



# From content standard to RDF

Gordon Dunsire

15. seminar Arhivi, Knjižnice, Muzeji  
23-25 Nov 2011, Poreć, Croatia

Published by Gordon Dunsire

Edinburgh 2012

# From content standard to RDF

---

*Gordon Dunsire*

## Abstract

The paper describes the functions and components of cataloguing codes as content standards, and of the Dublin Core abstract model and Dublin Core application profile. The components of a content standard are mapped to the Singapore framework for representing application profiles. The paper discusses the basic steps in developing an application profile, and the utility of application profiles in constructing and validating bibliographic records using a specific content standard. The paper concludes by observing that Dublin Core application profiles expressed in Resource Description Framework can improve the interoperability of metadata created by specific communities at the global level of the Semantic Web. Examples from the International Standard Bibliographic Description and RDA: resource description and access are given throughout the paper.

## Content standards

The Statement of cataloguing principles (ICP) issued by the International Federation of Library Associations and Institutions (IFLA) in 2009<sup>1</sup> states its goal as “serving as a basis for international standardization in cataloguing” and its scope as “to guide the development of cataloguing codes”. ICP itself does not offer a definition of “cataloguing code”, but Dictionary.com gives a definition based on the Random House dictionary as “the principles, rules, and regulations for entering and describing books or other library material in a catalog”<sup>2</sup>. ICP refers to “controlled forms of names”, “vocabulary used in descriptions”, and “data elements”, and notes that “a cataloguing code should take into account the entities, attributes, and relationships as defined in conceptual models of the bibliographic universe”. ICP “can be applied to bibliographies and other data files created by libraries, archives, museums, and other communities”.

A cataloguing code is therefore a standard which determines the structure and content of descriptions of items in archive, library, and museum collections: metadata consisting of elements whose values are determined by rules and regulations.

Two major international standards have been influenced by ICP: RDA: resource description and access<sup>3</sup>; and the International Standard Bibliographic Description (ISBD)<sup>4</sup>. The introduction to RDA says it “provides a set of guidelines and instructions on formulating data to support resource discovery”, while the introduction to the consolidated edition of ISBD says it “determines the data elements to be recorded or transcribed in a specific sequence as the basis of the description of the resource being catalogued”. In fact RDA also provides a list of data elements based on IFLA’s Functional requirements for bibliographic records (FRBR) and Functional requirements for authority data (FRAD), and ISBD provides rules for the content of its data elements. Both are content standards for specified metadata elements primarily of use in libraries.

Many other content standards are available for library and other domains, such as Describing archives: a content standard (DACS)<sup>5</sup>, and International core data standards for ethnology/ethnography<sup>6</sup>.

The general utility of content standards is to support the consistency, coherency, and completeness of metadata records used in resource discovery services. Such standards are generally different for different domains and contexts, and they usually change over time; for example, RDA is intended to replace the Anglo-American Cataloguing Rules (AACR) issued in two editions since 1967.

The components of a cataloguing code or content standard alluded to in ICP can be further determined by examining how rules are formulated in the text of the standard.

From the ISBD consolidated edition: “The place of publication, production and/or distribution is the name of the place associated on the prescribed source of information with the name of the publisher, producer or distributor.” In this example, “place of publication, production and/or distribution” is the element in focus and “name of the publisher, producer or distributor” is a related element; “associated” indicates a relationship between the two elements. The element in focus is defined as “the name of the place associated with the name of the publisher, producer or distributor” and “prescribed source of information” is the source of the content or value of the element.

From the RDA toolkit: “Record the type of carrier used to convey the content of the resource using one or more of the terms listed below.” Here, “type of carrier” is the element in focus, and “used to convey the content of the resource” is its definition. The value of the element is to be taken from the “terms listed below”, which form a controlled vocabulary.

Analysis of ICP and specific standards suggests that the components typically found in a content standard are:

- Data elements and their definitions.
- Relationships between data elements.
- Intended usage of data elements.
- Source of values for data elements.

