Chapter 1.

Introduction.

The aim of this project is to evaluate established and emerging resource description models on the World Wide Web, assessing their structural features and functional effectiveness as facilitators of resource description and retrieval.

Through evaluation of current metadata concepts and criteria, and investigation into user responses to metadata via practical research, the project will define essential metadata features and suggest improvements on current models.

1.1. Background to Project.

> 'The fundamental reasons for cataloguing remain. Within the system of information exchange, authors and creators want their documents to be found while users want their information relevant to their needs…'
>
> (Younger. 1997)

In the evolution of the Internet, a relationship has emerged between increasing network popularity and the initiative for GUI interfaces. The most recent stage of this process are the Web Browsers, such as Netscape and Internet Explorer which we use today.

What makes the WWW unique amongst other network interfaces is its accessibility to the general public. The growth of a massive user base on the network, with an estimated 700,000 Web sites in 1997 (Gill. 1998), is surely attributed to this drive for a user-friendly Internet format. What is clear, is that there has been a massive shift from academic and professional use to a commercial and personal interest user base as a result of increased Internet accessibility. Gwyneth Tseng, Alan Poulter and Deborah Hiom (1996), discuss this issue in *The Information Professional's Guide to the Internet:*

> 'The number of commercial sites connected has now overtaken the number of academic sites. The mid 1990s has seen a flood of companies making their first trial Internet connections, following an inexorable trend in the USA...'

The popular culture of the WWW has resulted in the proliferation of 'Web sites' on the Internet, and in the increasing demand for technology to index and retrieve sites for the end-user.

Despite the development of 'Web Crawler' software and the compilation of subject indexes by educational and commercial organisations, 'Internet surfing,' is still widely regarded as an inefficient means of information discovery on the Web.

This view is shared by many current authors and researchers interested in bibliographical structure on the WWW, including ADAM (Arts, Design and Media Information Gateway,) researcher, Tony Gill:

> 'The importance of descriptive catalogues for accessing and managing collections increases in proportion to the size of the collection being described. The lack of such a catalogue is one of the most serious drawbacks of the World Wide Web…'
>
> (Gill. 1998).

The emergence of Internet standards, such as the client/ server model, industry-standard communications protocols (such as TCP/IP and HTTP,) RDBMS database standards (such as SQL,) and application layer protocols, such as Z39.50, have facilitated the technical basis for new and improved methods of resource description and retrieval on the World Wide Web.

However, what is urgently required for the development of cohesive resource description, indexing and retrieval methods on the World Wide Web, are abstract cataloguing standards for the structure and format of resource description models, and the necessary tools and systems to facilitate those network standards.

Keith Trickey, Senior Lecturer at John Moores University, Liverpool, comments on the necessity for traditional cataloguing methods in imposing bibliographic structure and control on the World Wide Web:

> 'The desperate need for the classifier's sense of structure, and the cataloguer's grasp of consistent description are clearly evidenced in the opening pages of many sites...'
>
> (Trickey. 1998).

Networking standards are constantly evolving, and whilst all the information in this project is current at 10/08/'00, it is likely that the specifications and practices discussed in this project will undergo changes in the future.

Digital resource description is currently a lesser known technology amongst the majority of Web developers, and largely restricted to the information science community; however, with increasing publicity and support for this subject by the software industry and leading Information Providers, this situation could well change.

1.2. Project Scope.

This study considers the effectiveness of prevalent resource description structures on the World Wide Web, and evaluates their strengths and weaknesses as facilitators of resource description and retrieval.

The project begins by examining the most popular gateway or interface to the Internet, i.e.: Search Engines, evaluating how effectively these systems use resource description data to inform indexing and ranking processes. The project also defines *metadata*, in concept, and as a criteria for resource description models.

Additionally, the project evaluates abstract standards for expressing metadata, evaluating their scope and compatibility with recent cataloguing systems; the project also considers structures used to carry data, (i.e.: the transport layer), investigating markup languages used to express metadata content.

In addition, the project will analyse data from the practical research questionnaire, exploring responses to metadata use in the experiments, trends in metadata awareness, work practices and perceptions of metadata amongst HTML authors.

Finally, the project will provide an analysis of metadata as a whole, suggesting essential criteria for metadata, improvements on current models and conclusions on the present scope and effectiveness of metadata as a significant future technology.

1.3. Project Objectives.

The objectives of this project are as follows:

1. To assess the strengths and weaknesses of prevalent resource description and retrieval approaches on the World Wide Web.

2. To define current metadata concepts and criteria.

3. To broadly identify the various technical approaches to metadata.

4. To analyse emerging metadata formats, including markup languages and language-specific schemas, in terms of concept, structure, scope, depth, expansibility, interoperability and practical application.

5. To evaluate metadata initiatives, in terms of concept, scope, interoperability and practical functionality.

6. To assess awareness and perceptions of metadata amongst non-metadata specialist Web developers.

7. To assess metadata compilation and application accessibility for non-metadata specialist Web developers.

8. To define the scope for user-participation in metadata initiatives.

9. To suggest essential features for inclusion in evolving metadata standards.

Chapter 2.

Methodology.

This project comprises a literature search and a practical research element.

2.1. The Literature Search.

As a consequence of the growing demand for increased online bibliographic structure and control, many resource description standards have become widely used amongst Information Providers and the information science community.

As a result of increasing interest in this subject, a large volume of material on resource description initiatives, bibliographic standards and specialist commentary are now available on the World Wide Web.

Although the subject of online resource description has been discussed since the implementation of the first CERN network in 1989, the development of online bibliographic standards has not been widely documented in printed form.
Despite this, many information science journals have supported discussion on online bibliographic standards, and several books concerned with this subject are now appearing on the market. This study refers to a wide number of these journal articles,

and to several books concerned with online resource description, including *Introduction to Metadata,* edited by M. Bacha (1998), and *Designing XML Internet Applications*, by M. Leventhal (2000).

2.2. Media Formats.

A broad range of sources were consulted in the literature search, drawn from the following media formats:

- Specialist Books,
- Journal Articles,
- Online Resources.

2.3. Key Areas for Investigation.

The following key areas were investigated in the literature search:

- HTML script as a facilitator of metadata standards in the current environment of the World Wide Web.

- Metadata approaches amongst automated and directory-based Search Engine systems, including Alta Vista and Yahoo.

Continued Overleaf.

- The role of SGML-based markup languages as facilitators of resource description and retrieval, including XML and XHTML.

- Metadata structures expressed in XML markup script, including RDF and XML-Data.

- The role of schemas and script protocol standards, such as the XSL stylesheet standard and DTD architecture.

- Industry standard and proprietary metadata models, including the CIMI schema, the MESL specifications and the VRA Core.

- The role of metadata initiatives as facilitators of emerging metadata standards.

- The role of controlled qualifiers and vocabularies in the compilation and indexing of metadata.

