Common presentation of data from archives, libraries and museums in Denmark

Leif Andresen
Danish Library Agency
October 2007

Introduction

In 2003 the Danish Ministry of Culture entrusted the three national authorities for archives, libraries and museums to develop recommendations for data content, data formats and data transport. The goal was to facilitate presentation of joint information from archive, library and museum sectors (in the following ALM-sectors) for the public on the Internet. The initiative was in prolongation of ongoing local/regional projects.

Over the last years there has been an increased focus on co-operation between archives, libraries and museums – particularly in local and regional contexts. This was the reason for setting up the Danish ALM standard committee appointed by the three national authorities for the ALM sectors. The steering group consisted of the heads of the three national authorities and the work has been done by a working group with six members, two from each sector.

In February 2006 a report was completed by the committee. Dublin Core is recommended as the common format for content information to which selected parts of individual databases can be mapped. The selection is based on relevance of data according to search and presentation.

As a tool for interoperability a XML schema is composed collecting the XML schemas of the mentioned content formats. The name of this schema is *dkabm* – *dk* for Denmark, *a* for archives, *b* for libraries (biblioteker) and *m* for museums. The 15 basis DC elements are complemented with DC refinements, Administrative Components and supplements from the developed dkdcplus schema. XML is recommended as exchange format.

After a public hearing of the report, the final set of specifications was approved the heads of the three national authorities 15 October 2007. The specifications are publicly available (in Danish!) at: http://www.bs.dk/standards/abm/

Introductory considerations

The working group limited the requirements to:

- only data relevant to the public for search and presentation should be converted to the common format
- the requirement for representing details is not the same in a common database as in the sector specific ALM databases. But it is important for the user of the common data base to get a pointer (normally a link) to the original database, e.g. to retrieve more information, to ask for a copy or to send a loan request.

The working group didn't find alternatives to Dublin Core. But the group realized that mapping from more complex structures as used in museum and archive databases was not simple and that some information is lost in the translation.

Development of a Common content format

The basis for mapping to Dublin Core is the ALM-sectors' specific formats for registration of collections. These formats are developed by the individual sectors partly based on - or inspired by - international standards. A consequence of this is different traditions for the selection of registration units as well as variances in registration levels, which can cause problems when converting to a common format.

For archives and museums the unit of registration is often done on a collection level, which means that one registration contains several units. The coherence of a collection can be due to many different circumstances: same excavation, same donor or just collected in the past by somebody. For libraries the collections are normally registered on a document level, describing identical units. An exception is multilevel published books.

The challenge was to map this hierarchical structure to the flat Dublin Core structure with a workable result for search and presentation.

One of the original goals for Dublin Core as metadata format was to support 'Resource discovery'. In the development of mapping, support of the sector-specific functionality has not been in focus. The focus has merely been on identifying the existence of an object. The consequence is a relative simplicity in the mapped format. To exchange data with a common system will be easy notwithstanding inequalities between the original registration formats. Another consequence can be that the users of the common system can encounter problems with interpretation of data because the original context is missing.

Mapping schemes to Dublin Core has been developed for four domain formats:

- *Daisy* for governmental archives
- *Arkibas 4* for local archives
- *danMARC2* for libraries
- *Regin* for museums

To ensure a fundamental level of retrieval in the system, it is important that a basic part of the original semantic retains are preserved. This has been an important parameter for evaluation of what kind a data have needed to use refinements to the fifteen basic Dublin Core elements.

One way to handle this problem is to add information about the source of the information. When the common system handles a record, the source of the record can be used to improve the presentation of data. To meet this requirement together with the requirement of linking to the original registration it is necessary to go outside Dublin Core. These two data elements are not descriptions of the resource, but information about registration of the resource – and then outside the scope of Dublin Core. The AC - Administrative Components was employed to meet these kinds of requirements.