## **DCMI Abstract Model**

The abstract model developed by the Dublin Core Metadata Initiative (DCMI) specifies “the components and constructs used in Dublin Core metadata ... an information model which is independent of any particular encoding syntax”<sup>7</sup>. The model was designed to bridge the modern paradigm of unbounded, linked data graphs typified by the Semantic Web with the more familiar focus on metadata records which can be validated against a cataloguing code or content standard.

Dublin Core metadata elements are intended for simple, generic descriptions of resources, and it has been recognized from the start of the 21<sup>st</sup> century that specific implementations would use these elements together with other specialized vocabularies of elements and values. DCMI uses the term “application profile” for documentation which describes how standards can be tailored to particular applications, and has developed the Singapore framework for Dublin Core application profiles<sup>8</sup> to

maximise the interoperability and reusability of an application. The framework uses the DCMI Abstract model (DCAM) to formalize application profiles which can be processed by computers. Furthermore, a Dublin Core application profile (DCAP) supports the mixing of components from different metadata standards, none of which have to be Dublin Core itself. That is, a DCAP need not use any Dublin Core elements, and can therefore be created for implementations of specific content standards such as ISBD and RDA, using only elements from within the standard. Indeed, the ISBD/XML Study Group has created a draft application profile for ISBD<sup>9</sup> and the former DCMI/RDA Task Group<sup>10</sup> intended to develop an application profile for RDA based on FRBR and FRAD. The latter initiative has been delayed while awaiting publication of the RDA element sets, and has been subsumed by the successor DCMI Bibliographic Metadata Task Group<sup>11</sup>.

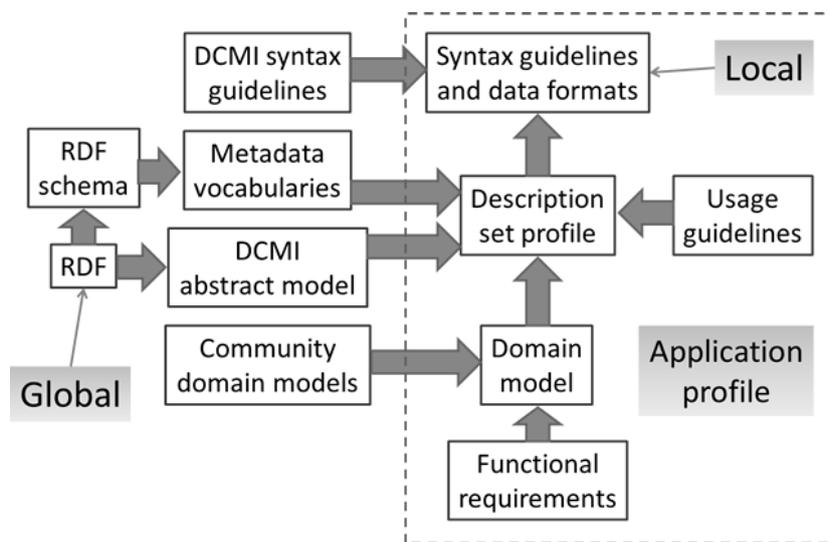


Figure 1: The components of the Singapore framework for Dublin Core application profiles.

Figure 1 shows the components of a DCAP related to Resource description framework (RDF) and the domain standards of models and specifications in broader use. RDF is the basis of the Semantic Web, and its related RDF schema provides a language for describing RDF vocabularies. These are the global foundation standards of a DCAP. The domain standards include generic DCMI guidelines on expressing an application profile in a specific syntax or data format, the metadata vocabularies and domain models from specific communities that may be included in an application profile, and the DCAM itself. A domain model is the set of entities or things of interest to a community, together with attributes or properties of those things and relationships between them; for example, the domain of ICP is the “bibliographic universe”. Although domain standards have broad application with a community, they are not truly global. The library community has other universes of interest, such as professional training, architecture and design of library buildings, collection management, and so on. An application profile can be even more local in its intended use for specific types of material, audiences, and functions.

