# A proposal for metadata formats for use in the Swedish Enviro–Net

*Sigfrid Lundberg*

Lund university library, NetLab
P.O. Box 3, S-221 00 Lund, Sweden

## Introduction

Metadata is, by definition, data on data. For printed material the production of metadata, i.e. cataloging, abstracting and indexing, has been an important task of librarians for more a century. Obviously it is an expensive enterprise to use trained staff for this, and because of the volatility of the Internet the trend has been to completely abandon metadata production, and completely rely on full text searching. As a consequence it is today impossible to look up *August Strindberg* in any of the popular WWW search services *and* restrict the search to material *about* him, *or* to material authored *by* him.

Although copyright issues are hot topics on the Internet, the Hypertext Mark up Language [1] does not provide facilities for informing document users about metadata like copyright ownership or author. Currently document title is the only piece of information. supported by HTML.

This document does is not intended to provide *general* recommendations for metadata production for electronic publications. This is the task of other projects, such as the Nordic Metadata Project [2]. Instead it is a customized proposal for the Swedish Environet. As far as the current standardization efforts are concerned, the state of affairs is not as stable as one would wish. Hence parts of the solution proposed has to some extent be proprietary, but we have largely built the package upon components that are *not* considered controversial today. We are confident that the solution proposed is at least fairly close to what will be a *de facto* standard in the near future.

## Dublin Core: DC—The *Jack-of-all-trades* metadata attribute set

Most metadata attribute sets have been proposed for the description of one or only a few kinds of data. They are thus optimized for providing good solutions to fairly narrow ranges of problems. Such a metadata system tailored for governmental information sources will be discussed further below.

Before we go into that particular niche, we will briefly mention the most general of them all: Dublin Core Set, DC for short [3,4,5]. Jack-of-all-trades is master of none; and DC does not cover all special cases. However, DC is optimized for ease of use in conjunction with electronic material, and, as we will see, for making it possible to use it as a sort of *lingua franca* metadata attribute set. The DC set contains only thirteen attributes (***Table 1***), all of which are optional and repeatable. The DC standardization effort, which is not yet completed, is endorsed by most workers involved in other more specialist standards.

***Table 1.*** The following comprises the reference definition of the Dublin Core Metadata Element set as of December, 1996. The elements or their names are not expected to change substantively from this list, though the application of some of them are currently experimental and subject to interpretation. Further, it is expected that practice will evolve to include sub-elements for certain of the elements.

| **Element Lable** | **Element Description** |
|---|---|
| TITLE | The name given to the resource by the CREATOR or PUBLISHER. |
| CREATOR | The person(s) or organization(s) primarily responsible for the intellectual content of the resource. For example, authors in the case of written documents, artists, photographers, or illustrators in the case of visual resources. |
| SUBJECT | The topic of the resource, or keywords or phrases that describe the subject or content of the resource. The intent of the specification of this element is to promote the use of controlled vocabularies and keywords. |
| DESCRIPTION | A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources. |
| PUBLISHER | The entity responsible for making the resource available in its present form, such as a publisher, a university department, or a corporate entity. The intent of specifying this field is to identify the entity that provides access to the resource. |
| CONTRIBUTORS | Person(s) or organization(s) in addition to those specified in the CREATOR element who have made significant intellectual contributions to the resource but whose contribution is secondary to the individuals or entities specifed in the CREATOR element (for example, editors, transcribers, illustrators, and convenors). |
| DATE | The date the resource was made available in its present form. The recommended best practice is an 8 digit number in the form YYYYMMDD as defined by ANSI X3.30-1985. In this scheme, the date element for the day this is written would be 19961203, or December 3, 1996. |
| TYPE | The category of the resource, such as home page, novel, poem, working paper, technical report, essay, dictionary. It is expected that RESOURCE TYPE will be chosen from an enumerated list of types. |
| FORMAT | The data representation of the resource, such as text/html, ASCII, Postscript file, executable application, or JPEG image. The intent of specifying this element is to provide information necessary to allow people or machines to make decisions about the usability of the encoded data (what hardware and software might be required to display or execute it, for example). |
| IDENTIFIER | String or number used to uniquely identify the resource. Examples for networked resources include URLs and URNs (when implemented). Other globally-unique identifiers,such as International Standard Book Numbers (ISBN) or other formal names would also be candidates for this element. |

| Element Lable | Element Description |
|---|---|
| SOURCE | The work, either print or electronic, from which this resource is derived, if applicable. For example, an html encoding of a Shakespearean sonnet might identify the paper version of the sonnet from which the electronic version was transcribed. |
| LANGUAGE | Language(s) of the intellectual content of the resource. Where practical, the content of this field should coincide with the Z39.53 three character codes for written languages. |
| RELATION | Relationship to other resources. The intent of specifying this element is to provide a means to express relationships among resources that have formal relationships to others, but exist as discrete resources themselves. |
| COVERAGE | The spatial locations and temporal durations characteristic of the resource. Formal specification of COVERAGE is currently under development. Users and developers should understand that use of this element should be currently considered experimental. |
| RIGHTS | The content of this element is intended to be a link (a URL or other suitable URI as appropriate) to a copyright notice, a rights-management statement, or perhaps a server that would provide such information in a dynamic way. No assumptions should be made by users if such a field is empty or not present. |

The intention is that the basic DC attribute set should be easy enough to use even for casual users. However, one of the major strengths of the DC metadata system is the sub-elements [6], which enables organizations to produce a detailed metadata. It is even possible accommodate metadata produced using more specialized attributes within a DC record. This is the case for the GILS metadata attribute set, treated in the next section.