2.4. Subject Specific Sources.

A variety of subject-specific sources were consulted on issues relating to the concerns of this project. Due to the limited availability of printed works on metadata, a broad range of IT related subject matter has been consulted to assist in the understanding of technology surrounding metadata. The broad categories of subject-specific material consulted are displayed below:

- Up-to-date writing on Information Communications Technology, with particular emphasis on Open System Interconnection standards, such as server database functions and Application Layer standards, such as Z39.50.

- Printed or online specifications for markup languages and related standards, including SGML, HTML, XML, XHTML, XSL, XPointer and XLink.

- Documentation on HTML development software, with particular emphasis on support for the Meta tag.

2.5. Journals.

The following journals were used to discover articles relevant to this subject:

- *The Journal of Librarianship and Information Science.*
- *Program.*
- *Information UK Outlooks.*
- *Vine.*
- *Library Trends.*
- *American Libraries.*
- *The Journal of Documentation.*
- *Library Review.*

2.6. Online Resources.

The following online resources were of particular importance in researching this project:

- The Dublin Core Homepage: http://purl.org/DC

This site contains the Dublin Core Specifications, outlining the history and structure of Dublin Core metadata; the site also provides an introduction to the Dublin Core Qualifiers, used to provide a standard syntax for Dublin Core content.

Continued Overleaf.

- The UKOLN Homepage: http://www.ukoln.ac.uk

This site contains links to a number of UKOLN funded projects and joint projects, including the UKOLN RDF (Resource Description Framework) Page, and Sirpac, an online editor to create RDF files; documentation is also available on the PURL project (Persistent Uniform Resource Locations), a technology to ensure stable URL addresses on the Internet.

- The WWW International Consortium (W3C): http://www.w3.org/Consortium/Process/#RecsW3C

The W3C is the most important standard-making body for the World Wide Web; this site contains the official W3C specifications for HTML (Hypertext Markup Language,) XML (Expansible Markup Language,) the Dublin Core and many related standards.

- *Ariadne* (Online Journal): http://www.ariadne.ac.uk

This online journal, funded by E-lib (the Electronic Libraries Programme), specialises in electronic information sources and related aspects of Information Technology. *Ariadne* also specialises in resource description and cataloguing standards in an online context.

2.7. Key Texts.

Key texts that particularly influenced this project included the following:

- Cromwell-Kessler, W. (1998). Crosswalks, Metadata Mapping and Interoperability. *In:* Baca, M. ed. *Introduction to Metadata.* USA, The Getty Information Institute.

This article provides a thorough overview of established and emerging resource description models, both in a traditional non-digital context, and in the context of online resource description; a detailed comparison of metadata formats is also provided in table form.

- Dempsey, L. and Heery, R. (1998). Metadata: a current view of practice and issues. *The Journal of Documentation,* 54 (2) March, 145-172.

This article provides a detailed explanation of metadata concepts and an overview of many metadata standards, including the Dublin Core, RDF and TEI. The problem of unreliable Search Engine indexing and ranking processes is also discussed.

- Gill, Tony. (1998). Metadata and the World Wide Web. *In:* Baca, M. ed. *Introduction to Metadata.* USA, The Getty Information Institute.

This article outlines key metadata concepts, criteria and structural approaches to metadata in an online context.

Continued Overleaf.

- Miller, E. An introduction to the Resource Description Framework. [Online]. (2000.) Cited 01/5/'00. http://www.dlib.org/dlib/may98/miller/05miller.html

This online text provides a detailed description of RDF, the *Resource Description Framework*, and outlines key features, such as *node* and *value* functions, and *resource mapping* to describe a range of resources.

- Trickey, K. (1998). Information Organisation on the Web? It is basically about respect and trust. *Library Review,* 47 (2) 135-137.

This article defines metadata concepts, and outlines many key problems with current bibliographic standards on the World Wide Web; in particular, this article suggests the necessity for traditional cataloguing and classification functions in an online context.

- Younger, J. A. (1997). Resource description in the digital age. *Library Trends*, 45 (3) 462-87.

This article defines current problems in Search Engine indexing and retrieval systems, using detailed case studies to illustrate levels of effectiveness in keyword searching. The article also considers the causes of inadequate indexing and ranking processes, including lack of Search Engine support for advanced metadata structures.

2.8. The Practical Research Element.

This project also comprises a practical research element, investigating the status of metadata amongst the HTML user community. Key issues for investigation include user awareness of metadata standards, user practices, user perceptions and user responses to metadata compilation following a practical experiment.

The practical research will be quantitative in approach, and will consist of a practical experiment, to be completed by the respondent, and a conventional questionnaire element. The experiment and questionnaire will comprise a single document, either completed in printed form or via email.

The objectives of the practical research element reflect the overall aim and objectives of this project. Issues for investigation in the practical element include the following:

- Awareness of metadata standards amongst HTML authors. (Objective 6.)
- Perceptions of metadata standards amongst HTML authors. (Objective 6.)
- Metadata compilation and application accessibility for HTML authors. (Objectives 1, 4, and 7.)
- Metadata format interoperability with prevalent networking standards and working practices. (Objectives 1, 4 and 7.)
- Script protocol transparency with prevalent markup language syntax, conventions and structures. (Objectives 1, 4 and 7.)

2.9. The Experiment.

The experiment consisted of two small experiments, *Experiment 1*, and *Experiment 2*. Each experiment introduced the respondent to a particular metadata standard, providing comprehensive examples and guidelines on the conventions and syntax of each format.

Following the introduction and examples, the respondent was asked to either write or type a description of any Web page using the conventions and syntax described.

The first section, *Experiment 1*, asked the respondent to create simple Dublin Core META tags to describe a Web page; two elements were included in the Dublin Core data: *keywords*, and *description.*

The second section, *Experiment 2*, asked the respondent to create a brief XML document describing their Web page, using simple XML *root* tags, and *element* tags to define element classes and their content.

In the following questionnaire section, the user was asked for responses to their experiences in the experiments, (A copy of the questionnaire with experiments is provided in Appendix B.)

2.10. The Questionnaire.

The questionnaire element followed the practical experiment, and consisted of four sections :

1. Questions about the respondent and their experience using HTML.
2. Questions about the respondent's contact with Metadata.
3. Questions about Experiment 1 (Dublin Core.)
4. Questions about Experiment 2 (XML.)

A combination of question types were used in the questionnaire, including scale-based questions, open-ended questions and multiple choice questions.

The Likert model (P. Marshall. 1997), was used to provide a standard format for scale-based questions providing ordinal data. Options were circled, or highlighted using a colour other than black if completed via email.
This scale format is expressed in numeric terms, using a scale of one (highest) to five (lowest), as illustrated by the following example:

7.) Generally, how comfortable were you creating Dublin Core data?
Please circle one of the following.

Very Comfortable 1 2 3 4 5 Least Comfortable

The questionnaire also used multiple choice questions, allowing the respondent to select more than one category. The proportion of positive instances within each category was discovered by counting the category total, thus transforming these nominal results to quantifiable numerical data.