The DCAP components include the functional requirements for the application ( for example FRBR and FRAD for RDA), the specific domain model of entities and relationships covered by the application (for example, ISBD has only the single entity Resource), usage guidelines applicable to data elements (for example, the FRBR and FRAD user tasks assigned to RDA elements), syntax

guidelines and data formats for elements and their assembly into a record (for example, ISBD's punctuation rules), and the description set profile (DSP) for the application.

A DSP describes a record as a set of descriptions for each entity or thing identified in the domain model. A description is a set of statements about a single entity; each statement refers to a single data element or property of that entity. A description also includes constraints on the data elements, such as their repeatability, mandatory status, and sequence, and the relationships of the entity to other entities in the application.

The content standard components identified above can be mapped to the components of an application profile. The data elements and definitions are part of the DCAP domain model; the data elements, relationships, and usage are part of the DSP; and usage and source are part of the Usage guidelines. The Singapore framework can therefore be used to represent the documentation of a cataloguing code or content standard in a form that can be processed by machine. The basic steps for achieving this are given in the Guidelines for Dublin Core application profiles<sup>12</sup>. The general order is to define the functional requirements of the application, then select or develop a domain model, then select or define metadata "terms" for the properties representing the attributes and relationships of entities in the domain model and their values, then determine the constraints on those terms. The structure of a valid record using these entities, properties, values, and constraints is then expressed in a DSP. Finally, usage guidelines are assigned to the entities, properties, and values. Syntax guidelines then allow developers to express the application profile as operational software.

## Constraints

The representation of constraints associated with the entities, properties, and values in a DSP is an essential factor in the determining the overall utility of a DCAP. Constraints determine the validity of records created using a specific content standard. They can be used to detect, for example, missing data which is mandatory, repeated data which should be single-valued, data which does not conform to a controlled vocabulary, and data which does not conform to a specified date or number format.

Mandatory and repeatability status is represented by minimum and maximum values for the occurrence of a property in a DSP description. A minimum occurrence value of 1 indicates that the property is mandatory, while a maximum occurrence value of 1 is used for a non-repeatable property. For example, the ISBD data element for media type is mandatory and non-repeatable, so the corresponding property has minimum and maximum occurrence set to 1.

The use of a controlled vocabulary is indicated by specifying a Vocabulary encoding scheme (VES), from which the value of a property must be taken. For example, the ISBD value vocabulary for media type is specified as the VES for the ISBD data element for media type.

A constraint on the format of the literal value of a text strings or number is specified using a Syntax encoding scheme (SES). There are many pre-defined encoding schemes for dates and numbers specifying, for example, that a date should be in day-month-year format or a number should be represented by a floating-point decimal. SESs are particularly useful for specifying aggregations of data elements which need to be processed as single elements. For example, ISBD area 0 for content

form and media type is an aggregation of elements for content form, content qualification, and media type. The area as a whole is mandatory and non-repeatable, and is the first in the sequence of areas in an ISBD record.

As well as validating records, a DSP with constraints can also be used to create records from existing data. The source of the original data is irrelevant providing the data element's property and value conform to the DSP, and there is a wide range of ontological properties that can be used in many cases to match properties from a domain outside of the application profile to a property in the profile's domain model.

## RDF

RDF enables global interoperability of metadata, so it is desirable to represent a DCAP entirely in RDF. ISBD, RDA, and a number of other library content or metadata format standards already have an RDF representation of the entities, attributes, relationships, and value vocabularies in their domain models. The RDF representation of description sets and constraints is not finalized, however, but the DCMI Architecture Forum<sup>13</sup> has been actively working on completing this work following a discussion on the ISBD application profile at the DC-2010 conference.

