Metadata for Web Portal

Nikos Manouselis
Informatics Laboratory, Agricultural University of Athens, Greece

Constantina Costopoulou
Informatics Laboratory, Agricultural University of Athens, Greece

Alexander Sideridis
Informatics Laboratory, Agricultural University of Athens, Greece

INTRODUCTION

Web portals can be defined as gateways to information and services from multiple sources (Tatnall, 2005). An important aspect of Web portals is the organisation, navigation, labelling, and indexing of their content in order to facilitate searching of information and services (that is, the resources stored as Web portal content). One of the aims of Web portals is to collect and categorize resources (otherwise called content objects), so that users can search, identify, and access the most appropriate resources for their needs.

Metadata plays a critical role in such systems. Metadata is defined as structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage a resource. It is often called "data about data" or "information about information" (NISO, 2004; Steinacker, Ghavam, & Steinmetz, 2001). Metadata is used to provide information about resources that do not necessarily need to be displayed on the screen. It can then be used by software such as search engines or content management systems. Examples of information commonly stored as metadata include authorship, publication date, modification date, copyright information, and subject keywords.

Metadata can be therefore used in the context of Web portals to describe resources, and thus to facilitate their categorization, storage, search, and retrieval procedures (Duval, Hodgins, Sutton, & Weibel, 2002; Miller, 1996). In this article, we provide an overview of what metadata is, and how it can be used for the description, categorization, and classification of Web portal content. Using the case study of an organic agriculture (OA) Web portal, appropriate metadata for describing OA electronic markets (e-markets) and developing an e-market directory service is presented.

BACKGROUND

Metadata is made up of data items that can be added to or attached to a resource. Such data items can be (1) machine-readable, giving software applications the data they need to interpret the information held on a resource, or (2) designed for human interaction, listing the creator, subject, title, and other data needed to find and manage the resource. These data items are better known as *metadata elements*. Three types of metadata elements have been identified (NISO, 2004):

- **Descriptive Metadata Elements:** Describe a resource for purposes such as discovery and identification. They can include elements such as title, abstract, author, and keywords.
- **Structural Metadata Elements:** Indicate how compound objects are put together. For example, how pages are ordered to form chapters.
- **Administrative Metadata Elements:** Provide information to help manage a resource, such as when and how it was created, file type, and other technical information, and who can access it. There are several subsets of administrative metadata elements, such as:
  - **Rights Management Metadata Elements:** Deal with intellectual property rights.
  - **Preservation Metadata Elements:** Contain information needed to archive and preserve a resource.

Metadata can be embedded in a resource or can be stored separately. More specifically, the following ways of associating metadata with resources have been identified (Duval et al., 2002): *embedded metadata* resides within the resource; *associative metadata* is maintained in files tightly coupled to the resource it describes; and *third-party metadata* is maintained in a separate database (termed as a metadata repository) by an organization that may or may not have direct control over or access to the content of the resource.

To use and benefit from metadata on the Internet, we need a common format for expressing it that should be designed for machines rather than humans (Steinacker et al., 2001). *Metadata schemas* (or metadata models) are sets of metadata elements designed for a specific purpose, such as describing a particular type of resource (NISO, 2004). The

definition or meaning of the elements themselves is known as the semantics of the schema. The values given to metadata elements are the content. Metadata schemas generally specify names of elements and their semantics. Optionally, they may specify content rules for how content must be formulated (for example, how to identify the main title), representation rules for content (for example, capitalization rules), and allowable content rules (for example, terms must be used from a specified controlled vocabulary). There may be also syntax rules for how the elements and their content should be encoded. A metadata schema with no prescribed syntax rules is called syntax-independent. *Metadata specifications* are well-defined and widely agreed metadata schemas that are expected to be adopted by the majority of implementers in a particular domain or industry. When a specification is widely recognized and adopted by some standardization organization (such as ISO), it then becomes a *metadata standard*.

There is no one all-encompassing metadata standard to be used in all applications. Rather, there are various metadata standards or specifications that can be adapted or "profiled" to meet community context-specific needs (Kraan, 2003). This conclusion has lead to the emergence of the *application profile* concept. An application profile is an assemblage of metadata elements selected from one or more metadata schemas, and the purpose of an application profile is to adapt or combine existing standards or specifications into a package that is tailored to the functional requirements of a particular application, while retaining interoperability with the original base schemas (Duval et al., 2002).

DEFINING METADATA FOR THE BIO@GRO WEB PORTAL

The Bio@gro Web portal provides online access to accurate and multilingual OA information and resources, as well as to mobile services for all actors involved in the OA value chain (e.g., organic farmers, distributors, retailers, food companies, agribusinesses, consumers, academics). This Web portal is being developed in the context of the European e-Content Programme project 11293 "Bio@gro" for information dissemination, and public awareness increase regarding OA (http://bioagro.aua.gr). The Bio@gro project is being implemented by a cross-European consortium, including nine partners from four European countries. The portal is currently under development, and it is expected to be launched in full operation by the end of 2006.

Metadata is used in the context of the Bio@gro portal to describe the OA-related resources, which are distinguished in several content categories. The resources are described and categorized in the portal databases. The metadata is being authored and maintained by the Bio@gro portal team. It is stored in a specially designed metadata repository, separately from the actual resources. Therefore, Bio@gro is engaging a third-party metadata storage approach. The main uses of the metadata descriptions in the portal are for descriptive and for administrative purposes; therefore, corresponding metadata elements have been defined for each content category. These elements have been adopted and specialized from existing metadata schemas, so that to facilitate both the reusability (e.g., in other Web portals) and the interoperability (e.g., with other database systems) of the resources' descriptions. Thus, the metadata schemas to be used in Bio@gro are application profiles of existing metadata schemas.

Bio@gro Content Objects

The major content categories of Bio@gro are OA information resources (such as news, events, reports, and other documents), educational resources (such as online courses, best practice guides, and educational videos on OA topics), e-government resources (such as online addresses of governmental organizations dealing with OA and Web sites of OA agencies), and e-markets with OA products and supplies. These content categories include resources that are termed as "Bio@gro content objects" or simply BCOs. A BCO is a single information unit that can be identified, collected, and described for the Bio@gro portal in a meaningful and useful (for the OA actors) way. The format of a BCO can be digital or nondigital. BCOs in a digital format are expected to be categorized in the portal, either collected from the Web or developed by Bio@gro. Other types of BCOs may be nondigital ones, including traditional information resources, such as books or articles in printed media like magazines or scientific journals. For nondigital BCOs, only their description will be provided in the Bio@gro Web portal. BCOs may also have related copyrights or permissions of use: some may be freely uploaded in the Bio@gro portal (no permission rights), and some may not (restrictive copyrights or permissions of use). For the latter ones, again only a description will be included in the portal. Interested users will have to access copyrighted BCOs according to the policy of the copyright holder, for example, through the Web site of the publishing house for a scientific paper.

Each content category requires the use of a particular metadata schema, to reflect the special properties of each type of BCOs. In Bio@gro, four different metadata schemas are being used for the description and classification of the BCOs (Bio@gro, 2005). These metadata schemas are the Dublin Core standard (DC, 2004), the IEEE Learning Object Metadata standard (IEEE LOM, 2002), the e-Government Metadata Standard (e-GMS, 2004), and the Dublin Core for E-Markets (DC-EM) metadata schema (Manouselis & Costopoulou, 2006). Each one of these metadata schemas has been specialized in order to become appropriate for the needs of the Bio@gro portal, creating four corresponding Bio@gro application profiles.