z39.50 the local data will be provided via the local Z39.50 servers in TEL-XML but in case that is not supported, MARC is also accepted. For the http/XML sources the SRU server will perform some conversions if the original data is not conform DC-Lib.

In the TEL portal the final conversions will take place to provide the user presentation. This conversion makes use of the metadata registry in which it is defined how the metadata should be processed.

In the above overview there is an additional Z39.50 gateway, which could transform the Z39.50 output records to TEL-XML. This is not implemented or even planned but is just an extra possibility that would allow the portal to deal with one single XML format instead of XML and MARC.

As there were no portals available that supported distributed search conforming to the SRU protocol at the start of the project, a test-portal was developed for this purpose. This test-portal is based on XSL and javascript and runs locally in the browser. It was found to be quite convenient to have everything at one place during the test phase and therefore the metadata registry is initially implemented in tables in the same javascript that makes up the local portal.

10. Handbook

A TEL metadata handbook will be made available to help the partners in submitting collections and new metadata elements to TEL. This handbook will contain information on metadata mappings, conversion schemes, standards, relevant links and TEL requirements with respect to metadata. It will help the TEL partners to define an ontology in developing an integrated glossary (or specific vocabularies) related to the different systems and services.

11. Conclusion

Projects working to limited timescales in the rapidly changing world of digital resource sharing have to adopt a pragmatic approach to development activity. The parallel development of the testbeds and the metadata schema has allowed work to proceed on both these aspects of TEL without delay to either activity and will allow the development of each to incorporate the findings of the other. Adoption of emerging standards minimises the need to spend time developing customised solutions and should assist wider interoperability as well as adding an element of future-proofing to the resulting system. In

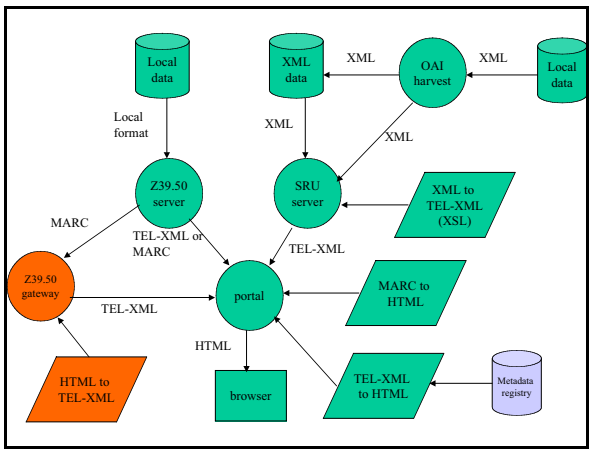

Figure 3. Metadata formats in different processing stages

using current developments in Z39.50 and Dublin Core TEL is confident that it has chosen well-founded namespaces with which to work in this respect. Resource sharing across nations necessitates working with heterogeneous collections – the use of OAI harvesting and central indexing of XML records facilitates integrated access to these. Looking to the future means developing systems that are open to evolution – the proposed metadata registry is an attempt to build more openness into the system.

Although the current TEL project will not result in a fully operational European Library system, the results of the project will constitute the groundwork on which that ambitious vision can be realized.

References

[1] Baker, Thomas, A Grammar of Dublin Core, 2000, http://www.dlib.org/dlib/october00/baker/10baker.hml

[2] DC Collections Working Group http://dublincore.org/groups/collections/

[3] DCMI Libraries Working Group http://www.dublincore.org/groups/libraries/

[4] Functional Requirements for Bibliographic Records, International Federation of Library Associations and Institutions, March 1998, http://www.ifla.org/VII/s13/frbr/frbr.pdf

[5] Lagoze, Carl and Hunter Jane, The ABC Ontology Model, in Journal of Digital Information, 2001 http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/

[6] Powell, A., RSLP Collection Description Schema http://www.ukoln.ac.uk/metadata/rslp/schema/

[7] The European Library
http://www.europeanlibrary.org/

[8] The Open Archive Protocol http://www.openarchives.org/OAI/openarchivesprotocol.htm

[9] The Open Metadata Registry http://wip.dublincore.org:8080/registry/Registry

[10] Z39.50 international Next Generation
http://www.loc.gov/z3950/agency/zing/