Metadata for digital resources

Working Group on Metadata for Digital Resources

Copenhagen, August 2009

Table of contents

1. Introduction

There are obvious reasons for analyzing and trying to improve the ways that metadata for digital resources are generated, managed, used and reused:

- The same digital resources are made available to - and used by – different individuals and communities.
- Different individuals and communities produce both identical and different kinds of metadata – based on different practices, rules, formats and schemes.
- Reuse and value adding of metadata among different communities is very limited.
- Common standards and protocols are seldom employed across communities

This leads to the assumption that handling and mediation of digital resources can be eased and improved by strengthening cooperation between those communities, who creates, manages and uses digital resources and metadata.

In 2006 the IFLA Standing Committee on Cataloguing started a working group that should recommend guidelines for those who create, share and manage metadata (including: elements, tools, formats and protocols).

This report summarizes the work and results of the Working Group on Metadata for Digital Resources. The report is aimed at the Standing Committee on Cataloguing and the Cataloguing Section of IFLA and is not suitable for a bigger audience in its present form. I.e. descriptive chapters on e.g. metadata and digital resources are omitted, since these concepts should be well known to the intended audience.

The results of the working group are scarce, and consist of a few recommendations for the Cataloguing Section, some general models for the creation, management and use of Metadata for digital resources and some arguments for improving cooperation among communities that works with or influence metadata for digital resources.

With the delivery of this report, the Working Group on Metadata for Digital Resources has finished its work, and the working group is closed.

1.1. Charge and background for the report

The Working Group was constituted in 2006.

In 2005 the background for the work was described as follows:

"Digital publishing is rapidly spreading into different publishing domains, at present especially within the academic world and among government and municipal bodies. For libraries that have started collecting and organising this digital material it is obvious that it is much more time consuming and complicated to handle these kinds of resources than the traditional ones – if bibliographic data are not delivered together with the resources.
The majority of these resources are text documents, and like traditional resources, the majority of them carry bibliographic data – they have a title,

they have an author or it is made clear which organisation has issued the text. The context of the text, when and where and by whom it has been produced, can usually be derived from information scattered over the document. The bibliographic information is there, but it is not explicit and cannot be efficiently exploited."

The original 2005 draft charge for the Working Group focused on the establishment of a common set of metadata elements applicable for different communities. The charge has been changed several times during the work, and the final charge – decided by IFLA Cataloguing Standing Committee in 2007 is as follows:

Charge of the Working Group:

To recommend guidelines for those who create, share and manage metadata: elements, tools, formats, protocols

The working group will be guided by the following principles: The guidelines are envisaged as high level guidelines, meaning:

1. *they should serve as a pivot for semantic mapping, that is, they should serve as a data element dictionary to which different communities may map their own sets of bibliographic data*
2. *they should serve as guidelines on how to arrange the bibliographic data in a text document*
3. *they should not prescribe any specific tagging system, but*
4. *they should serve as a model against which different tagging systems may be measured, and*
5. *they should be easy to adopt in at least the most commonly used text editing programs*
6. *they should define a set of attributes for things like language, script, capitalisation practice, etc., to facilitate further processing*

Although the charge has been changed several times during the work, it still has some ambiguities and vagueness:

- The 6 dots in the charge deals with bibliographic data in a manner that covers both embedded metadata and common text structures. *The working group focused on metadata and not text structures.*

- The charge is not specific about, what types of resources should be included (text, sound, images, moving images etc.). *The models described in chapter 2. and 3. covers all types of digital publishing and media.*

1.2 Working group membership

The working group had the following members

> *Anders Cato*, Head of Cataloguing. Kungl. Biblioteket - National Library of Sweden

> *Francoise Bourdon*, Deputy Director of Bibliographic and Digital Information Department. Bibliothèque nationale de France

> *Marit Vestlie*. Director Culture and Public Programs. National Library of Norway

> *Erik Thorlund Jepsen*, Head of Library Development. The Danish Agency for Libraries and Media. *(Chair)*

> *Miriam Säfström*. BTJ Sweden *(since 2007)*

> *Gunilla Jonsson*. Former Head of Department for Collection Development & Documentation. Kungl. Biblioteket – National Library of Sweden *(2006)*

> *Erik Oltmans*. Former Manager for Acquisitions & Cataloguing at Koninklijke Bibliotheek *(2007)*

The report is based on discussions and input from the working group, but it is structured and written by Erik Thorlund Jepsen.

