Abstract

Metadata are an indispensable tool in environmental information systems, especially for finding and understanding environmental data. Despite their importance, however, collecting metadata for public environmental information systems is still an issue. After recapitulating the basics, this paper focuses on alternative approaches for collecting metadata and reviews related application areas. Eventually, several theses conclude the paper which aims at fostering the discussion in order to find out in how far metadata in environmental information systems may benefit from both proven and new approaches in other application areas.

1. Introduction

Metadata are an indispensable tool in environmental information systems, especially for finding and understanding environmental data. However, metadata are not restricted to environmental information systems, they are as well important in data warehousing, e-business standards and even in every day tools and gadgets such as digital cameras. Hence, this paper aims at encouraging the environmental community to have a view across the fence and try to find out what they can learn from best practices in further areas. In so far, it takes just the opposite perspective as the invited talk „Metadaten in verteilten Umweltinformation- und Erdbeobachtungssystemen“ on the 11th workshop of the IT team of the Global Biodiversity Information Facility [GBIF-D 2008].

The paper is organized as follows. In Section 2, we first explain the well established measures recall and precision from information retrieval which can be used for any retrieval task. We then recapitulate categorization and collection approaches for metadata. In Section 3, we review other areas in which metadata are used. Instead of an abstract, several theses conclude the paper in Section 4 which will hopefully stimulate the discussion.

2. Basics

In ISO 11179 metadata are defined as “The information and documentation which makes data sets understandable and sharable for users”. Metadata are used in order to preserve and to identify the underlying “real” data uniquely. Furthermore, they are the underlying source for information retrieval. Hence, it makes sense to recapitulate the basic measures in information retrieval.

2.1 Measures in information retrieval: recall and precision

In information retrieval, two measures, namely precision and recall are important. In order to understand these measures, we have to distinguish

- the complete set of documents, including both relevant and irrelevant documents,
the set of documents relevant for the specific query and
the set of documents that is actually retrieved.

Note that this set of documents retrieved might include irrelevant documents and that relevant documents might be missing. Hence, the challenge for any information retrieval system is to minimize both, the number of irrelevant documents included in the set of documents retrieved (so-called “false positive”) and the number of relevant documents that are not in the set of retrieved documents (so-called “false negative”). And this is exactly what the measures precision and recall reflect:

- Precision is defined as the fraction of relevant documents retrieved of all documents retrieved.
- Recall is defined as the fraction of relevant documents retrieved of all relevant documents.

Thus, high values of precision (close to 1.0) minimize “false positive” results whereas high values of recall (as well close to 1.0) minimize “false negative” results. As can be seen already intuitively, it is not possible to optimize both measures simultaneously: when trying to increase precision, i.e., trying not to miss any relevant documents, we will include more irrelevant documents, thus decreasing recall, and vice versa.

### 2.2 Categorizations

Metadata can be categorized as either dependant on the content or independent from the content. Content dependant metadata may either describe the content or may be required for correct interpretation of the data (e.g. measurement device used, its calibration). Content independent metadata is used for identification (e.g. keys) or administration purposes such as author and date.

Another important categorization is the distinction between using a controlled vocabulary (e.g. a thesaurus) or an uncontrolled one, such as free key words.

### 2.3 Collecting Metadata

There are several approaches how to collect metadata:

- personal specifically educated for the purpose (e.g. librarians),
- the “authors” of the data, or
- the users (“user generated meta data” [Mathes 2004])

may provide the metadata.

Specifically trained personal will ensure consistency, maintain a high quality and represent the view of a neutral 3rd person. On the other hand, the mere amount of data and the costs incurred by such 3rd persons are shortcomings of this approach.

Authors are familiar with their data, hence describing them with metadata is easy for them. On the other side, authors are likely to reflect only their potentially very specific view of the data which might not be sufficient or not even useful for (all) users.

With the advent of Web 2.0, user generated content in general and user generated metadata more specifically gained increasing popularity. As in the case of authors, a single user is likely to reflect only his or her specific view of the data. Thus, the challenge with any user generated content is to attract a sufficient number of active users in order to be more likely to achieve a certain quality of the content or – in our case – of metadata.

We conclude that there is no single golden bullet that fulfil all needs. Rather, a combination of different approaches is in place. With respect to environmental metadata, one of the challenges is to motivate a sufficient number of users to provide “their” metadata.
3. **Metadata in further application areas**

Libraries and their catalogues are probably the most well-known application area of metadata. However, metadata are indispensable in technical areas as well. Examples are location information in mobile phone networks and metadata in data ware housing that indicate the source and the date of the data.

### 3.1 E-Business

In e-business, business standards or (to put it broader) content related standards can be categorized [Berlecon 2003] bottom up as being used for
- product identification,
- product classification and description,
- exchange of catalogue data,
- transactions (exchange of business documents) and
- business processes.

The left side of Fig. 1 visualizes this hierarchy; the technical standards on the right hand side are out of scope for this paper.