Multiple choice questions of this kind were used to discover which formats respondents had heard of and had used, as shown in the following example:

3.) Please indicate which of the following *metadata* models you have heard of by ticking the corresponding boxes below.

HTML <META> tag	☐	RDF	☐
The Dublin Core	☐	XHTML	☐
XML	☐	None of these.	☐
		Others (Please specify below)	

_____

Open-ended responses required a process of interpretation and categorisation to allow transformation from nominal to quantitative data (Fink. 2000). In most cases, comments could be grouped into distinct categories, allowing proportion calculations.

In the following example, the respondent was asked for opinions on the XML experiment:

> 22.) Please describe any issues of interest or difficulties encountered during this experiment.
> **If no, please go to question 23.**
>
> _____

Questionnaire results and analysis will be presented in chapter 10.

2.11. Sampling Methods.

The sample for the practical research element consisted of 10 individuals.
The choice of the sample was purposive and structured (Mason. 1996), due to the specialist criteria for individuals required.

2.11.1. Sample Population Criteria.

The sample was composed of Web developers with at least a working knowledge of HTML. This category of individuals included both professional and amateur HTML authors, and those using HTML for a variety of personal and work-related purposes. Whilst individuals from an information services background were invited to participate, participation by information scientists and metadata specialists was avoided.

The reasons for my choice of target group are as follows:

- To include a wide variety of HTML users, ranging from IT professionals to private HTML users, thus discovering responses to metadata amongst a realistic sample of the Web development community as a whole.

- To exclude information scientists, and metadata specialists, since the practical research aims to discover the status of metadata amongst the majority of Web developers, rather than the much smaller information science community, whose responses would inevitably be more informed on metadata technology, and would not reflect the responses of the majority.

2.11.2. Demographics.

Web developers were chosen from personal acquaintances living and working in the following locations:

- Buckley, North East Wales.
- Wrexham, North East Wales.
- Chester, Midlands.
- Liverpool, North West.
- Sheffield, North Central.

2.12. The Pilot.

The experiment and questionnaire was piloted amongst a small group of Web developers based in the Wrexham area; consultation with the respondents revealed several potential problems with the questionnaire, which resulted in several amendments, including the following:

- Changes to phraseology used in the scale-based questions.
- Additional explanations and examples provided in questions on the experiments.

2.13. Methods for Analysis..

Primary, or unidimensional results (Fink. 2000) were compiled from questionnaire responses. Where appropriate, nominal responses were categorised and transformed into quantitative data. The following methods were used to compile basic unidimensional statistics:

- Descriptive Statistics.
    - Frequency counting,
    - Proportions,
    - Range,
    - Percentages.

Primary results were interpreted and analysed using the following methods:

- Measures of Central Tendency and Dispersion.
    - Mean averages,
    - Mode averages,
    - Median averages,
    - Range,
    - Standard Deviation.
- Measures of Correlation.
    - Cross-Tabulation,
    - Proportional Comparison.

2.14. Statistical Data Software.

Microsoft Excel was used to record and graphically present results. Excel was chosen for its interoperability with other Microsoft Office applications, and wide range of internal and user-defined functions.

Excel formulae was also used to calculate equations for the questionnaire results and analysis, (an overview of the most commonly used equations is provided in Appendix G.)

Chapter 3.

Search Engine Resource Indexing and Control.

Search Engines are the most popular form of gateway or interface to the World Wide Web today; they provide an intuitive graphical interface for users to query resource indexes, and obtain hyperlink listings for resources matching their query.

Search Engines can be though of in terms of two distinct components, comprising the *Parser Interface*, and *Resource Indexes* (Gill. 1998.)

The *Parser Interface* comprises the user-input stage of the search process, where keywords or controlled vocabulary is entered into an online applet application; user-input is used as criteria for selecting matching resources, or *hits*, and these are ranked according to relevance factors.

Resource Indexes contain a variety of data elements extracted from Web resources; indexes may comprise a single database, or several sub-indexes containing distinct resource data, such as full HTML content, resource titles or resource keywords.

The main difference between most Search Engine indexes is the preference of either manual resource discovery and indexing (called the *directory* approach,) or the use of autonomous resource discovery and indexing systems.

Directory Search Engines, such as Yahoo, use manual resource selection, indexing and ranking methods, based on the discretion of a human, rather than machine agent,

and may also use proprietary resource description standards to classify resources within useful subject-specific indexes. The only disadvantage with manually-compiled indexes is the inevitable omission of many new resources, due to limited manpower.

Figure 1. The Yahoo Parser Interface.

(The preference for directory based architecture is evident in the use of categories to direct the casual browser.)

3.1. Indexing Processes.

Most Search engines use the automated approach, where resources are drawn for indexing on a monthly basis by autonomous resource discovery software called *Web Crawlers*, or *Bots*.

The indexing process comprises two distinct aspects: firstly, the method of discovering Web pages themselves, and secondly, the method of storing information extracted from the Web page in resource indexes.

In the case of directory-based systems, the resource discovery method relies on either manual submission by Web authors, or Web page selection by a team of administrators; in automated systems, however, the discovery process in based entirely on the interrogation of URL links contained in Web pages already archived. When a Web page is indexed by an automated system, the URLs contained in the page, in the form of hyperlinks, are added to a single database. It is this URL database which the Web Crawler uses to discover and index pages each month (this process is known as *spidering*.)

Automated indexing systems are known to harvest millions of Web pages each day, as in the case of Alta Vista, which harvests up to 10 million pages each day using the automated 'Scooter' system (Morris. 2000.)

The viability of the hyperlink method for page discovery is controversial; on one hand, many thousands of new URL hyperlinks are discovered each time indexing takes place, on the other hand, pages indexed in this way lack any kind of human or

automated quality assurance; this problem is accentuated by the sheer volume of additions.

In the words of the Alta Vista homepage, (2000) this is a ‘blind’ and ‘random’ procedure, lacking any serious form of quality control:

> ‘Scooter, (the Alta Vista Web Crawler) sends out thousands of HTTP requests simultaneously like thousands of blind users grabbing text, pulling it back, and throwing it into the indexing machines so that text can be in the index. No human filtering or judgement is involved... a random game with hundreds of millions of Web pages.’

The second stage in the indexing process involves the method of data extraction and storage itself. In the case of both directory and automated systems, the entire page content is usually stored in a file and added to the server index. The file comprising the indexed resource usually contains a combination of extracted words and formal markup syntax.

As an example, the Alta Vista system indexes all the words in every page, while the Google system excludes some common words and HTML script.

Following the indexing procedure, some Search Engines perform a resource categorisation or description process, this is most commonly seen in directory systems, such as Yahoo, which assigns each page to one of 25,000 categories.

In other systems, however, there is often very little formal categorisation or description of content beyond the use of structural categories already existing within the HTML script itself.

