

Table of contents

Introduction	2
What is metadata?	2
Definition	2
Example.....	2
Origin	2
Purpose.....	2
Example.....	3
Metadata have potential.....	3
Metadata buzz.....	3
The Semantic Web	3
A word of warning though	4
Metadata in practice	4
Metadata for a purpose	4
Define metadata.....	5
Overview of metadata standards and projects	5
Schema versus metalanguage	5
<i>Metadata schema (semantics)</i>	5
<i>Metalanguage (syntax)</i>	6
The Dublin Core Metadata Initiative (DC).....	6
<i>Overview of DC elements</i>	6
<i>The use of DC within HTML</i>	8
The Resource Description Framework (RDF).....	9
Topic Maps.....	9
PRISM.....	10
IMS.....	10
Bibliography and links	11

Introduction

In recent years we have seen the rise of several, often conflicting, projects and standards for cataloguing electronic resources. Though created for different purposes and by different groups, some of them are closely related and can be of great benefit when applied wisely. Welcome to the fast evolving world of metadata.

What is metadata?

Definition

Metadata is generally defined as 'descriptive information about information' and refers to any data used to support the identification, description and location of an information object, such as a document. Simply put, metadata is the collection of labels that describe a piece of information.

Example

Consider your collection of CDs or videotapes, If you had to play each one to verify what was recorded on it, then finding a particular song or film would be a long tedious job. If you have a label on each CD or tape with the title and a list of contents, then selecting the right one becomes easy. With a larger collection, then if you have a searchable list or a subject index, and the recordings are in groups with similar locations, then it becomes easier again. This is the basic motivation behind metadata.

Origin

The term metadata is not new. As long as people have been collecting information – be it in a library, a museum, or any other institution – they had to get hold of ways to properly organize that information. The catalog that originated in the traditional library world is the classical example of metadata. And its most commonly known fields are 'Author', 'Title' and 'Subject'.

Purpose

In the online world it is important to know whether a document is up to date or what the format of a file is so we have the correct software to open it. Metadata typically deal with such issues. Roughly speaking they serve, sometimes simultaneously, three functions:

- **semantic analysis**
data explaining the content of the information object: title, subject (or subject categories, taxonomies, ontologies), keywords, intended audience, content rating, and so forth.
- **administration**
data used for managing the information object: author(s) of the resource, reviewer(s), the version number, date to be reviewed, property rights, and so forth
- **access and publishing**
data that can be extracted directly from an information object: file name, size, extension, creation date and so forth.

These three categories are not mutually exclusive: administrative metadata could be used for access and publishing and we could also debate whether 'intended audience' is solely semantic and not administrative.

Moreover, even the distinction between 'data' and 'metadata' is not an absolute one; many times an information object will be interpreted in both ways simultaneously.

Example

The most commonly known examples of metadata are HTML <meta> tags and Microsoft Office Properties.



Figure 1: Properties window in Word

Up till now HTML-metatags on the Web have not been very successful: they have been so commonly abused for the purpose of higher ranking, that they are ignored by most search engines.

Metadata have potential

Metadata buzz

In the context of the fast growing number of digital resources, metadata have gained serious weight. Today, the Internet abounds with metadata projects. It is even tempting to proclaim 'metadata' as the next buzzword.

There is also a company - the Metadata Corporation – that has claimed a copyright on the terms "Metadata" and "METADATA" (with the first or all letters capitalized). The copyright is on pretty shaky ground in my opinion, nevertheless some have already started hyphenating the term (meta-data).

The Semantic Web

An important driver for this metadata activity is Tim Berner's Lee vision of the Semantic Web:

- 'The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.'
- 'The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people. (...) The Semantic Web is a vision: the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications.'

One of the key elements of Semantic Web technology is the use of well-defined, machine-understandable metadata. No doubt they could be an invaluable means for classification, browsing (metadata can be used to create navigational elements) and searching in the future:

- Consider the way searching on the Web works today; this (still) largely is a matter of matching query words with words in the text of a document (so-called free-text searching). Anything that makes the matching process easier or more standardized is bound to give results that are perceived to be of much higher quality than is the case with today's search engines. Metadata is expected to do so.
- Think of security issues on the Web; with a common format for metadata, a browser could be able to get an assurance, before imparting personal information in a Web form on how that information will be used. (In this regard, W3c refers to the 'Web of trust'.)
- In a call centre environment, where operators need to be able to answer client's questions rapidly and accurately, metadata-enriched information is a 'must have'.