1.3 Working group activities and accomplishments

1.3.1. Activities

The working group has held 3 meetings (at the annual WLIC conferences in Seoul, Durban and Quebec) and 2 two-day workshops in Copenhagen and Paris.

Originally, the workgroup was intended to be a joint group between the following sections of IFLA Division IV: Cataloguing section, IT section and Bibliography section. It has not been possible to engage the two other sections in the work.

Prior to the establishment of the working group, there were discussions about the understanding and formulation of the charge. These discussions involved several persons from all sections of Division IV. Some of the discussions (e.g. concerning understanding and limitations of the concept bibliographic data in the WG context) have never been closed, and it is the opinion of the working group that this has lead to a final charge, which is still unclear and to broad in scope.

The working group started out very ambitiously by trying to map:
- Important actors concerned with or influencing the creation, refinement and use of metadata
    i. Their interest, needs and behaviour concerning metadata for digital resources.
- The most common tools and standards for those communities and actors, who are involved in or influencing the creation, refinement and use of metadata
    i. The content and structure in these standards
- Metadata needs of different actors in comparison with the common standards.

Unfortunately, we were not able to follow this line of work, since the group did not have sufficient expertise on:
- The interest, needs and behaviour of other communities (e.g. publishers and system developers)
- Existing metadata standards employed in non-cataloguing communities.

Instead, the working group focused on outlining the field of interest and existing problems and possibilities on a more general level.

1.3.2. Accomplishments

The working group has not fulfilled the charge.

All through the work, the group has lacked knowledgeable members on IT and metadata standards and needs and behaviours in communities outside library cataloguing. Furthermore, the group has never managed to come up with appropriate methodologies for fulfilling the charge.

Another reason for not fulfilling the charge is that the group agreed that prior work (report on metadata removed from IFLA-net *Guidance on the Nature, Implementation, and Evaluation of Metadata Schemas in Libraries: Final Report of the IFLA Cataloguing Section Working Group on the Use of Metadata Schemas, 2005*) has tried to come up with solutions similar to the 6 bullets in the charge, and we were (with our competences) not able too come up with new solutions.

Instead, the working group has produced:
- Some arguments for corporation between communities that work with digital resources and metadata
- Two general models that can be used in arguing for increased cooperation
- A few recommendations for the Cataloguing Section

1.4 Recommendations for the Cataloguing Section

The workgroup recommends:

1. *That* the Cataloguing Section and The Standing Committee on Cataloguing studies and evaluate the report. *Are the recommendations feasible and in proportion with the present understanding and handling of metadata?*
2. *The* chair (or appointed members from) of the Standing Committee contacts other sections for follow up initiatives. *Who inside IFLA should be involved in refining the recommendations and in planning a strategy for further work inside IFLA*
3. Eventually, that the Standing Committee on Cataloguing lay out a plan for further work. *Who outside IFLA needs to be involved if initiatives, besides local institutional efforts, should be started up (e.g. with system developers, organizations of publishers etc.)*

It is not recommended to publish the report in its current form, but feel free to use whatever content could prove usable.

1.5 Structure/Outline of the report

In chapter '2. Resource life cycle' we introduce a model for the life cycle of digital resources and discuss how metadata production, use and reuse can be improved at different stages of the life cycle.
The resource life cycle consist of 4 stages characterized by what is being done to the resource by what kind of actor. Actions are creation, publishing, access provision and use.
In chapter '3. Actors, Roles and Metadata' we describe a model that can be employed in all sorts of projects that try to optimize metadata production, corporation and reuse of metadata. The model comprises description of the concepts actors, roles, common behaviour, interest, tools, metadata needs and metadata created.
In the conclusion we summarize the results, arguments and recommendations.