![Fig. 1: Classification of e-business standards](https://example.com/figure1)

When we apply these layered categories to environmental metadata standards, we find nearly all of them in the bottom layers as they are used
- for identification, e.g. of measurement data, and
- for data description, e.g. based on environmental thesauri.

Exchange of catalogue data has been addressed in environmental metadata standards as well, basically by harmonizing catalogues on different levels of administrative hierarchies, e.g. on the European, on national and on regional levels, and by ensuring the interoperability of different geo data infrastructures.

However, there are no distinct environmental (metadata) standards on levels 4 and 5 addressing either transactions or business processes. In this context, transactions are still fairly simple sequences of actions such as placing an order, receiving the goods and then transferring the money. In contrast, business pro-
cesses encompass more complex workflows as they specify correct reactions upon certain actions, e.g. partial fulfilsments of orders, changing or even revoking orders partially or completely. Of course, a simple answer to the question why there are no environmental (metadata) standards addressing transactions or business processes might be that at least currently there is no need for them. But as current technologies such as Web services [W3C 2008] explicitly support autonomous (i.e. no human interaction required) communications among different systems and as Web Services are used in metadata systems such as NOKIS [NOKIS 2008], we can easily imagine complex interactions that could be automated or at least be better supported by protocols that already foresee correct reactions in given situations. Hence, we suggest to have a look at the respective e-business standards and to carefully analyse what we can learn from them for upcoming and future environmental metadata standards.

3.2 Metadata in Web 2.0

The notion of Web 2.0 basically tries to reflect a much more active user involvement than what was possible in Web 1.0. It is not, however, a strict definition [WikiPedia 2008, OReilly 2006]. Folksonomies or user generated metadata [Mathes 2004] as already mentioned in section 2.3 are directly related to our subject. Hence, we should ask how to motivate users of environmental data sources to tag them so that other users may profit.

One of the most successful examples of Web 2.0, both commercially and with respect to user involvement, are social networks such as Xing [Xing 2008] and LinkedIn [LinkedIn 2008] for business professionals, and facebook [Facebook 2008]. Obviously, there will never be a single social network that captures all of a person’s interest. Thus, interoperability among such networks as addressed by Google’s OpenSocial [OpenSocial 2008] is an important issue. Technically this approach is very similar to interoperability approaches for environmental data catalogues. Its economic importance, however, is significantly bigger.

3.3 Multi media

When using a digital camera, we maintain metadata without necessarily being aware of it: technical information on individual shots (e.g., camera, exposure time, aperture, focal length) are captured in EXIF headers or IPTC core headers. The so-called IPTC core is maintained by the International Press Telecommunication Council [IPTC 2008] whereas the Exchangeable Image File Format (EXIF) [EXIF 2008] is a de-facto standard that is not even maintained by any official organization.

The benefit of these approaches to metadata management is that technical metadata is maintained as an integral part of the data itself, namely the photos. Compared to a purely separate metadata management, this metadata cannot get lost.

A prototype system for the distributed generation of metadata for multimedia documents is described in [Rust, Prabel 2008]. [Multimedia-Metadata 2008] provides a more extensive overview of this field of multimedia metadata, esp. MPEG-7.

4. Some final theses

Several theses which will hopefully stimulate the discussion conclude this paper:
1. Technical problems including those ones related to distribution can be solved. In the past, CORBA [OMG 2008] was used. Currently, Service Oriented Architectures (SOA) and Web Services [W3C 2008] are the method of choice. We will see what comes next.
2. There will not be a single universal standard for a specific application area that is both comprehensive and practicable. Instead, it is important to ensure the semantic interoperability among several standards, e.g., by agreeing on a core (labelled as “profile” in several standards).

3. Currently, environmental metadata do not yet explore the full technical potential as compared to e-business standards.

4. Handling metadata as integral part of the data offers distinct advantages. However, to support e.g. retrieval sufficiently, an additional separate metadata handling (e.g. separate catalogue) might be necessary.

5. Technical metadata, i.e. metadata that are independent from the content, can initially be collected automatically. In case these metadata are (additionally) managed separately from the original data, they must then be propagated from the original data system to a separate catalogue system.

6. In contrast, collecting content dependant metadata initially is an intellectual task which, at least currently, cannot be automated.

7. For updating content specific metadata, though, automation is possible.

8. To take full advantage of Web 2.0 for environmental metadata, the challenge is motivate users to participate actively.

References

[Berlecon 2003] Berlecon Research: E-Business-Standards in Germany: Assessment, Problems, Prospects; A research project commissioned by the German Federal Ministry of Economics and Labor (BMWA); Final report (Short version); Berlin, April 2003


[GBIF-D 2008] Abschlussbericht der IT-Fachgruppe (GBIF-D), Januar 2004 bis Dezember 2007; Workshops zur Informationstechnik im Bereich der Biodiversitätsinformatik; Universität Bayreuth 2008


[GBIF-D 2008] Abschlussbericht der IT-Fachgruppe (GBIF-D), Januar 2004 bis Dezember 2007; Workshops zur Informationstechnik im Bereich der Biodiversitätsinformatik; Universität Bayreuth 2008