3.2. Parser Functions.

Another critical feature of Search Engine technology is the *parser interface*; these range from systems providing limited, or no control syntax, as in the case of Yahoo, to those providing rich and complex command structures for query input.
Often, the user has the choice of using either a standard or an advanced interface.

Standard interface features normally include natural language input, where the user query is broken down, or tokenized into interpretable data for resource selection and ranking.
Additionally, standard features include *wildcard truncation*, using an asterisk to search using string elements, and *phrase searching* to discover text contained in quotation marks.
Often, Boolean and other logical operators are available in the standard interface, although additional logic operators are reserved for the advanced interface only.
The logical operators available in Lycos are perhaps the richest available, providing the following:

AND To require the inclusion of the string prior to, and following this operator.
OR To require either the string prior to, or following this operator
NOT To exclude the string following this operator.
NEAR The last string specified must be within 25 characters of the next string.

Continued Overleaf.

ADJ The last string specified must be next to the following string.

- To exclude the string following this operator.

+ To require the string following this operator.

( ) [Parentheses] Used to construct logical routines; for example, the following query will retrieve pages containing the words ‘Austen’ and ‘criticism,’ and pages containing the words ‘Austen’ and ‘reviews,’ combining what is in effect two sets of search results for the same query:

(Austen and criticism) or (Austen and reviews)

Many Search Engines allow the searching of elements contained in HTML script; this is mainly achieved using *metastatements* which are inserted into the parser box, sometimes alongside conventional text queries.

Alta Vista, Google and Hotbot provide a wide range of metastatements; the common syntax for metastatements is an element type, such as ‘applet’, followed by a colon and the required content.

Although metastatements are used to interrogate the full text of indexed pages, and do not constitute any form of descriptive classification, they do provide a means of querying pages by structural content.

These are a few of the most commonly used metastatements:

host: [hostname]	The host computer, usually specified using an IP address. (See Glossary, *TCP/IP*.)
domain: [domain name]	The name of the domain where resources are located.
link: [URL]	A URL contained in the page.
anchor: [hyperlink]	A link to any Web resource, such as a picture or MIDI file.
image: [filename]	An image filename, such as a GIF, JPEG or Bitmap file.
title:[title name]	A title string from the HTML TITLE tag.
applet: [filename]	A Java applet filename, such as a JAR file.

The resource language may also be specified, usually using pull-down menus to select from a range of languages; this data is usually extracted from the <LANG> tag, containing an ISO language abbreviation; however, since HTML is not language dependent, and few HTML documents contain this tag, the effectiveness of language preferences are often limited.

Whist logical operators, metastatements and other preferences can be helpful in discovering resources, it must be noted that these features are only as efficient or practical as the index architecture and content they support.

It must be remembered that the use of keywords and logical operators have the potential to match incidental words or phrases having no relation to the sense or semantics of the query terms.

In addition, the metastatements are useful only to locate resources where the exact filename or URL is known.

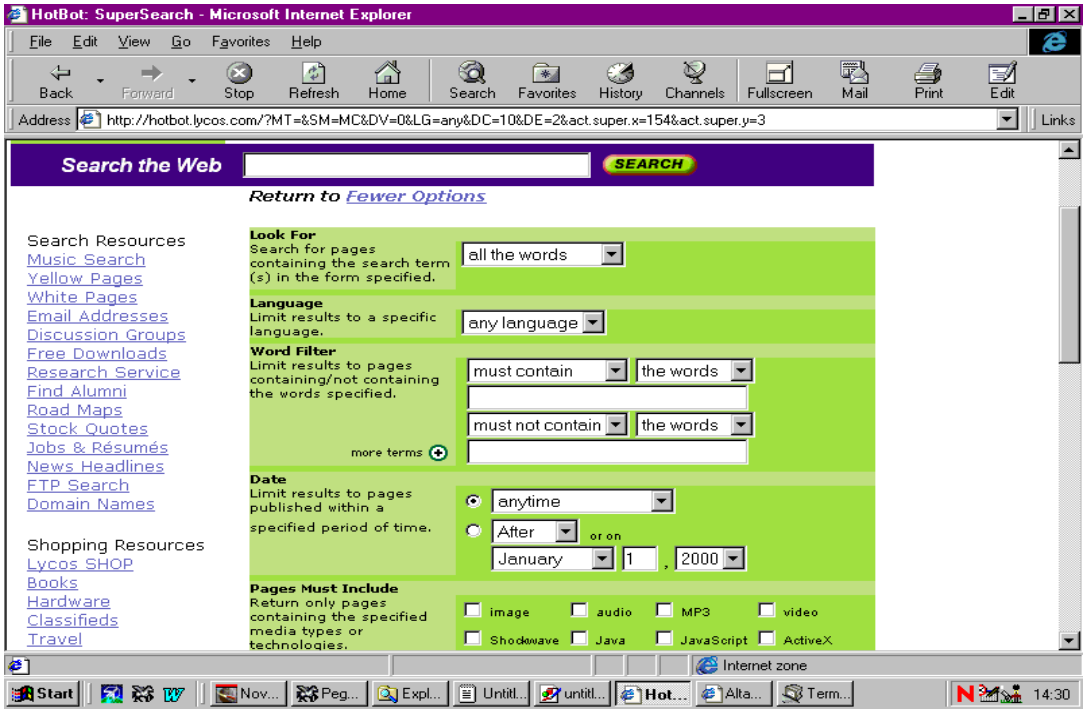


Figure 2. The Hotbot Advanced Interface, featuring language, date range and media type selection options.

3.3. Ranking Procedures.

Following a user query via the parser interface, the Search Engine scans through millions of full text data from indexed pages, attributing a relevancy rank to each page. In some cases, as in Lycos, the user may refine the importance of ranking factors themselves; in others, such as Yahoo, there are options to refine the search to commercial or non-commercial sites, and by location.

In most cases, the user may specify the number of results that will be found, as opposed to simply reducing the number of items displayed, thus hopefully reducing the number of irrelevant matches.