A word of warning though ...

There are many metadata initiatives today and their number is growing rapidly, as different communities seek to meet the specific needs of their members. Some approaches are quite simple in their description, others complex and rich. Viewed on a continuum of increasing complexity, they range from the basic records like Microsoft Office Properties, through relatively simple formats like the Dublin Core, to highly specific formats like the Encoded Archival Description, an SGML DTD .

None of them has gained a standard status in the online community yet. In some cases there is even a duplication of effort. As a result, there is a lot of confusion among web content developers about which standards and approaches to follow. The resource description framework (RDF) for instance is a nice idea but search engines and web content developers are not using it. Even a basic catalog vocabulary like the Dublin Core is still uncertain. In fact, nothing beyond simple metatags has caught on with web developers today.

Some refer to the victory of Google, and its brute force technology, over metadata. Others argue: 'Why bother about metadata if full text search can do the trick?' In future, it will then also be interesting to see if the complexity of human language raises the bar high enough so that Semantic Web technology is actually simpler to work with.

Fortunately in the last two years, many people have realised that it is important to avoid or reduce duplication of effort and start looking for ways to either harmonize, align, map, or combine metadata standards and approaches.

There is also a significant school of thought that **metadata should be absolutely minimal**.

Metadata in practice

Metadata for a purpose

Metadata can be generated automatically, or created by humans. They can be queried by a user, or they can be used by software agents in service of a user.

Metadata can be associated with resources in various ways:

- They can be embedded directly in the information object: e.g. Office-properties or the HTML-metatags in a web page.
- They can be a separate entity linked to or from the object they describe.
- They can be stored in a remote database. The record in the database may either have been directly created within the database or extracted from another source, such as a web page.

Define metadata

To define metadata involves selecting a representative sample of the information, analysing the content and structure of the information and the information flow (author, target group, versioning...), and labelling the metadata. For each label establish whether:

- A value can be chosen freely.
- A value has to be chosen from a list.
- A unique value or more values are possible.
- The value is tied to a certain format (e.g. date).
- The value is compulsory or not.

Overview of metadata standards and projects

Schema versus metalanguage

To get a grip on the myriad of metadata projects that deal with online resources, let's make a distinction between **semantic metadata schemas** (dealing with semantics) and **meta-languages** (dealing with syntax).

Metadata schema (semantics)

Typically a **metadata schema** (or vocabulary) consists of a limited number of elements, and a name and a meaning for each element.

Some of the popular schemas include:

- DCMI (Dublin Core Metadata Initiative)
Set of 15 elements to describe document-like objects. More further in this note.
- GILS (Global Information Locator Service)
Set of Internet search requirements that can serve many purposes. Comparable to the Dublin Core.
- PICS (Platform for Internet Content Selection)
A set of labels to be associated with Internet content. Originally designed to help parents and teachers control what children access on the Internet, but it also facilitates other uses for labels, including code signing and privacy.

Syntax is not strictly part of the metadata schema. So schemas are unusable, unless they are embedded in a metalanguage. An example of a metadata schema that is embedded in a metalanguage is the TEI-header.

- TEI-header
The TEI-header is developed by the Text Encoding Initiative as part of an XML-DTD. The DTD is comparable to DocBook, but more academic (originally designed for poems, plays and literary texts).
The TEI-header consists of elements describing bibliographic metadata, and has a rich level of granularity. (TEI headers can be large and complex or simple).
The concept has been copied by other DTDs (EAD most recently), and could become a de facto standard implemented in other DTDs in the future.

Metalanguage (syntax)

A metalanguage is a syntax that defines rules on how a document can be described in terms of its logical structure (headings, paragraphs or idea units, and so forth). The best-known metalanguages are SGML, XML and HTML. Other important schemes are:

- **RDF (Resource Description Framework)**
See further in this note.
- **MARC (MACHINE Readable Cataloguing)**
A (complex) structure for encoding Machine Readable Cataloguing data (most often bibliographic and authority, but, as of late, several other kinds of data); widely adopted by librarians.

The Dublin Core Metadata Initiative (DC)

Probably the best-known metadata project today is the Dublin Core Metadata Initiative. The project, started in 1995 by a group of information experts at the Online Computer Library Centre in Dublin, Ohio, is the product of extensive consultation and experience in managing information resources. The result today is the DC Metadata Element Set, composed of semantic descriptions of 15 basic descriptive elements. The element set is considered a natural starting point for the definition of a more precise metadata standard to support resource discovery on the Web. At this moment the **DC is widely adopted in the government sector**.

