Discussion Paper

A UK e-Government Metadata Framework

Andy Powell <a.powell@ukoln.ac.uk>

www.ukoln.ac.uk

6 April 2000

Version 1.3

Introduction

Metadata is structured data about data. In the context of the development of UK e-government portal services, metadata is crucial. Metadata will underpin much of the functionality that a portal will deliver. Portal-type services effectively bring descriptions of documents, collections, services, people, organisations and other resources together with the particular needs of an end-user and use that information to broker access to a subset of the network services available to that user in the government sphere.

This document has been prepared as input to a meeting of the Metadata Working Group of the Information Age Champions. It is intended as a discussion paper, though it is hoped that it may also form a useful basis for the development of a UK e-Government Metadata Framework document.

The intention is to briefly enumerate a list of the key entities, or classes of objects, that need to be described in order to support the development of portal-type services. The relationships between those entities are also discussed. Having enumerated the list, an initial set of candidate metadata schema for each of those entities, based on existing metadata 'standards' wherever possible, is also offered. The list of entities, relationships and metadata schemas is not intended to be exhaustive, rather it is a starting point for discussion. It should also be noted that there is not currently widespread agreement about the metadata schemas that should be used to describe some of the entities discussed below.

This document primarily focuses on metadata schemas - structured sets of descriptive attributes and their associated semantics. It briefly discusses syntax but does not discuss the protocols that might be used to share metadata descriptions between the software components that will make up the e-government portal architecture. It is anticipated that such discussion will take place during the development of the UK e-government Interoperability Framework document.

It should be noted that metadata can support a number of functions including:

Resource discovery - metadata describing what is available, where it is, how it is accessed and how it is used.

Content ratings - metadata describing who a resource is aimed at and what quality it is.

Administration - metadata supporting the management of resources by their administrators and curators.

Preservation - metadata that ensures the long term maintenance and availability of resources.

Rights management - metadata recording the ownership, copyright, access conditions, etc.

E-commerce - metadata related directly to e-commerce (for example, describing how much a resource costs).

The function of the metadata discussed here is primarily to support resource discovery, though metadata supporting the other functional areas is mentioned in some cases. The intention is to keep the metadata schema as simple as possible. Clearly, metadata schemas supporting all these functional areas will be crucial to the development of Information Age Government services. Those areas that are not discussed in any detail here will need separate consideration.

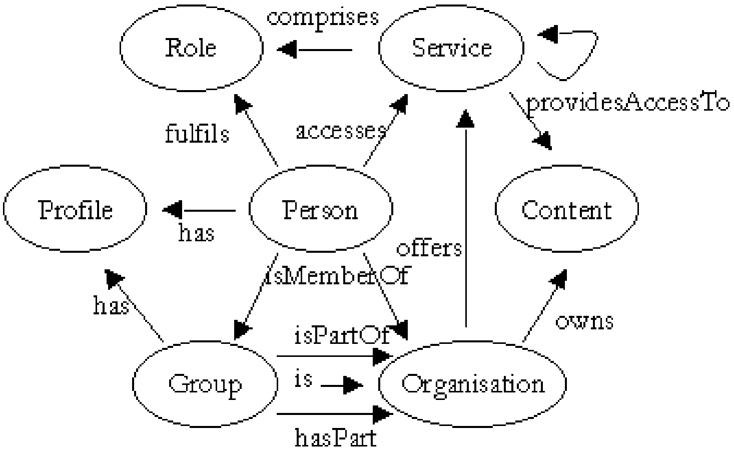
Finally, it is worth noting that the metadata described here is not targeted solely at human end-users. It is also targeted at software. Metadata will allow e-government portals to dynamically select which resources are required by the end-user. It will allow the portal to interact with those resources in an automated way on behalf of the end-user, based on the task in hand, the descriptions of those resources and knowledge of the end-user and the end-user's personal preferences.