2. Resource life cycle and metadata:

In this chapter we present a general model of the lifecycle of a digital object.

Metadata for the same entity is created, used and reused by different actors at different points of time in the life cycle of the resource.

2.1. Resource Life cycle

Figure 1. show a model that outlines the life cycle of the resource, from creation to consumption.

Figure 1. Resource life cycle
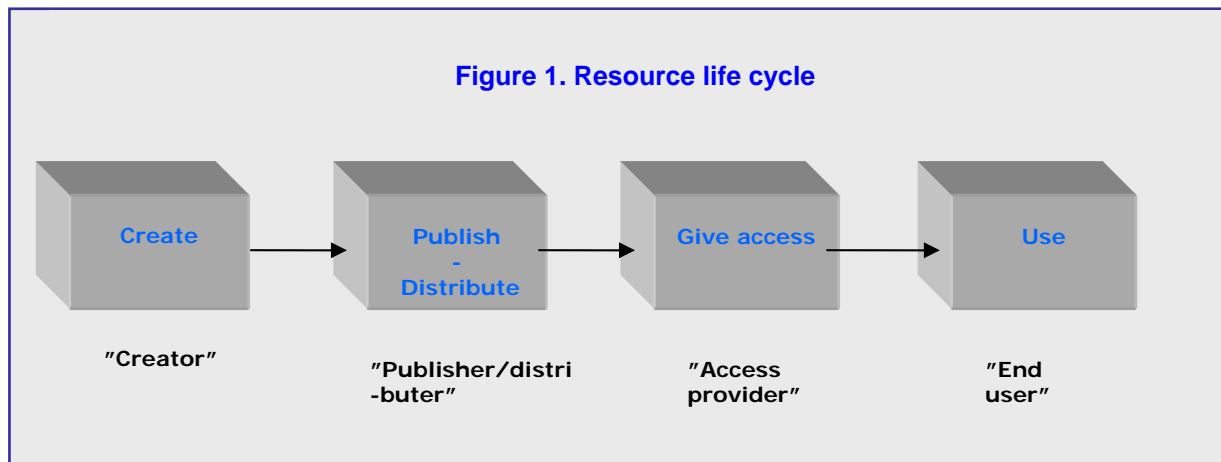
The different stages in the life cycle of the resource are characterised by what is being done to it (actions).
A resource is created, published/distributed, given access to and consumed (used).
Different parties can be active at different stages of the life of the same resource. E.g. a dissertation is written by a researcher (created), digitized by his/her university (published), given access to via the library catalogue and/or the digital repository at the university and finally printed and read by a student (used).
Several actions (e.g. creation and publishing) can be performed by the same party (person or institution).

2.2. Resource lifecycle and metadata

Metadata for the resource can be created and used at any stage in the model. Although some metadata elements will be directly associated with the interests of e.g. an access provider, the assumption of the working group is that the metadata can be refined during the lifecycle of the resource, and in this way benefit actors further down the line.
Reuse and value adding of metadata are key concepts in this refinement process.

Ideally, you could try to identify metadata value chains, where a basic set of metadata, added by the creator(s) or generated automatically, is refined and/or added new metadata along the life cycle of the resource.
Probably this is only plausible in certain fields of publication (e.g. in traditional scholarly publishing in journal articles and papers and in commercial music), but analyses of the possibilities performed by joint groups (e.g. publishers and access providers to scientific articles) could lead to mutual benefits for those who create, publish, provide access to and use the resources.

It is not uncommon that the same or basically the same metadata is created at different stages of the life of the resource. By regarding the whole cycle or chain it would be possible to avoid this double work and also some redundancy and creation of non-usable metadata. Partnerships, dialogue and shared

standards are also ways to work around the differences in normative forms and specificity of metadata.