The DC is different from other metadata schemas due to its **ease of use** and **interpretability**. The DC Metadata Element Set is meant for simple 'document-like objects' and not bound to strict rules: just the name (or identifier) of the element and the value of the element. The value consists of free text or it may be taken from a standardized resource (database). The description may reside in a separate file or it may be part of the resource itself. All elements are optional and may be repeated without any constraint.

Overview of DC elements

- **Element: Title**
Name: Title
Identifier: Title
Definition: Name given to the resource.
Comment: The name by which the resource is formally known.
- **Element: Creator**
Name: Creator
Identifier: Creator
Definition: Entity/authority primarily responsible for making the content of the resource.
Comment: Examples include a person, an organisation, a service, a company, a department.
- **Element: Subject**
Name: Subject and Keywords
Identifier: Subject
Definition: The topic of the content of the resource.
Comment: Typically, a subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal glossary.
- **Element: Description**
Name: Description
Identifier: Description
Definition: An account of the content of the resource.
Comment: Examples include: an abstract, an introduction, table of contents, reference to a graphical representation of content or a free-text account of the content.

- Element: Publisher**
Name: Publisher
Identifier: Publisher
Definition: An entity responsible for making the resource available
Comment: Examples of a publisher include a person, an organisation, or a service. Typically, the name of a publisher should be used to indicate the entity.
- Element: Contributor**
Name: Contributor
Identifier: Contributor
Definition: An entity responsible for making contributions to the content of the resource.
Comment: Examples of a contributor include a person, an organisation, or a service. Typically, the name of a contributor should be used to indicate the entity.
- Element: Date**
Name: Date
Identifier: Date
Definition: A date associated with an event in the life cycle of the resource.
Comment: Typically, Date will be associated with the creation or availability of the resource (date of publication). Recommended best practice for encoding the date value is defined in a profile of ISO 8601 and follows the YYYY-MM-DD format.
- Element: Type**
Name: Resource Type
Identifier: Type
Definition: The nature or genre of the content of the resource.
Comment: Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the working draft list of Dublin Core Types). To describe the physical or digital manifestation of the resource, use the FORMAT element.
- Element: Format**
Name: Format
Identifier: Format
Definition: The physical or digital manifestation of the resource.
Comment: Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types defining computer media formats).
- Element: Identifier**
Name: Resource Identifier
Identifier: Identifier
Definition: An unambiguous reference to the resource within a given context.
Comment: Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Example formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).
- Element: Source**
Name: Source
Identifier: Source
Definition: A reference to a resource from which the present resource is derived.
Comment: The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.
- Element: Language**
Name: Language
Identifier: Language
Definition: A language of the intellectual content of the resource.
Comment: Recommended best practice for the values of the Language element is defined by RFC 1766, which includes a two-letter Language Code (taken from the ISO 639 standard), followed optionally, by a two-letter Country Code (taken from the ISO 3166 standard). For example, 'en' for English, 'fr' for French, or 'en-uk' for English used in the United Kingdom.

- Element: Relation**
Name: Relation
Identifier: Relation
Definition: A reference to a related resource.
Comment: Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.
- Element: Coverage**
Name: Coverage
Identifier: Coverage
Definition: The extent or scope of the content of the resource.
Comment: Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names) and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges.
- Element: Rights**
Name: Rights Management
Identifier: Rights
Definition: Information about rights held in and over the resource.
Comment: Typically, a rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses *Intellectual Property Rights* (IPR), *Copyright*, and various *Property Rights*. If the rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource.'

The use of DC within HTML

```
<html>
  <head>
    <meta name="DC.Title" content="Metadata" />
    <meta name="DC.Creator" content="Geert Allegaert" />
    <meta name="DC.Subject" content="Metadata, Dublin Core, RDF, PRISM" />
    <meta name="DC.Description" content="This research note discusses metadata
standards and approaches" />
    <meta name="DC.Publisher" content="www.namahn.com" />
    <meta name="DC.Contributor" content="Bart Azijn" />
    <meta name="DC.Date" content="2002-04-24" />
    <meta name="DC.Type" content="Text" />
    <meta name="DC.Format" content="text/html" />
    <meta name="DC.Identifier" content="urn:my-nid:my-nss" />
    <meta name="DC.Source" content="http://www.namahn.com/ex-draft.html" />
    <meta name="DC.Language" content="en" />
  </head>
  <body></body>
</html>
```

The Resource Description Framework (RDF)

RDF files are XML files that conform to a restricted set of XML tags. As such RDF can be considered a syntax-model for processing metadata. RDF can accommodate metadata-schemes from different sources.

