

# Appendix B: Technical aspects of web-based searches for resources

---

## B.1 Introduction

It is relevant to consider a range of technical issues about searching for resources on the web. It is helpful to begin by considering the similarities with a physical library.

A library *holds* resources, such as manuscripts and journals. The defining characteristic of these resources is that they contain materials with which people engage—content. The data which makes up the content is characterised here as *Type 1* data. On the web, *Type 1* data are also resources, but the range of resource types is more diverse. The analogy to a library holding on the web is a *Resource Site*.

The difference between a library and the web is that in a library the resources are usually physically close, and their location and characteristics are well-defined. This is not the case with the web, which is like a library with millions of rooms, each with a number of ‘collections’. Not all collections are clearly labelled, and even when they are, they may be labelled in a way unique to each room.

In a real library, one usually finds a resource by looking it up in the library catalogue. The catalogue contains information about the physical location (room and shelf) where the resource may be found. The data which makes up the library catalogue is characterised here as *Type 2* data. On the web, *Type 2* data are provided by search engines. The analogy to a library catalogue is a *Search Site*.

Libraries long ago decided that standardised systems would enable people to locate resources more efficiently; hence the Dewey Decimal classification and Library of Congress subject heading schemes were established. These schemes established rules by which *Type 2* data could be organised to aid searching. These are characterised here as *Type 3* data. On the web, one instance of *Type 3* data is the metadata concept (see below).

While acknowledging the similarities between physical libraries and (teaching and learning) resources on the web, there are also significant differences.

- Unlike libraries, where holdings and catalogues are maintained at the same site, Resource Sites and Search Sites are independent entities, with two important implications:
  - The owner of a Resource Site has no obligation to notify Search Sites about changes to the resource. In fact, the Resource Site owners may not even know that a Search Site has exposed their Resource Site to Internet users by hyperlinking the resource within search results. This may lead to a ‘now it is here, now it is not’ situation.
  - There is very little incentive (if any) for a Resource Site owner to provide a proper description of the resource, even if they had the skill to do that correctly.
- Digital information can be reproduced at almost zero marginal cost. Unlike a library with physical holdings, the availability of a resource on the web only depends on the existence of the material. Copyright, therefore, becomes an issue, as discussed in chapter 5, Policy.
- The hyperlinking inherent in the web hides the location of a data resource. When a Search Site finds a resource matching the search criteria, hyperlinks give the illusion that the resource actually originated at the Search Site.

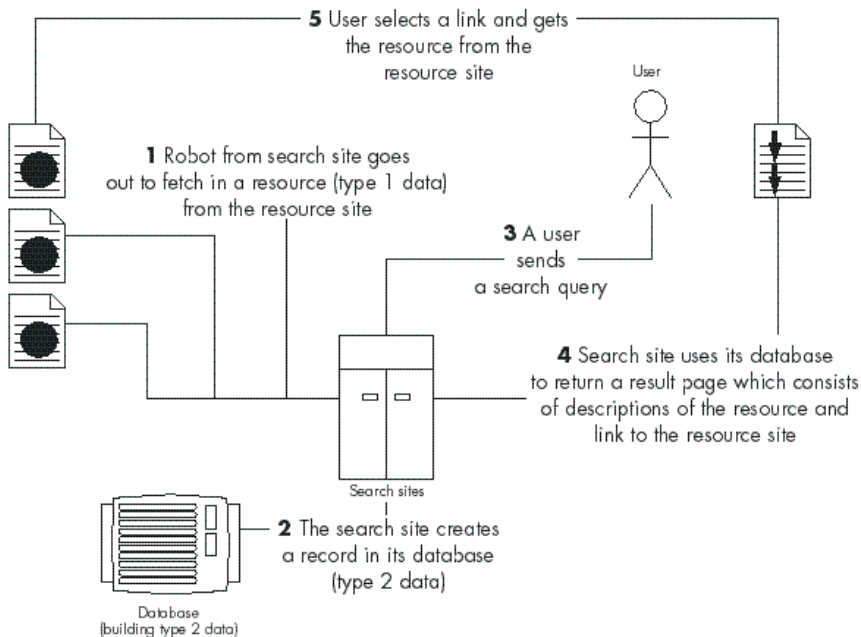
---

## **B.2 The anatomy of search sites—a data model**

Teaching and Learning Resource Databases are generally not Resource Sites, because they do not hold nor store the teaching and learning resources. Instead, these databases provide reasonably effective mechanisms for their users to discover useful and relevant teaching and learning resources. Teaching and Learning Resource Databases are Search Sites.