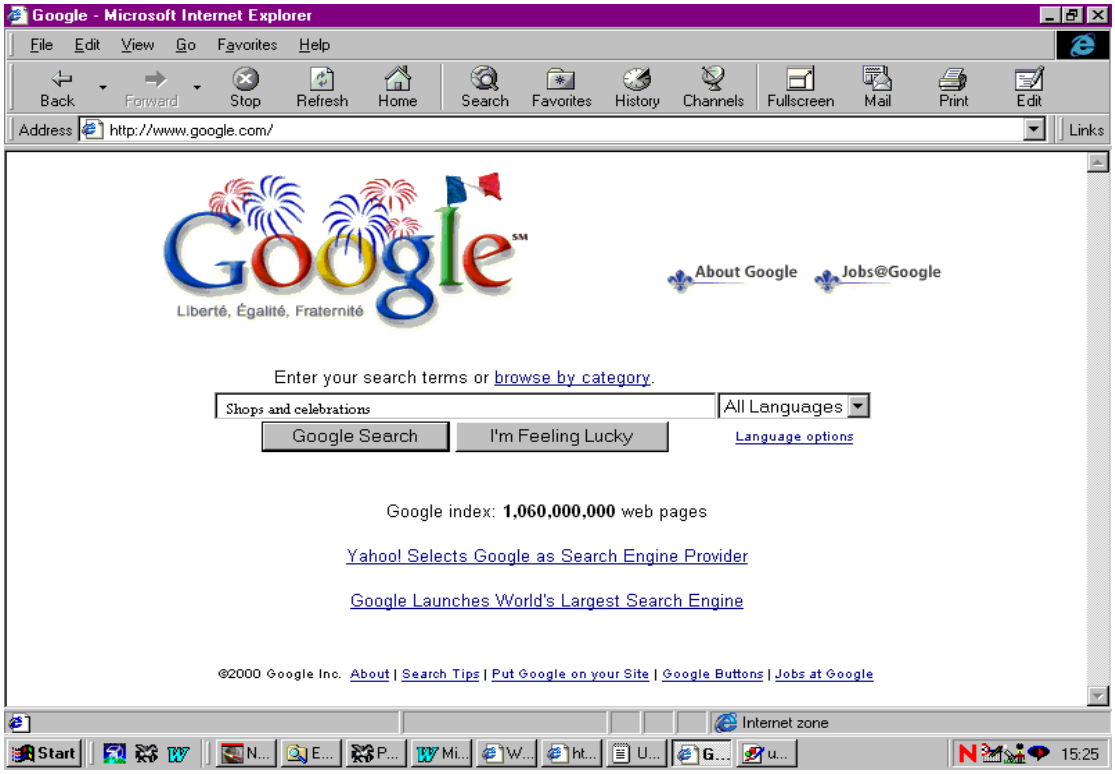


Figure 3. The Google Standard Parser Interface, with Language Pull-Down Menu.

One of the most common methods of relevancy selection and ranking used by Search Engines, is the frequency of query elements contained in the TITLE tag. Of secondary importance to this factor, is the frequency of user query elements contained in the entire HTML content of the Web page. Some, but not all Search Engines also factorise the frequency of query elements in the <H> header tag, in the resource URL, and in the first two to three lines of text found in the Web page (usually the HEAD section.)

Other Search Engines use different approaches to ranking. The Google system attributes a static value to each page indexed; this value is based on the number of

times indexed pages contain hyperlinks to other indexed resources. This arbitrary score, or *Page Rank* is used as an alternative to the element frequency approach. Another ranking method used by Search Engines is *metasearching*, i.e.: searching several Search Engine indexes for a user query; this approach is used by systems such as Savvysearch and Dog Pile, providing a large number of resource hits of mixed quality. Inevitably, the relevance of resources is mixed, since several Search Engine indexing methods are used to match the query.

3.4. The META tag.

Another ranking factor used by some Search Engines, is a simple resource description schema using the META tag supported in HTML v.4.
These tags enable the HTML author to include class-specific descriptive elements in their document for potential discovery and indexing by compatible Search Engines. *Meta tags* are inserted into the 'head' of HTML script; two standards for the use of this tag have gained widespread use amongst leading Search Engines, these are the *Keywords and Description* tags; these tags are illustrated below:

<META> NAME="Keywords" CONTENT="Culture, History, Music"</META>
<META> NAME="Description" CONTENT="A page devoted to favourite music and literature."</META>

The *keywords* tag is used to provide basic descriptive elements for use in page indexing and ranking procedures.

The *Description* tag is used to provide a resource summery for each hit displayed in the search ranking table.

Whilst the META tag has provided a standard for Web resource indexing and description processes, it has been widely criticised by Search Engine administrations; the main reason for this lack of support lies in the widespread practice of *keyword spamming,* i.e.: including large numbers of keywords on many subjects to facilitate inclusion in search hits. Commercial organisations have been cited as the main culprits in this activity.

At present, only Hotbot is known to attribute significant importance to the *keywords* meta tag in the ranking process, whilst other Search Engines claim to factorize this tag, but actually attribute more importance to other factors, such as the TITLE, HEADER and other HTML elements, such as phrased text (in quotation marks). Many Search Engines have included limited support for META content in ranking processes, mainly due to pressure from the Web developer community for a resource description standard; however, as Search Engine staff point out, there is still not enough evidence to suggest that enough people are using the tag to give it ranking precedence, and it is too open to abuse in the form of spamming to make it a viable alternative to present methods. Alta Vista (2000) confirms this view :

> 'Consider the opportunity for abuse and spamming... Those words are worth little more than any other word in the main text of the page.'

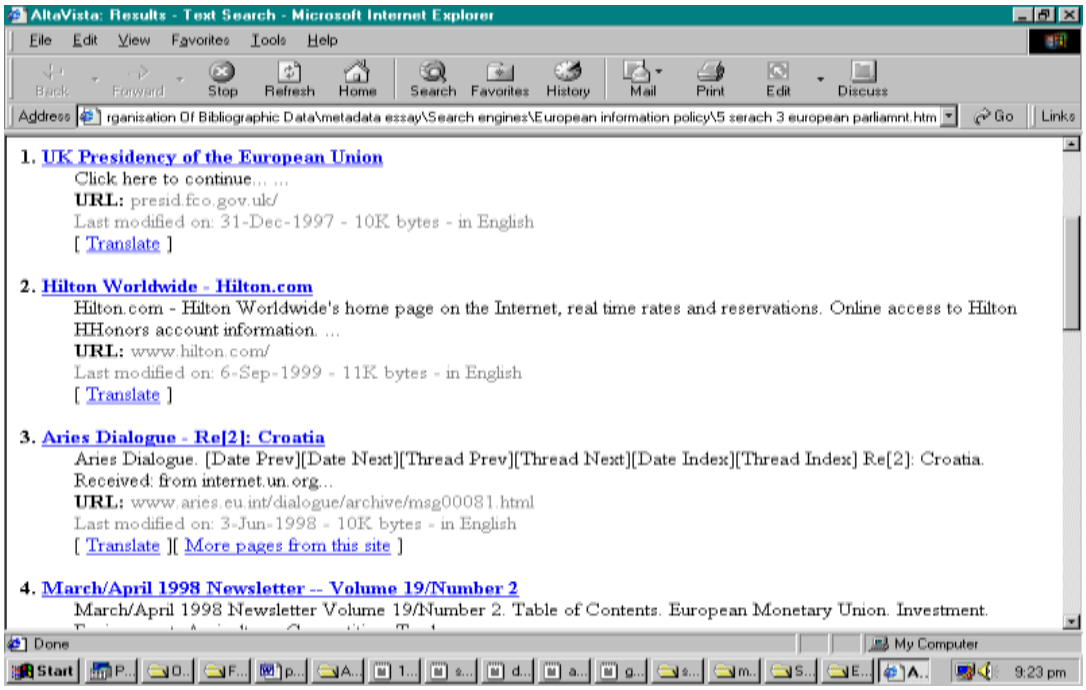


Figure 4. An Alta Vista Results Table. Because none of these pages contain the META *description* tag, the first few lines of HTML are used to provide a summary.