Example:
1. *Already, when you are writing a text in e.g. Word, the system generates a quite large set of metadata on title, author, institution, size and more.*
2. *When the publication is accepted and published, the editor creates a new set of metadata following a given registration standard like Onix or similar or a local solution.*
3. *When the resource is then registered by access providers like libraries or digital repositories, a whole new set of metadata is created.*
4. *Nowadays, also consumers can add metadata (tag and evaluate). And in some instances these user added metadata is employed to facilitate other consumer's access and evaluation of relevance of the resource.*

There are some obvious reasons, why metadata sets and wording of metadata is not identical at the 4 stages above. Administrative metadata will almost certainly differ.

Nevertheless, all areas of publication could probably identify – or choose - a metadata core element set, consisting of a few central metadata that are consistent and useful throughout the lifecycle of the resource.

Often, it is quite easy for actors to consider evident metadata needs of the next actors in the life cycle. Yet, actions are necessary to make them see this. Actions can be:

1. Dialogue that convinces the first actor that benefits are mutual or that implementation is without significant cost.
E.g. In Denmark a common exchange format are developed for – and adopted by central institutions in the ALM sectors. This work was initiated by government institutions but is now adopted by almost all levels of production, distribution and mediation in the public area.

2. Implementation of feasible business models that serves two or more actors.
E.g. a common metadata set and format is developed for and used by distributors of digital music recordings and the central library website for digital music recordings (who are a major customer). The metadata set and format was a joint initiative between the two partners.

E.g. at several universities in Scandinavia, students are forced to deliver given metadata in given wording in digital form when they deliver their thesis'. These metadata are then reused and refined by university libraries, university digital repositories and other institutions (e.g. the national bibliographies) that reuse the university records.

3. Implementation of demands for given metadata in given forms by those who acquire resources from others (e.g. that libraries only buy resources from distributors that add certain metadata in given schemes)

> *E.g. The Danish State Library 'Statsbiblioteket' prioritize distributors that alongside the resource delivers metadata in MARC.*

The working group has also discussed, if it is possible to interact with system developers like e.g. Microsoft, so that metadata created while creating and storing the resource (by creator) are tailored to fit the needs of other actors in the resource life cycle and the standards employed by these actors.
There is a slight tendency towards such solutions. E.g., Microsoft has recently adopted Dublin Core Metadata element set in their xml based exchange format Office Open XML.

In the next chapter, we will look into, what determines metadata needs and introduce a model that can be used for analyzing problems and possibilities concerning metadata in most situations.

3. Actors, roles and metadata

When analyzing potential benefits of reuse and refinement of metadata, we found it necessary to look into who creates - and benefits from the creation of – metadata.

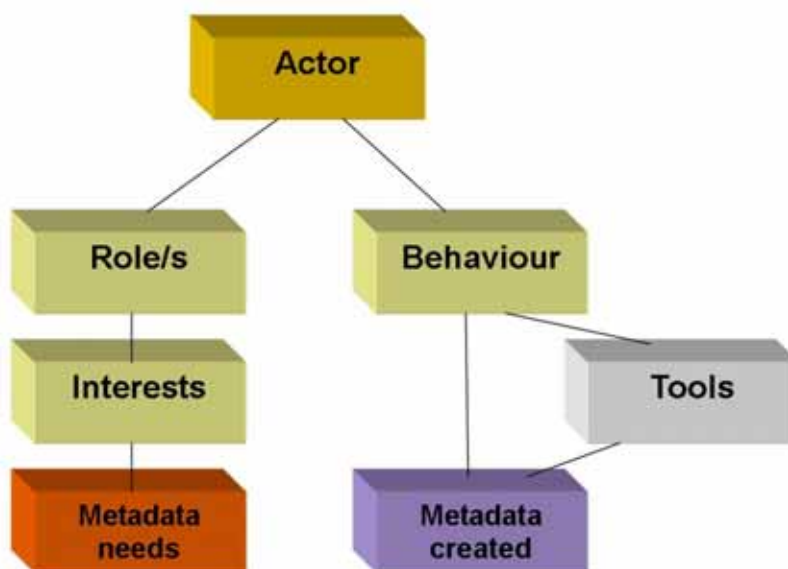**Figure 2: Model for analysis: high level**