The logical structure of a Search Site is illustrated in figure A2.1. The functionality of Search Sites can be logically divided into three operations: *gathering*, *indexing* and *reporting*.

Figure B 2.1 Logical structure of a search site



Search Sites require mechanisms for locating resources (*gathering*). Some commercial sites (e.g. Yahoo!) depend on manual discovery of resources and encourage users to suggest resources. Other Search Sites (e.g. Alta Vista) send out software robots, known as 'spiders' or 'crawlers' to collect resources by following hyperlinks on documents (label 1 of figure A2.1). These robots locate web resources (Type 1 data) for further processing.

While some projects endeavour to store a snapshot of all currently available web resources, it is generally impossible for Search Sites to physically host all Type 1 data, because of the space implications of duplicating the files. Therefore, Search Sites generally create *indexes* from the gathered Type 1 data (label 2 of figure A2.1). The Type 2 data (locations and descriptions of resources) stored in the index is used to create a *report* when a user submits search criteria (labels 3 and 4 of figure A2.1). Users then have the option of retrieving the resource from the Resource Site (label 5 of figure A2.1).

In summary, instead of going out into the web to locate resources, by methodically exploring all Type 1 data, Search Sites scan the Type 2 data created by their robot(s) in order to find matching resources, and then report this to the user.

Table B 2.1 specifies clearly the conceptual differences between the three different types of data.

Table B2.1 Definitions of the three different conceptual types of data on web sites

Type 1 data	Type 1 data are resources in which users are interested. CFL resources are Type 1 data stored on Resource Sites. (In some ways, Type 2 and Type 3 data are themselves Type 1 data, because they are resources on the web.)
Type 2 data	Type 2 data are derived directly from Type 1 data. They are locations and descriptions of Type 1 resources. There are two main subtypes of Type 2 data: indexes and metadata (see below). The size and comprehensiveness of Type 2 data, and the relevancy of the collection to the site's users are the primary assets of a Search Site. It is unlikely that any collaborative framework that jeopardises the ownership of Type 2 data will be supported by Search Sites.
Type 3 data	Type 3 data cannot be derived from a single piece of Type 1 or Type 2 data. Instead, it is meta-meta information, like the Dewey Decimal classification scheme. Type 3 data includes the data that describes hyperlinks between documents; various metadata standards, such as the Dublin core and IMS standards; the usage logs in proxy servers about web pages; and the popularity rating of a web page among similar pages. Type 3 data are typically associated with a group of resources, identifying the relationships between resources. Type 3 data is analogous to the telephone. One telephone has limited functionality, but a collection of telephones enables communication and collaboration.

### B.3 Metadata

Many teaching and learning resources are not text-based, or may not have embedded hyperlinks, making the production of indexes through the gathering process a real challenge. It is necessary to have other mechanisms of describing such resources.

Technically speaking, indexes are inverted text indexes of HTML (or text) pages. Search Sites can produce a list of all the words found in a resource, removing all common words which may not be characteristic of the resource (such as 'a', 'an', 'the', 'is') and storing them in computer-readable format. When a user submits a keyword query, the keyword is matched against such word lists and the resources that contain the keyword are returned as the search result.

However, some Resource Site owners add their own descriptive keywords (metadata) to their resources, and some commercial Search Sites search on this metadata. Metadata is defined by (Milstead & Feldman 1997) as '...data about data. It describes the attributes and contents of an original document or work'. The DESIRE project (IBM 1998) describes metadata as 'data associated with objects which relieves their potential users of having to have full advance knowledge of their existence and characteristics'.