## Governmental Information Locator Service: GILS—The metadata attribute set for governmental information

The adoption the GILS metadata system by federal governmental organizations with a few notable exceptions (such as operational data at the Central Intelligence Agency) is mandatory according to U.S. Public Law 44 USC 3511. The purpose of GILS has been described as follows:

> *As part of the Federal role in the National Information Infrastructure, GILS identifies and describes information resources throughout the Federal government, and provides assistance in obtaining the information. GILS supplements other government and commercial information dissemination mechanisms, and uses international standards for information search and retrieval so that information can be retrieved in a variety of ways* [7].

Although GILS originated as a federal initiative in the United States to give the tax payers easy access to information which collection they have funded, there are similar initiative in many democratic countries. The most important international is the G7 Enivronment and Natural Resource Management Project [8,9]. And many of them are using GILS in order to facilitate interoperability.

Since GILS has a strong and enthusiastic user community among governmental organizations world wide, we suggest that the Swedish Enviro–Net's locator service should use that attribute set in its main database.

### Embeddable metadata

The are two methods of metadata delivery: It may be delivered together with the object it describes (like the Library of Congress and British Library records printed in many text books today). Or it can be found in databases. Needless to say, many libraries use the information provided in the printed record to create upgraded bibliographic records.

It would be a logical step to do the same for electronical documents, and it is possible to embed metadata inside documents of some types. The metadata will then reside in the document for the benefit of robot based search services. Although the facilities provided by HTML for the purpose are poor, it is possible to do so using:

```
<META NAME="SOME_ATTRIBUTE" CONTENT="SOME_VALUE">
```

The problem with this scheme has so far been that there have been few and poorly standardized attributes available for the purpose, and no method for providing information about the attribute sets used. This problem has recently addressed at a workshop sponsored by the World Wide Web Consortium [10]. The current recommended scheme is described in ***Appendix A***.

For obvious reasons it is much more difficult or impossible to do so for multi–media objects like images and sounds. There are simple methods to avoid those problems, and the same is true for objects that are not document like. Swedish Enviro–Net project will provide metadata for databases, which also fits into this category. The workaround would be to embed metadata into that HTML document which is carrying, the so to speak, canonical link to the object or service in question (e.g., the databases search forms), or to deliver multimedia objects as inlined objects, and embed its metadata in the page in which it is inlined.

Although it is possible to embed a GILS record directly in HTML (see ***Appendix A*** for a real world example of this), we suggest that the Swedish Enviro–Net takes the step to use DC for this purpose. In all but the most complex cases it will be possible to allow document authors create DC records, and convert these at the Enviro–Net server to GILS records [11]. The gains of this procedure are that DC is considerably easier to use, and that the DC community exerts a strong pressure upon search service providers (Lycos, AltaVista etc.). It is most likely that these services will start supporting DC metadata before any other metadata system. Hence there is a fair chance that the metadata produced with in the project eventually will reach a very wide audience.

### Other relevant metadata systems

It is not unlikely that the DC/GILS combo might turnout to be wanting, in that various kinds scientific data data collections may call for more detailed descriptions. An appropriate description of such data might require specialized scientific, geospacial or biological metadata attribute sets. There are accepted or evolving metadata standards within these areas: The Scientific and Technical Attribute and Element Set (STAS) [12], FGDC [13], which got its name from the body that proposed it, the Federal Geographic Data Committee, and NBII, National Biological Information Infrastructure [14], of the U.S. National Biological Survey. Of these, the STAS and FGDC standards seem to by the most advanced stage of development. STAS is used by Chemical Abstracts and the first version of FGDC is released and is used by military (primarily the by the U.S. Navy) and governmental agencies in the United States.

We have no firsthand experience of using any these and are currently not prepared to give advice, but urge the Swedish Enviro–Net to undertake a study to assess the needs for facilities provided by these standards within the project.

### Summary of recommendations

The Swedish Environ–Net project has the potential to set the standards for similar efforts in Sweden and perhaps also in the Nordic Countries. From our point of view, we would welcome a formal statement from the project that it will build its information structure using open standards, in order to facilitate interoperability. We suggest:

1. The use of GILS attribute set for describing resources in the Environet.
2. That metadata are produced by document authors or publishers and not by the Environet itself. To this end
3. Most metadata should be embedded in document using DC. This will make maintenance of the database easier, since records may be revised together with the documents they describe.
5. The Environet's database will be possible to produce using a harvesting robot.
6. Quality ratings, or statements can be maintained by Environet as external metadata objects.
7. More detailed studies are needed on the need metadata systems with special features. I suggest that the Swedish Enviro–Net should review the needs for advanced metadata formats like FGDC, STAS and NBII metadata standards.
8 Although this paper does not cover the classification of resources, it is of utmost importance for the success of the project (if the service is to include a browseable information structure). Again I suggest that this problem should be addressed by trained librarians at member organizations.