Figure 2. shows how an *Actor* plays one or many *Role/s* in his/her interaction with the digital resource. Each role has a set of *Interests* connected with it, and these interests are what determine the *Metadata needs* of immediate interest.

The model also shows *Metadata created*, which is something connected with the actors "traditional" *Behaviour*, and also as a result of which *Tools* (chosen in accordance with behaviour) have been used to create the metadata.

The model allows for analysis of metadata production patterns. Discrepancies between the *Metadata needs* and the *Metadata created* are of course interesting, since analyses and models like FRBR in the library community are not necessarily developed in other communities.

Explanation of elements in the model (Fig.2.):

Actor: An entity that plays an active part in the lifecycle of the resource and that can create, refine, use and circulate metadata. Every actor can act in different roles. Actors can be:
- Individuals
- Publishers
  - Private
  - Universities
  - Public administration
- Libraries
- Repositories
- Bookstores

Role/s: A role is defined by action. Each action is connected with a certain stage in the life cycle of the resource. Roles are creator, publisher, distributor, access provider and end user.
- Creator: an entity primarily responsible for making the content of the resource.
- Publisher/Distributor: an entity who issues or makes available publications to the public.
- Access provider: one or more institution(s) or system(s) that give end users access to resources.
- End user: one or more individual(s) or group(s) that consume (use, read, listen to, reuse) the content of the resource.

All actors can take on all roles.

Interests are the expected benefits (of the metadata). Interests are connected with the role but differ slightly depending on who is playing the role. E.g. when an individual (actor) publishes (role) something on MySpace his/her interests will vary from them of a university library (actor) who publishes (role) a digitized manuscript in a hand print project.

Behaviour is the general professional/practical behaviour of the actor. In this high level model we deal mainly with "traditional" behaviour (e.g. libraries have a strong metadata tradition and create elaborate descriptions; individuals generally have no knowledge of cataloguing principles and use whatever tools available).

Tools are used to create and manage metadata. Tools can be formats (e.g. MARC, DC), rules (e.g. AACR2, RDA, controlled vocabularies), protocols (e.g. OAI-PMH), software and applications (e.g. forms, metadata extraction applications, Voyager).

Metadata needs: the metadata elements needed to support the interests of an actor in his/her role(s).

Metadata created: the metadata actually created by an actor in one of his/her roles.
What metadata is created depends on the actor, his/her behaviour and which tools are used.

The model shows actors guided by their own "traditional behaviour", playing all roles in the same way and choosing tools out of habit.
Eventual discrepancies between metadata needs and metadata created pose the questions whether this is a good idea. (E.g. is MARC really the best format/tool for a library serving as a publisher/access provider? Could the individual using MySpace to publish/give access gain from being able to create more "library-like" metadata? Etc.)
For some communities there are pretty good ideas about common needs and interest (e.g. FRBR and FRAD) other instances seems more based on habitude than up-to-data needs and interest e.g. Marc formats and ONIX.


3.1. Potential use of the model

The model described in Figure 2. can be employed in the analysis of metadata production and reuse at all levels:

- Institutional / local level
  e.g. the metadata production and reuse at a university (including creation, publishing and access providing)
- National level
  e.g. the implementation of multi sectored metadata solutions or negotiations/joint activities between government bodies and bodies presenting e.g. publishers on the national level
- Higher / International level
  e.g. joint activities between IFLA and IPA

The strength of the model is that it emphasises several evident areas of investigation when trying to improve metadata production, reuse and refinement. Whether these improvements take place inside a given community or as joint ventures between different communities (e.g. different roles in the resource life cycle)

A crucial element in using the model, is that all parties are (or becomes) knowledgeable of the right side of the model (behaviour, tools used and metadata created: elements, formats and wording), and are willing to investigate – in a critical manner – the left side of the model (interest and metadata needs) in order to change or refine the right side.