A note on terminology

The term 'resource' is used in this document in a very general sense to mean anything that has identity and that is of interest. Two kinds of resources are of particular interest - resources that are 'content' and resources that are 'services'. Content includes documents, Web-pages, images, books, CD-ROMs, databases, etc. Services perform a function. In some cases that function is simply to provide access to content, for example a Web service, in other cases it is not, for example banking services, photocopying services, printing services, authentication services, etc. It should be noted that services may be physical (libraries, museums, telephone help-lines, etc.) or on-line (network services). Network services may be structured - supporting structured queries and returning structured results sets - or unstructured. Typically, current Web servers are unstructured network services, in that HTTP does not provide a 'standard' query language and HTML does not provide a way of marking up data (other than in the form of a document). This may change in the future with the development of an XML query language and the increased delivery of XML-encoded information.

e-Government entities

The diagram below shows the key entities (and a partial list of their relationships) that need to be described for the effective development of e-government portal services.



This entity-relationship model clearly needs further work! Note that all relationships are shown as uni-directional in the diagram. In reality they are all bi-directional, with converse relationships going the other way. For example, the converse of 'owns' is 'isOwnedBy'.

With reference to the terminology used above, service and content are both resources. A service may provide access to content, access to other services or it may simply provide a function. A portal is a network service that delivers a range of functions including providing access to a range of other structured or unstructured network services. Services may be on-line (e.g. Web or Z39.50 based services), physical (e.g. a library, museum or local government refuse-collection service) or hybrid (e.g. Inter-Library Loan with documents delivered by snail mail).

Person, group and organisation are represented in a simplified way in the above diagram. It's worth noting that each of these entities may be:

> an end-user of e-government portal services,
>
> the fulfiller of a particular role (e.g. telephone support line) within a service,
>
> the creator, contributor, publisher, owner or administrator of content and services.

A person may be a member of one or more organisations and/or other, more loosely defined, groups. A group may be composed of one or more organisations, may be part of a larger organisation or may simply be an organisation.

A person may have an associated profile. A group may also have an associated profile. A group profile may be inherited by any person that is a member of the group.

A person, group or organisation may own a variety of content and offer one or more services. In some cases, the same content will be accessed through several services offered by different people, groups or organisations, none of which is the owner.

Metadata schema

This section proposes some metadata standards that may be used to describe the entities listed above.

Content

Content resources need to be considered at a number of different levels. At the 'item' level, individual entities - Web-pages, images, documents of various kinds, sounds, videos, etc. - need to be described. However, resource discovery (and resource management) often happens at the 'collection' level, where groups of related resources are treated as a single unit. The relationship between collections and items may need to be made explicit in the metadata, allowing end-users to navigate between the two. The relationships between content and the services that provide access to that content also need to be made explicit in the metadata, enabling users to obtain and interact with the content they have discovered and allowing portals to do that on their behalf.

Dublin Core (DC) is a simple metadata attribute set that is primarily targeted at item level description of both physical and digital resources. DC provides 15 high-level descriptive attributes. There is currently some activity within the Dublin Core Metadata Initiative to allow more complex DC descriptions to be created than is possible by simply using the 15 attributes. DC activity is now quite widespread and there is government related DC activity in a number of countries around the world.

Although targeted at item level description, recent work within the Research Support Libraries Programme Collection Description Project in the UK has demonstrated that DC can be used successfully to provide collection level descriptions. Some additional, collection-specific, metadata attributes are required in addition to the DC 15.

There are certainly some alternatives to DC for the description of both items and collections. Many of them are able to provide richer, more detailed, descriptions though some might be limited to describing particular types of resources (e.g. books). However, DC provides a baseline level of resource description, suitable for supporting resource discovery. Experience shows that the richer metadata schemas can be mapped to DC with relative ease.

DC provides little support for the description of the rights (access rights, copyright and ownership) associated with resources. Further work will be required to specify e-government requirements in this area and to identify suitable a metadata schema. The European-funded INDECS project may provide a sensible place to start research in this area.

For content intended as an 'educational' resource, the schemas developed within the IMS framework could be used to enhance the baseline DC description.

Content ratings metadata should be based on the W3C Platform for Internet Content Selection (PICS). An e-government specific PICS vocabulary may need to be developed.

It is difficult to make firm recommendations about administrative metadata and metadata related to preservation currently. Work is ongoing in these areas, particularly in the area of metadata for preservation. The UK Cedars project should provide a useful starting point for research into metadata for preservation.

Recommendation: Resource discovery metadata based on Dublin Core plus additional collection-related attributes as necessary. IMS as appropriate. Rights management metadata based on INDECS. Ratings metadata based on PICS labels.