Metadata can be either embedded within the resource or stored separately in a database. It is typically authored by humans and is represented in some standard format, such as the Dublin Core <<http://purl.oclc.org/dc/>> or IMS <<http://www.imsproject.org/>> metadata specifications. Metadata is often represented in machine readable format, such as metatags in HTML documents or the Resource Description Framework (RDF) specification (Mason & Ip 1998). Table A2.2 compares the functionality of the index and metadata forms of Type 2 data, showing the advantages possible with the use of metadata.

Table B 2.2 Comparison of the functionality of indexes and metadata

Applicability	Index	Metadata
Applies to text resources	yes	yes
Applies to non-text resources, such as software or digital video	no	yes
Describes intellectual property ownership of the resources	no	yes
Describes conditions of use of the resources	no	yes
Supports machine understanding of the resource	no	yes

For metadata to be widely applicable, it needs to conform to standards of semantics and syntax, which allow, where possible, the use of natural language, and are flexible enough to cover a wide range of circumstances. Two major international initiatives are Dublin Core and the emerging Instructional Management Systems (IMS) standard. The EdNA metadata standard is an Australian initiative in the educational domain and is based upon Dublin Core. EdNA and IMS are currently collaborating in an effort toward harmonising the two standards.

### B3.1 Dublin Core

The Dublin Core metadata standard <<http://purl.oclc.org/dc/>> is an internationally recognised standard, which has largely been developed by experts from the library and information technology communities. It consists of a set of 15 elements, separated into categories of *content*, *intellectual property* and *instantiation*, as shown in table A2.3. The semantic meanings of these 15 elements are well defined and some of elements accept values only from a 'controlled vocabulary'. Under the Dublin Core standard, a metadata element's meaning is unaffected by whether or not the element is embedded in the resource that it describes.

Table B 2.3 Dublin Core metadata elements

Content	Intellectual property	Instantiation
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Source	Rights	Identifier
Language		
Relation		
Coverage		

### B3.2 IMS Standard

The Instructional Management Systems (IMS) <<http://www.imsproject.org/>> standard is also impacting the online education scene. This US-based initiative is in its third year of development. In the last twelve months its metadata foundation has become more closely aligned with the Dublin Core standard, although technically speaking it goes much further, in terms of the 'granularity' of metadata, in terms of interoperability, and in terms of complexity. IMS metadata is currently considerably more complex in conception than Dublin Core, and involves three 'schemas': *categories* (9) *data elements* (57) *abstract data types* (17). The IMS initiative claims that it is focused on the flexible management of online courseware, though some would argue that it is not well-matched to Australian pedagogical needs.

The overall stated goal of the IMS project is to 'enable an open architecture for learning'. IMS stakeholders are identified as learners, teachers, coordinators and providers. Key design considerations for online learning are identified as:

- granular content;
- scalability;
- interoperability;
- customisability and extensibility; and
- facilitation of and support for collaboration.

Importantly, the IMS specification involves not just metadata standards but standards that also relate to user profiling and other technical issues involved in the delivery of online education. However, while it is a significant international initiative with support from EDUCAUSE (an amalgamation of the US-based organisations EDUCOM and CAUSE in 1998), it is still very much in the prototype stage. Its importance is recognised by DETYA, and an Australian IMS Centre has been established at the University of New England to

spearhead Australian involvement in the development of IMS standards, coordinate IMS activities being conducted around Australia, disseminate information about IMS developments to the Australian educational community, and promote the use of IMS standards in Australia.

### B3.3 EdNA Metadata standard

Overarching strategies, such as the Dublin Core and IMS standards, cannot be all things to all communities and alone cannot cover all the data needs of Search Sites. Thus, other metadata standards, such as EdNA <<http://www.edna.edu.au/EdNA/genericpage.html?file=/edna/aboutedna/metadata/index.html&sp=eec099eccccb>> or MetaChem <<http://metachem.ch.adfa.edu.au/>> exist to define further, finer-grained elements which are relevant to the communities they support. As identified by case study participants, a teaching and learning resource database will need to provide information about a range of issues, including teaching context, referee's reports and evaluation data.