Figure 3. Joint effort to improve metadata production and reuse
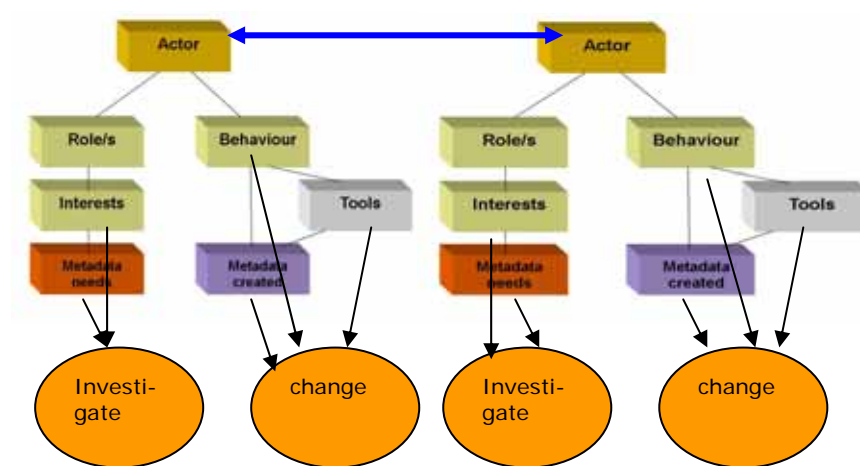
4. Summary and conclusions

The working group has not fulfilled the charge.

Instead, the working group has produced:
- Some arguments for corporation between communities that work with digital resources and metadata
- Two general models that can be used in arguing for – and analyzing potential benefits of - increased cooperation in metadata production, use and reuse
- A few recommendations for the Cataloguing Section

Models and arguments
The life cycle of digital resources consist of 4 stages: Creation; Publishing/distribution; Access providing; Use.

Often, it is quite easy for actors at each stage of the life cycle to consider evident metadata needs of the next (or foregoing) actors in the life cycle. Yet, actions are necessary to make them see this. Actions can be:

- Dialogue that convinces the first actor that benefits are mutual or that implementation is without significant cost.
- Implementation of feasible business models that serves two or more actors.
- Implementation of demands for given metadata in given forms by those who acquire resources from others (e.g. that libraries only buy resources from distributors that add certain metadata in given schemes)

Supplementing the life cycle of digital resources, we have developed a model containing central elements that must be considered when trying to improve

metadata production, use and reuse. The strength of this model is that it emphasises several evident areas of investigation when trying to improve metadata production, reuse and refinement. Whether these improvements take place inside a given community or as joint ventures between different communities (e.g. different roles in the resource life cycle)

Recommendations

The workgroup recommends:

- That the Cataloguing Section and The Standing Committee on Cataloguing studies and evaluate the report. *Are the recommendations feasible and in proportion with the present understanding and handling of metadata?*
- The chair (or appointed members from) of the Standing Committee contacts other sections for follow up initiatives. *Who inside IFLA should be involved in refining the recommendations and in planning a strategy for further work inside IFLA. In best case, this report can be used as a starting point for such efforts.*
- Eventually, that the Standing Committee on Cataloguing lay out a plan for further work. *Who outside IFLA needs to be involved if initiatives, besides local institutional efforts, should be started up (e.g. with system developers, organizations of publishers etc.)*

It is not recommended to publish the report in its current form, but feel free to use whatever content could prove usable.

With the delivery of this report, the Working Group on Metadata for Digital Resources has finished its work, and the working group is closed.

Copenhagen, 25-8-2009

Erik Thorlund Jepsen
Head of Library Development. The Danish Agency for Libraries and Media