### References

1. Berners-Lee, T. and Connolly, D., 1995. *Hypertext Markup Language – 2.0.,* Request for Comments 1866. http://www.it.kth.se/docs/rfc/rfcs/rfc1866.txt.
2. Hakala, J., 1986. *The homepage of the Nordic Metadata Project 1996–1998..* http://linnea.helsinki.fi/meta/.
3. Weibel, S. L. and Miller, E. J., 1986. *Dublin Core Element Set Reference Page.* OCLC Online Computer Library Center, Inc. http://purl.org/metadata/dublin_core.

4. Tkac, V. M., 1996. *OCLC/NCSA Metadata Workshop: The Essential Elements of Network Object Description..* http://www.oclc.org:5046/oclc/research/conferences/metadata/metadata.html.

5. Miller, P., 1996. Metadata for the masses: what is it, how can it help me, and how can I use it? *Ariadne* 5 (http://www.ukoln.ac.uk/ariadne/issue5/metadata-masses).

6. Knight, J. and Hamilton, M., 1986. *Dublin Core Sub-Elements..* http://www.roads.lut.ac.uk/Metadata/DC-SubElements.html.

7. Christian, E., 1986. What is GILS? *Government Information Locator Service (GILS)* (http://www.usgs.gov/gils/intro.html) U.S. Geological Survey.

8. —— , 1996. Environment and Natural Resources Management.In: *G7 Global Information Society (http://www.g7.fed.us/),* U.S. Geological Survey. http://www.g7.fed.us/enrm/enviro.html.

9. ENRM, 1997. *G7 Enivronment and Natural Resource Management Project.* Centre for Earth Observation (http://www.ceo.org/). http://enrm.ceo.org/free/info.html.

10. Schwartz, M., 1996. *Report of the distributed indexing/searching workshop..* http://www.w3.org/pub/WWW/Search/9605-Indexing-Workshop/.

11. Miller, E., 1996. *Dublin Core Element Set Crosswalk.* OCLC Online Computer Library Center, Inc. http://www.oclc.org:5046/~emiller/DC/crosswalk.html.

12. Nassar, N., 1986. *Scientific and Technical Attribute and Element Set (STAS)..* http://www.cnidr.org/ir/stas.html.

13. The Federal Geographic Data Committee, 1986. *Metadata Standards Development..* http://www.fgdc.gov/Metadata/metahome.html.

14. National Biological Survey, 1986. Building the Infrastructure: Standards, Protocols, and Tools Relating to the NBII. *National Biological Information Infrastructure* (http://www.its.nbs.gov/nbii/infrastructure/).

### Appendix A.

The general recommended syntax for metadata embedding is to use the HTML META tag for distributing attributes and values:

```
<META NAME="META_DATA_SCHEME.attribute"
      CONTENT="Value of that attribute">
```

Then a LINK tag is used for providing a definition of the META_DATA_SCHEME in question:

```
<LINK REL="SCHEMA.META_DATA_SCHEME.attribute"
      HREF="URL to document describing the attribute">
```

The following is a part of a detailed DC record, complete with links to descriptions of each attribute:

```
<META NAME="DC.title"
      CONTENT="(TYPE=long) Metadata for the masses:
      what is it, how can it help me, and how can I use it?">
<LINK REL=SCHEMA.dc
      HREF="http://purl.org/metadata/dublin_core_elements#title">

<META NAME="DC.title"
      CONTENT="(TYPE=short) Metadata for the masses">
<LINK REL=SCHEMA.dc
      HREF="http://purl.org/metadata/dublin_core_elements#title">

<META NAME="DC.subject"
      CONTENT="(SCHEME=keyword) Dublin Core,
      Metadata, Warwick Framework, Resource Description,
      Resource Discovery">

<LINK REL=SCHEMA.dc
      HREF="http://purl.org/metadata/dublin_core_elements#subject">
<META NAME="DC.author" CONTENT="(TYPE=name) Paul Miller">
```

The URLs used in the LINK tag should go to an official repository of metadata schemes. In the case of Dublin Core Set, such a repository has been established at OCLC in Dublin, Ohio.

The following is partial GILS record, which has its repository at USGS.

```
<META NAME="AUTHOR" CONTENT="Eliot Christian">
<LINK REV="made" HREF="mailto:echristi@usgs.gov">
<META NAME="GILS.Originator"
      CONTENT="U.S. Geological Survey">
<LINK REL="schema.GILS"
      HREF="http://www.usgs.gov/gils/prof_v2.html#core">
```