The use of DC within RDF:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <rdf:Description rdf:about="http://doc">
    <dc:creator>Geert Allegaert</dc:creator>
    <dc:title>Metadata</dc:title>
    <dc:description> This research note discusses metadata standards and
approaches </dc:description>
    <dc:contributor>Bart Azijn</dc:creator>
    <dc:date>2002-04-24</dc:date>
  </rdf:Description>
</rdf:RDF>
```

Topic Maps

Probably the most intriguing metadata project today is Topic Maps: a tool to organize and present complex information in a way that is optimised for **visual exploration**. The aim of Topic Maps is to combat information overload or, as Steve Pepper (the founding father of Topic Maps) refers to it, 'Infoglut'.

Many people have been through how similar Topics Maps are to RDF, and indeed:

- RDF and TMs have XML syntaxes
- RDF and TMs can address similar applications.
- RDF and TMs can be easily mapped

Topic Maps have a richer model than RDF that includes the visualisation of the **relations** and **associations** between concepts.

XTM, or how does it work?

XTM stand for XML Topic Maps. The main concept within XTM is a **topic**. A topic can be anything: a person, a process, a product, an entity, a concept, ... Each topic is linked to information resources that are relevant to that topic. Such resources are called **occurrences** of the topic. It is also possible to describe relationships or **associations** between topics.

```

<topic id="hamlet">
  <occurrence>
    <instanceOf>
      <topicRef xlink:href="#xml-version"/>
    </instanceOf>
    <resourceRef
alink:href="http://www.hypermedic.com/style/shakespeare/hamlet.xml"/>
    </occurrence>
  </topic>

<association id="plays-the-role-of">
  <instanceOf>
    <topicRef xlink:href="#casting"/>
  </instanceOf>
  <member>
    <roleSpec>
      <topicRef xlink:href="#role"/>
    </roleSpec>
    <topicRef xlink:href="#hamlet"/>
  </member>
  <member>
    <roleSpec>
      <topicRef xlink:href="#actor"/>
    </roleSpec>
    <rolSpec>
      <topicRef xlink:href="#sir-lawrence-olivier"/>
    </rolSpec>
  </member>
</association>

```

XTM today

Though Topic Maps are becoming an international (ISO-) standard, Topic Map technology is – at the time of writing – not yet mature enough for productive use.

If you want to have a glimpse of Topics Maps at work, then have a look at <http://www.kartoo.com/>.

PRISM

PRISM (Publishing Requirements for Industry Standard Metadata) is an RDF/XML-based metadata project, set up by the **publishing industry**. Some of the players involved are Adobe, Interwoven, Condé Nast Publications and Time Inc.

The PRISM-specification builds on Dublin Core by adding more detailed elements for discovery, rights management and distribution. As such it is particularly apt for news, magazines, catalogs and journals.

IMS

IMS is an all-in metadata approach that is being **used widely within the education sector**. Open standard, XML-based. Developed by The IMS Global Learning Consortium, Inc. (IMS) for facilitating online learning activities. Particularly apt for locating and using educational content, tracking learner progress, reporting learner performance, and exchanging student records between administrative systems.

Bibliography and links

- Demystifying metadata, nice introduction to metadata: <http://mappa.mundi.net/trip-m/metadata/>
- Metadata and search tools: <http://www.searchtools.com/info/metadata.html>
- Metadata principles and practicalities : <http://www.dlib.org/dlib/april02/weibel/04weibel.html>
- Metadata watch : <http://www.schemas-forum.org/metadata-watch/>
- Dublin Core Metadata Initiative homepage: <http://dublincore.org/>
- Dublin Core and RDF: <http://www.dlib.org/dlib/october00/baker/10baker.html>
- RDF and metadata : <http://www.xml.com/xml/pub/98/06/rdf.html>
- Topic Maps: <http://www.topicmaps.org/>
- GILS homepage: <http://www.usgs.gov/gils/>
- PRISM homepage: <http://www.prismstandard.org/>
- IMS Global Learning Consortium homepage: <http://www.imsglobal.org/>
- Critical note on metadata, detailing seven reasons while even the most promising metadata schemes will fail on the public web: <http://www.well.com/~doctorow/metacrap.htm>