The EdNA metadata standard was publicly released in August 1998. In achieving agreement upon a standard suitable to all sectors, the EdNA metadata specification was based upon a minimalist approach. At that time, some debate prevailed as to what constituted standard use of 'qualified' Dublin Core. Also, there was considerable interest in deploying Dublin Core throughout other communities within Australia, such as government, libraries and museums. A wide acceptance of the Australian Government Locator Service (AGLS) standard <<http://www.naa.gov.au/govserv/agls/>> took place quite quickly and EdNA supported this whole-of-government approach. However, in releasing its first metadata standard, EdNA also flagged that further discussion focused on defining pedagogical information in later versions would have to take place.

In summary, the metadata specifications are Type 3 data which, when applied appropriately on Type 2 data, can enhance the machine understanding of Type 1 data. The primary asset of Search Sites is their collection of Type 2 data. However, while Type 3 data is useful in its own right by enhancing the service of Search Sites, standardisation is required to enable the efficient creation of Type 2 data, so that Search Sites can access the widest range of resources.

---

## B.4 Making use of metadata

Whichever metadata standard is used, it is necessary to associate the metadata (Type 2 data) with the Type 1 data to which it applies. This is achieved by either:

- embedding the metadata within the Type 1 data through the use of the <META> tag in HTML 3.2 and HTML 4.0 documents (only appropriate for text-based resources); or
- storing the metadata in a separate, detached resource linked to the Type 1 data (suitable for all types of resources).

The creation of metadata deserves special attention. Indexes can be created relatively simply by software robots, but are only appropriate for text-based resources. For other media, such as software modules and computer-aided learning packages, metadata is the only Type 2 data a Search Site can use. As shown in table A3.2, metadata is currently the only mechanism by which a database of CFL resources can contain the richness desired by participants in this study.

Unfortunately, metadata cannot easily be automatically created. It needs the knowledge and judgement of human beings, and metadata generation will be very labour-intensive. On the one hand, the original resource owner knows the resource best, and so may be seen as the most qualified to create the metadata associated with the resource. On the other hand, case study participants strongly made the point that they did not value information supplied by the developer, preferring instead an unbiased, third-party view. Special purpose metadata-editing software will need to be developed to isolate the academic subject expert from the semantics and syntax of the metadata. There are strong implications for staff development if widespread use of metadata elements is to occur.

The process of creation of metadata can be simplified by the use of metadata creation and maintenance tools, such as those available from the Distributed Systems Technology Centre (DSTC) <<http://www.dstc.edu.au/>>, a Cooperative Research Centre located primarily within the University of Queensland. Improved metadata tools are also being developed by EdNA, in collaboration with the higher education sector, in projects on: the improvement of the current metadata tool sets; and the production of a new tool to create metadata using retro-fit techniques for HTML documents.

Prior to the wide provision of metadata by Resource Sites, the primary asset of Search Sites was indexes (a form of Type 2 data) gathered by software robots. The quality of the result returned to end users depended largely on the characteristics of the Type 2 data owned by the Search Sites. That is, how

the Type 2 data had been created and how it had been presented to users, for example by the use of rating scales.

As metadata becomes more widely used, the role and assets of Search Sites will change significantly, because the Search Site may no longer own its Type 2 data. If the metadata is stored at the Search Site (detached from the Type 1 resource, but linked to it), then the Search Site owns the data, as before. However, if the individual metadata is embedded within the Type 1 resource, then it is owned by the Resource Site.

If a Search Site does not own the metadata, then it has no rights to modify it to add value to it. The asset of the Search Site will become the collection of the Type 2 data which it has created, coupled with the rating service it uses to help users to find appropriate resources. The collection of Type 2 data created by the Search Site is likely to qualify as a 'compilation' under copyright law (see paragraphs 2.8 and 4.1.6 of Appendix A). However, the expanding use of metadata stored separately from the original resource will require careful analysis of the legal situation. The enacting of moral rights legislation will also have an impact on the content of the metadata created by Search Sites (section 5 of Appendix A).