3.5. Conclusions.

What all the Search Engines seem to have in common, is a dependency on structures within the HTML script itself, such as the title, header, URL or tag contents used either in the ranking process, or specified using *metastatements*.

Whilst conventional Search Engines are good at extracting structural information about Web pages, i.e.: elements comprising the page, such as pictures, applet files etc. and some descriptive information, i.e.: domain, host and country of origin, they are very poor at extracting even the simplest descriptive information about the page content itself.

As a result of very arbitrary, even random methods of resource selection, Search Engines are unable to support extensive structural association between index data elements and their content, e.g.: resource title, author, publisher, language etc. These Search Engines cannot tell us who created the page, which organisation published it (unless this matches the domain in the URL,) the date of original compilation, or a whole range of details common to even the simplest library catalogue.

Similarly, Search Engines are unable to relate the semantics of page content with user input. As an example of this problem, the user cannot guarantee that a query for 'St. Michael's Library' will provide sites containing the defined string; on the contrary, the search could result in any sites containing words or even elements of words specified, (eg: a hospital called St. Michael's, a church, an individual etc.)

Jennifer A. Younger (1997) has commented on the limited indexing methods of Web Search Engines:

> '…constructed without reference to relationships among documents and little or no control over names or concepts… There is an ever increasing amount of material to index, and all of it just uncontrolled text.'

It is these issues which advanced and emerging resource description models are intended to address, enhancing the bibliographic structure and integrity of the World Wide Web.

Chapter 4.

Metadata: Definition and Overview.

The current technical and fashionable term for bibliographic/ descriptive data on the World Wide Web, is *metadata*. Metadata has its origins in the non-digital environment, in areas such as archival and conservation record formats.

As a definition of resource description standards, *metadata* is universally defined by information scientists and professionals as 'data about data,' (Gill. 1998.)

Jennifer A. Younger (1997) has defined *metadata* as:

> '…documentation about documents and objects. They describe resources, indicate where resources are located, and outline what is required to locate them successfully.'

4.1. Concepts.

Metadata concepts are influenced by the functional aims of a wide range of organisations working in information services and the conservation industry, including libraries, museums and archives.

Anne J. Gillard-Swetland (1998,) has defined the following functions of metadata in a digital context:

- Increased Accessibility.

  To provide more effective resource searching and retrieval standards, across a wider range of indexes and systems.

- Multi-Versioning.

  Recording multiple versions of resources, their relation to similar resources and other attributes, such as location and geographical coverage.

- Legal Issues.

  Documenting resource rights, copyright information and proprietary interests.

- Preservation.

  To preserve existing digital and printed resources in digital form.

Information professionals have also agreed basic components or attributes of metadata; these attributes reflect the necessity for compatibility and interoperability between network and software standards, the need for secure and reliable data, and the relationship between abstract standards and digital structures used to carry metadata content.

Jennifer A. Younger (1997,) has defined essential metadata attributes as follows:

- To comprise data element standards for resource description.
- To comprise data structure and format standards for resource description.
- To support reliable resource locations for resource retrieval.
- To provide interoperability between metadata formats for indexing interpretation, compilation and conversion.

A central problem in defining metadata criteria is cross-format compatibility; this is due to the disparate number, definitions and allowable content of resource description elements for each format. Jennifer A. Younger (1997,) has suggested three aspects of traditional cataloguing methods that could be used in defining metadata elements:

- Copy information.

  Resource characteristics, e.g.: version history, historical value or scarcity.

- Publication / Manifestation.

  Bibliographic features, e.g.: title, publication date, author name, place of origin, dimensions or type of digital resource (See *MIME* in Glossary.)

- The Work.

  Content specific data, e.g.: an abstract or summary, a subject heading or formal classification definition.

4.2. Approaches.

At present several different mainstream approaches to the form and uses of metadata exist on the WWW. All have their advantages and disadvantages, although some stand out above others for precision, extensibility and compatibility with current indexing and retrieval systems.

Metadata models vary regarding their accessibility to the general user; the HTML Meta tag, (used for Dublin Core elements) uses simple HTML syntax accessible to any HTML competent Web author.

Other models, however are more complex, based on several inter-dependent script languages and interpretable by only a few compatible systems; this is seen in RDF (*Resource Description Framework*), a standard protocol for expressing XML (*Expansible Markup Language*.)

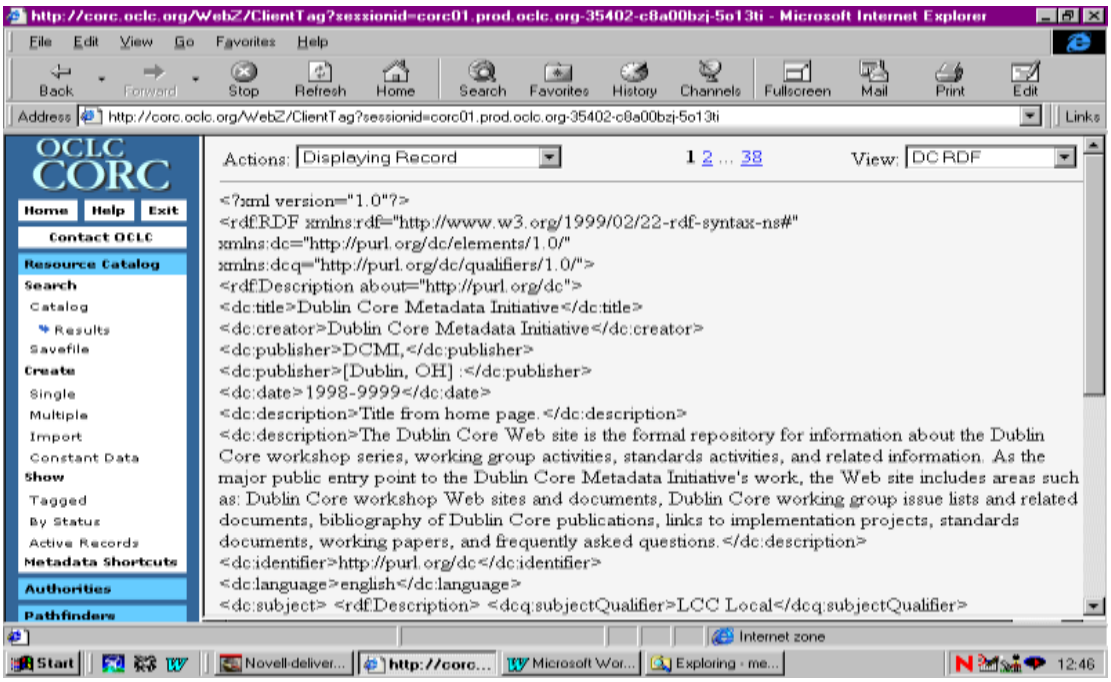


Figure 1. The Dublin Core, expressed using RDF protocol in XML markup script.