Illustration of the dissimilarity based on source of the registration data:

- Using DC.Creator can be different for different kinds of originators. For museums creator can be the one responsible for composition of a museum's file. For a library the creator can be the author of a book, the composer of music or the band playing on a cd. For an archive the creator will often be the institution or part of the institution who established the archive. To make it possible to present the Creator-information in the right context it is necessary to use some refinements of Creator to preserve what kind of responsibility the described creators have. These refinements are described in part 5.
- Also using DC.Title will reflect different kind of titles. For libraries a title will normally be mandatory for all documents, but for archives and museums a title will often be constructed only for the common database – normally the source system does not contain a title in the same sense as a bibliographic title.

When trying to convert data for DC.Type and DC.Format some inconsistency must be anticipated. The different vocabularies seem not to cause problems because they reflect heterogeneous appearance forms.

Some of the basic fifteen elements are only used for libraries. DC.Publisher and DC.Language are not relevant as target for conversion because the Danish original databases for archives and museums do not contain these kinds of data. The lack of use for these elements is not assessed to give any problems for the use of a common database because they only exist for libraries.

In the model for the mapping of data from the different ALM-systems two levels of description are defined: a level for registration of collections and a level for registration of individual units. This way of splitting up into levels was chosen because archives and museums use these levels in their original registration. The different level is reflected in DC.Type using Collection for the first level and for museums normally Physical Object.

The combined schema

Used elements for exchange of ALM data for shared presentation:

Element **- refinement**	**Namespace**
DC.Title	dc
- alternative	dcterms
DC.Creator	dc
- preferredName	dkdcplus
- alternativeName	dkdcplus
- actPeriod	dkdcplus
DC.Subject	dc
DC.Description	dc
- version	dkdcplus

DC.Publisher	dc
DC.Contributor	dc
DC.Date	dc
DC.Type	dc
DC.Format	dc
- extent	dcterms
- medium	dcterms
DC.Identifier	dc
DC.Source	dc
DC.Language	dc
DC.Relation	dc
- isReplacedBy	dcterms
- replaces	dcterms
- isReferencedBy	dcterms
- references	dcterms
- isPartOf	dcterms
- hasPart	dcterms
DC.Coverage	dc
- spatial	dcterms
- temporal	dcterms
DC.Rights	dc
AC.Identifier	ac
AC.Source	ac
AC.Location	ac

The scheme Period from dcterms is used for ActPeriod and Temporal. A list of values SubjectType is used for DC.Subject. This list include the values DK5 (library classification), DBCS, DBCF and DBCM (library keywords) and SRKM (museums classification) in dkdcplus.

XML Schemas

dkabm.xsd
http://www.bs.dk/standards/schemas/dkabm_2006-11-23.xsd
packs dkabm-records into a file.

dkabm_types.xsd collecting schemas mentioned below.
http://www.bs.dk/standards/schemas/dkabm_types_2006-11-23.xsd

dc.xsd defines the fifteen basic Dublin Core elements.
http://purl.org/dc/elements/1.1/dc.xsd

ac.xsd defines Administrative Components.
http://www.bs.dk/standards/schemas/ac_2005-09-01.xsd

dkdcplus.xsd defines Danish elements and subject lists.
http://www.bs.dk/standards/schemas/dkdcplus_2006-11-24.xsd

dcterms_ext.xsd import Danish elements from dkdcplus.xsd together with dcterms and dc.
http://www.bs.dk/standards/schemas/dcterms_ext_2006-01-13.xsd

dcterms.xsd defines Dublin Core refinements.
http://purl.org/dc/terms/dcterms.xsd

dcmitype.xsd defines Dublin Core resource types.
http://purl.org/dc/dcmitype/dcmitype.xsd

ISO639-2.xsd defines valid language codes.
http://it.ojp.gov/jxdm/iso_639-2b/1.0/iso_639-2b.xsd

Exchange of data

The working group recommended XML as exchange format and chose the DCMI XML schema. This is also in accordance with recommendations from the Danish *National IT and Telecom Agency* attached to *Ministry of Science, Research and Innovation*. Concerning data transport, the working group found it important not to freeze to a particular model. The recommendations for content and exchange are not linked to a specific model.