Services

On-line services can be described in a general way using DC. Protocol specific information typically can not be provided using DC, unless this information can be encoded in a protocol specific URI. This information is needed so that e-government portals can interact with services in an automated way. The information required typically includes 'location', the host name or IP address of the machine offering the service and a port number, availability and other protocol-specific information such as query syntax and result format.

Similarly, the description of physical services needs to include location information (a postal address for example), details about access (particularly for those with disabilities), hours of opening, etc.

Rights of access to content may be further limited by the rights of access imposed by a particular service. Likewise, any costs associated with access to content may be service specific. Portals may need to be able to determine an individual end-user's right to access services and content based on their profile (i.e. based on the metadata about that end-user).

Recommendation: Resource discovery metadata based on Dublin Core plus additional protocol-specific and rights-specific attributes.

People, groups, organisations and roles

Descriptive information about people, groups, organisations and roles is often referred to as 'white-pages' information. The development a white-pages metadata schema happened originally within the framework of the X.500 (and related) standards. More recently, development of the Lightweight Directory Access Protocol (LDAP) and the Virtual Business Card (vCard) specification have taken this work forward.

The vCard specification provides a schema for providing 'contact' information about people. The information is more extensive, though broadly similar, to the kind of information found on business cards. An associated vCard syntax is primarily targeted at embedding such descriptions within the body of email messages. The vCard specification does not currently support the description of groups, roles or organisations though future work may support this. Such descriptions are supported by the schemas developed for use with LDAP and X.500, though further work is required to determine which schemas are in widespread use.

Recommendation: vCard plus additional attributes based on existing LDAP or X.500 schemas or others as necessary.

Profiles

There are no widespread standards for user-profiles, though clearly, many Web-based services have developed 'proprietary' mechanisms for storing such information. In an e-government context, end-user profiles probably need to support a wide variety of information about the end-user. For example, home and business postal addresses, email addresses, preferred content delivery mechanisms (email, Web or postal for example), martial status, number of children, educational and health records, subject-area interests, payment preferences, willingness to travel, disability status, etc., etc.

Clearly there are significant privacy and data protection concerns in the storage of such information. However, profiles form part of the description of a person. In some cases, a personal profile will be inherited from the profiles of the groups of which that person is a member.

The e-government portal requirements for such profiles will need further investigation. It is not possible to make firm recommendations in this area at the current time.

The W3C P3P activity may provide useful starting point for research into privacy issues.

Syntax and management

This document did not set out to consider syntax issues in any detail. However, to state the obvious, multiple syntaxes are likely to be required. At the very least, the ability to embed a metadata description within the resource being described (where the content format allows it) seems desirable. With this in mind, it should be noted that DC metadata can be embedded within the head section of HTML Web-pages using the HTML <meta> element. Other content formats are less suitable for carrying embedded resource descriptions, though recently developed image formats may provide such support.

The Dublin Core Metadata Initiative is currently developing an encoding syntax for DC using XML, the eXtensible Markup Language, based on the Resource Description Framework (RDF). RDF is the W3C recommended metadata standard. Such descriptions will be suitable for embedding within HTML Web-pages, for providing external resource descriptions and as a metadata interchange format between the various software components that make up the e-government portal architecture.

This document makes no recommendations about how metadata should be created, stored and maintained. Just because metadata can be embedded into the resources does not mean that is a sensible place to create or manage the metadata. In most cases it is probably sensible to create and manage metadata separately from resources, using some kind of metadata database, embedding it into the resource on-the-fly as the resource is delivered to the end-user. Support for metadata in many content-management tools is limited currently, though this situation is likely to change in the future.

Conclusions

An e-government metadata framework needs to:

consider the entities that need to be described and the relationships between such entities in order to deliver e-government services,

consider the description of those entities with respect to each of the functional areas outlined above,

develop of a number of example usage scenarios to inform the above,

enumerate a set of metadata schemas that can describe the required entities in the required functional areas,

develop quidelines that encourage consistent cross-government usage of those metadata schemas,

identify suitable syntaxes for the encoding descriptions based on the above metadata sets,

consider the creation, storage, management and sharing issues associated with the application and use of metadata in the context of e-government systems.