There has been little or no consideration of expressing usage and content source guidelines in RDF, as they are primarily intended for consumption by human cataloguers rather than machine processing. However, the rule used to determine the value of a metadata statement expressed as an RDF triple is likely to be of significance when determining the provenance of the triple. Provenance needs to answer questions such as "why did they say that" as well as "who said that"<sup>14</sup> and "when did they say it", particularly when the content standard behind the statement is changing. It may be important to know that a value was derived using AACR, AACR2, or RDA because the evolution of those standards involved major changes in how certain elements were treated. The date of creation of a triple is not a reliable guide to which version of a standard was used, as libraries do not switch from one version to another simultaneously. RDF representations of content standards as a whole, their versions, and the individual rules, including options, governing specific data elements are going to be important, and will be well worth developing.

Content standards are necessarily diverse, meeting the needs of specific communities for specific functions which change through time. The Singapore framework and Dublin Core application profiles are rooted in RDF, and provide a bridge between such local diversity and the global interoperability of the Semantic Web.

## References

<sup>1</sup> IFLA. Statement of international cataloguing principles. 2009. Available at:

[http://www.ifla.org/files/cataloguing/icp/icp\\_2009-en.pdf](http://www.ifla.org/files/cataloguing/icp/icp_2009-en.pdf)

<sup>2</sup> Dictionary.com. Catalog code. 2012. Available at:

<http://dictionary.reference.com/browse/catalog+code>

<sup>3</sup> RDA toolkit. RDA 0.0. Available at: <http://www.rdatoolkit.org/>

- <sup>4</sup> ISBD: International Standard Bibliographic Description. Consolidated edition. Edited by the Standing Committee of the IFLA Cataloguing Section. Berlin/Munich: De Gruyter Saur, 2011
- <sup>5</sup> Society of American Archivists. Describing archives: a content standard (DACS). Available at: <http://www.archivists.org/governance/standards/dacs.asp>
- <sup>6</sup> CIDOC. International core data standards for ethnology/ethnography . Edited by CIDOC Ethno Working Group. 1996. Available at: [http://icom.museum/uploads/tx\\_hpoindexbdd/CIDOC\\_EthnoGroup\\_CoreDataStandards\\_42249.pdf](http://icom.museum/uploads/tx_hpoindexbdd/CIDOC_EthnoGroup_CoreDataStandards_42249.pdf)
- <sup>7</sup> Andy Powell and others. DCMI abstract model. 2007. Available at: <http://dublincore.org/documents/abstract-model/>
- <sup>8</sup> Mikael Nilsson, Thomas Baker, Pete Johnston. The Singapore framework for Dublin Core application profiles. 2008. Available at: <http://dublincore.org/documents/singapore-framework/>
- <sup>9</sup> Mirna Willer, Gordon Dunsire, Boris Bosančić. ISBD and the Semantic Web. J LIS.it Journal of Library and Information Science. Italy, vol. 1, no. 2, (2010), pp. 213-236. Available at: <http://dx.doi.org/10.4403/jlis.it-4536>
- <sup>10</sup> DCMI/RDA Task Group. Historical archives. 2012. Available at: [http://wiki.dublincore.org/index.php/DCMI/RDA\\_Task\\_Group\\_Historical\\_Archives](http://wiki.dublincore.org/index.php/DCMI/RDA_Task_Group_Historical_Archives)
- <sup>11</sup> DCMI. Bibliographic Metadata Task Group. 2012. Available at: [http://wiki.dublincore.org/index.php/Bibliographic\\_Metadata\\_Task\\_Group](http://wiki.dublincore.org/index.php/Bibliographic_Metadata_Task_Group)
- <sup>12</sup> Karen Coyle, Thomas Baker. Guidelines for Dublin Core application profiles. 2009. Available at: <http://dublincore.org/documents/profile-guidelines/>
- <sup>13</sup> DCMI Architecture Forum. 2012. Available at: <http://dublincore.org/groups/architecture/>
- <sup>14</sup> Gordon Dunsire. Tko je to rekao? = Who said that? In: 13. Seminar Arhivi, Knjižnice, Muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastructure. Zagreb : Hrvatsko knjižničarsko društvo, 2010. English original available at: <http://www.gordondunsire.com/pubs/docs/whosaidthat.pdf>