Data transport can be handled in two different ways:

- search in the specific databases and establishing common presentation on the fly, including conversion on the fly
- search in one common data base, which implies data exchange beforehand

The conversion on the fly implies several problems. The defined conversions need to be converted including change of data (e.g. from code to text) and selecting only parts of databases. These kinds of requirements to be implemented on running systems are much more complex than regular export of data. The working group recommended a common database, but the metadata schema and XML schema work for a distributed model too.

The specification includes guidelines for exchange based on harvesting (OAI-PMH), search (SRU) and file transport (mail/ftp).

Mapping from ALM domains

Mapping from museums

The Regin data model consists of *primary entities* describing objects or concepts in the museums' domain and *secondary entities* describing properties and aspects of the primary entities. Furthermore the model consists of relation tables describing various connections between entities.

In the mapping most of the primary entities have been mapped to Dublin Core records:

- Archive file

- Magnetic mediums
- Case
- Photo
- Ships
- Objects
- Literature
- Reports
- Big formats.

The primary entities Case and Archive file are mapped to collection level and all the others to item level. The secondary entities are mapped as attributes to the above primary entities and together with two primary entities (Artist and Player).

Mapping from governmental archives
The Daisy data model used for governmental archives is a relational model and consists of three main parts:

- Agents
- Heuristic units
- Archive store units.

The occurrence of a heuristic unit causes a Dublin Core record with the connected agents as DC.Creator's. The Daisy system doesn't contain keywords. To ensure reasonable possibilities for searching the names of the heuristic units are also mapped to DC.Subject and to DC.Title. Because the name of an agent often reflects a geographic area Agent is also mapped to Spatial as refinement of DC.Coverage.

Mapping from local archives
The format for the local archives is record-based and the mapping is from an Arkibas record to a Dublin Core record. The originator of an archive file is mapped to a DC.Creator refinement.

Mapping from libraries
Bibliographic records in danMARC2 are mapped to Dublin Core records. MARC records are supposed to be well known. Note that a top level record in a multilevel record is mapped to collection level.

Underlying metadata formats

The Common content format dkabm consist of Dublin Core – namespace *dc* – and DC refinements – namespace *dcterms* – supplemented by *ac* and *dkdcplus*.

AC - Administrative Components
To handle connection between the original data in the ALM databases and the resulting Dublin Core records it has been vital to register this information not about the described resources but about the metadata itself. This kind of data is out of scope for Dublin Core.

To solve this problem the metadata schema for administrative information about metadata Administrative Components was selected.

Three of the AC elements are used:

AC.Identifier	Identification in original system. Can be used for linking to all information
AC.Source	Identify the delivering organisation/institution
AC.Location	An unambiguous reference to the content metadata within a given context

dkdcplus

To handle the additions to DC and AC namespaces was defined a namespace and a XML schema called *dkdcplus*.

The specific Danish supplement contains three refinements for Creator, one refinement for Description and a list of values for Subject and a Danish-language version of DCMI type.

For Creator *preferredName* is used for the agreed version of a name. Several institutions have over time many names and variations of names and for archives it is suitable to collect documents from the same institution – also in presentation together with libraries and museums. As a logical consequence *alternativeName* is used for other variations and actPeriod to delimit the 'on duty' period.

For Version/Edition the element *version* is defined as a refinement of Description. It is important for description of library books to distinguish between different editions, and it is no less important in presentation together with archives and museums.

A reason for making a schema dkdcplus is to establish a Danish reference point for extensions to existing metadata schemas to ensure the interoperability between Danish metadata systems.

Leif Andresen
Danish Library Agency
Tel: +45 3373 3373
lea@bs.dk