RDF requires familiarity with XML conventions, RDF syntax and special RDF features (such as relational resource mapping.) Additionally, there are currently few practical opportunities to use RDF, or XML for the non-specialist, beyond a few online catalogues such as CORC.

Metadata may be associated with digital resources in a number of ways.
One of the simplest and arguably most effective methods is the inclusion of metadata information within the HTML script of Web pages themselves.
This method is supported by prevalent metadata models, including the HTML Meta tag and XML markup (as a subset of HTML.) HTML metadata may be inserted manually, or using a metadata-compliant tool or application (such as Frontpage, which supports the *keywords* and *description* Meta tags.)

```
<META NAME="DC.Title" CONTENT="Tudalen Catref Gwledig ">
<META NAME="DC.Creator" CONTENT="Paul Catherall">
<META NAME="DC.Description" CONTENT="A page devoted to Welsh culture.">
<META NAME="DC.Publisher" CONTENT="The Globe.com">
```

Figure 2. Dublin Core META tags in the HEAD of an HTML file.

At present, some Search Engines and Information Portals do recognise the HTML Meta tag, but are not compliant with advanced uses of this tag, such as the Dublin Core; instead, any advanced metadata of this kind contained in the resource HTML is merely included in free-text indexing processes or ignored.

Another approach to metadata resource-association, is via the use of external data files, usually created during the resource description and classification process, using an online cataloguing system.

XML-based metadata is intended for use as an external file separate from page content; in this capacity, XML may be used to describe HTML documents.

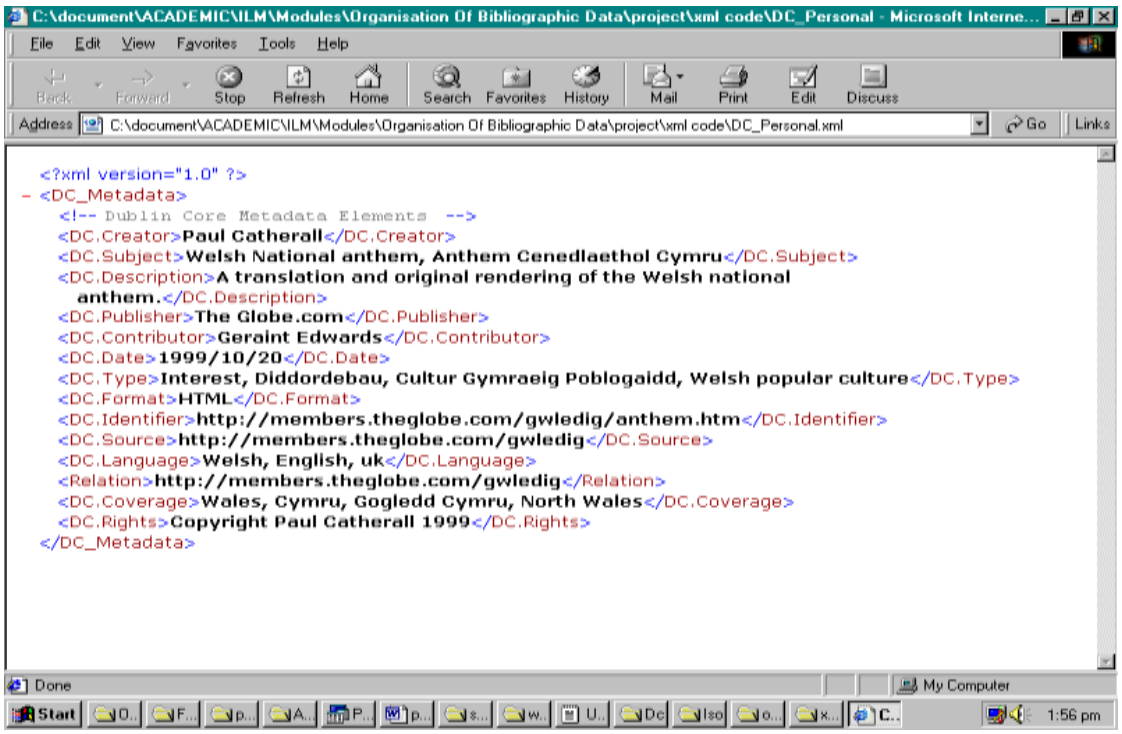


Figure 3. An XML data file, used to contain Dublin Core elements for a Web page.

Finally, metadata creation and management may involve many layers of manual and automatic processes, and many schematic layers within data itself.

As an explanation of this point, consider the RDF (Resource Description Framework) format; this script must be compiled either manually, using a text editor, or via an automated system, using a variety of script format and schematic standards.

This process of metadata layer interdependence is illustrated in the diagram below:

1.	HTML script using *http* protocol provides basis of metadata.
2.	XML script supported by HTML declared using *prolog* tag.
3.	An RDF schema is declared for recognition by an external system.
4.	XML script with RDF schema formatting present in HTML.
5.	RDF schema uses Dublin Core elements.
6.	Online schema-compliant system interprets RDF script.

Figure 4. Metadata Layer Elements for RDF Script.

Chapter 5.

Main Participants In Metadata Development.

This chapter introduces a few of the key organisations responsible for the standardisation and implementation of metadata standards.

5.1. W3C - The World Wide Web Consortium.

http://www.w3.org

The World Wide Web Consortium, founded in 1994, is an international association of industrial and service companies, research laboratories, educational institutions, and information services.

The mission of W3C is 'to lead the evolution of the Web,' (W3C. 2000,) in terms of information standards and resource description models.

W3C has co-ordinated the development of XML, RDF and related standards, and is responsible for the following metadata specifications:

- The XML 1.0 Specifications.
- The XML Schema Support Specifications.
- The RDF Model and Syntax Specifications.

5.2. UKOLN - The UK Office for Library and Information Networking.

http://www.ukoln.ac.uk/

UKOLN is a Bath based organisation, founded in 1977 by the British Library and funded by JISC (The Joint Information Systems Committee.)

UKOLN's mission is to facilitate research and development in online information management systems and standards. UKOLN also contributes to current awareness services for the UK information industry.
The UKOLN organisation has contributed extensively to the development of recent metadata models, including XML and RDF.

5.3. OCLC - The Online Computer Library Centre.

http://www.oclc.org

An international organisation founded in 1967 and based at Ohio State University. OCLC defines itself as, 'a non-profit, membership, library computer service and research organization dedicated to the public purposes of furthering access to the world's information and reducing information costs.' (OCLC. 2000.)

At present OCLC has developed several substantial metadata initiatives including the Dublin Core Element set, the Persistent URL project (PURL,) and the online cross-format cataloguing project, CORC (Cooperative Online Resource Catalogue.)

OCLC members have also participated in the development of metadata formats and W3C standard schemas.

5.4. The IETF (Internet Engineering Task Force,) or Network Working Group.

http://www.ietf.org

The IETF is an informal network of IT and network technology specialists, who work with other official bodies, such as OCLC and the W3C to design and propose networking standards.

The RFC (Request for Comments) service allows individuals or organisations to consult the IETF on a range of networking issues; suggestions by IETF members are published on the IETF Web site for further debate and development by other standard forming bodies.

Many RFC recommendations have become recognised networking standards, including the RFC specifications for the WHOIS++ indexing protocol (RFC1835). Other proposed RFC standards are now approved for use by OCLC and other organisations, these include standard qualifiers for metadata content, such as the RFC MIME (Multimedia Internet Mail Extensions) specifications for data types (RFC2045), and the ISO language and country definition types (RFC1766.)

5.5. The ISO International Standards Organisation.

http://www.iso.ch/

The ISO organisation, founded in 1947, is an international standards body, composed of representatives from principal national standards bodies around the world (such as the British Standards Institute.)

On the ISO homepage, (2000) the mission of this organisation is defined as follows:

> 'To promote the development of standardization and related activities in the world… developing cooperation in the spheres of intellectual, scientific, technological and economic activity.'

ISO's work results in international agreements which are published as International Standards. Significant ISO standards related to metadata include SGML, the Standard Generalized Markup Language (ISO 8879,) which is the parent language for HTML and XML, the language definition types (ISO 639,) and the Country Code List (ISO 3166.) A detailed overview of ISO standards is provided in Appendix F.

5.6. The Getty Institute.

http://www.getty.edu/gri/standard/

The Getty Institute was founded by J. Paul Getty in Malibu, near Los Angeles in 1953. The institute was originally founded as a museum, but has since become a wide ranging organisation, involved in conservation research and digital archiving.

The Getty Standards Programme is an international research project, funded by the Getty Trust in association with many leading conservation bodies, such as the Foundation for Documents of Architecture.

Several widely used models have been developed as a result of the Getty programme, including the following:

- The CDWA categories (Categories for the Description of Works of Art).
- The GDAD specifications (Guide to the Description of Architectural Drawings).
- The Object ID Specifications.
- The Getty Vocabulary Program.

Chapter 6.

Overview of Metadata Formats.

In the context of this chapter, a *format* broadly corresponds to an abstract standard of data classes or elements, and allowable content or qualifiers for describing a resource of some kind within a digital context.

In the case of the Dublin Core, data is intended for use within the HEAD of the resource HTML itself, although the Dublin Core has been implemented as an external database file format by projects such as CORC (8.1).
In other instances, no formal method of data storage is suggested, and the choice of metadata carrier rests with the organisation implementing the format.

6.1. The Dublin Core.

The Dublin Core (or DC) metadata elements grew out of the October 1994 'International WWW conference', at Dublin, Ohio, and were later developed into the metadata standard known as the Dublin Core by OCLC software developers and supporting information professionals from a number of major information networking organisations.

Like the *keywords* META tag, the Dublin Core is intended to facilitate class-specific indexing and retrieval by automated Search Engine technology. The Dublin Core has not been adopted by any prevalent Information Provider, although the format is supported by a few recent online projects developed by the information science community. The Dublin Core model uses conventional HTML META tags, containing the *DC* schema definition, followed by an element definition and element content. The example below illustrates the DC element for a resource 'creator':

<META NAME="DC.Creator" CONTENT="Paul Catherall"

The Dublin Core is intended for use by both information professionals and non HTML programmers, this is reflected in the DC manifesto (OCLC. 2000):

> 'The first target is to provide a generally acceptable, simple resource description format, hospitable to the description of a wide range of sources…'

At the March OCLC conference, 1995, 13 'Core' elements were arrived at for inclusion in the element set, these grew to 15 in subsequent conferences.

The Dublin Core is limited in some respects. As a permissive format, intended for popular use, the Dublin Core is entirely subject to the discretion or imagination of the user in using recommended content.

Element	Element Description.
Title	Name of the resource
Creator	Primary author.
Subject	Keywords and phrases defining the resource subject.
Description	Summary, contents list or abstract describing the resource.
Publisher	Organization responsible for resource publication.
Contributor	A contributor.
Date	The date of resource creation.
Type	Genre or nature of resource content.
Format	MIME multimedia type (Multipurpose Internet Mail Extensions.)
Identifier	A formal code identifier: URI, URL, DOI, ISBN etc.
Source	A resource from which the present resource is derived
Language	Language of content.
Relation	A related resource.
Coverage	Administrative or geographical parameters of resource.
Rights	Intellectual property and copyright declarations.

Figure 1. The Dublin Core Elements.

(Based on Dublin Core element descriptions from the Dublin Core Home Page, 2000.)

It is conceivable that were the Dublin Core adopted by Search Engine systems, this standard would be open to the same kinds of abuse as seen in the existing *keywords* META tag. The Dublin Core site (2000,) admits these limitations:

> 'However, there are limitations in what can be achieved using HTML META tags. It is not possible to group sets of META tags in HTML, nor is it possible to represent any hierarchical structure that may be present in the metadata...'

(Overleaf: an example Dublin Core record.)

```
<HTML>
<HEAD>
<TITLE>Tudalen Catref Gwledig</TITLE>
<META NAME="DC.Title" CONTENT="Tudalen Catref Gwledig ">
<META NAME="DC.Creator" CONTENT="Paul Catherall">
<META NAME="DC.Subject" CONTENT="Wales, Cymru, Welsh, Cymreig,
Interests, Diddordebau, Culture, diwylliannol">
<META NAME="DC.Description" CONTENT="A page devoted to Welsh culture.">
<META NAME="DC.Publisher" CONTENT="The Globe.com">
<META NAME="DC.Contributor" CONTENT="">
<META NAME="DC.Date" CONTENT="1999-10-20">
<META NAME="DC.Type" CONTENT="Interest, Popular Culture, Wales">
<META NAME="DC.Format" CONTENT="HTML / TEXT">
<META NAME="DC.Identifier"
CONTENT="http:\\members.theglobe.com/gwledig">
<META NAME="DC.Source" CONTENT="The Globe.">
<META NAME="DC.Language" CONTENT="English, Welsh, uk">
<META NAME="DC.Relation" CONTENT="http:\\members.theglobe.">
<META NAME="DC.Coverage" CONTENT=" Gogledd Cymru, North Wales">
<META NAME="DC.Rights" CONTENT="Copyright Paul Catherall. 1999">
</HEAD>
```

Figure 2. Example Dublin Core elements; this record was compiled using a Dublin Core editor I recently wrote, allowing the user to input plain text for HTML embedding as META tags (an overview of this program is provided in Appendix C.)

The Dublin Core qualifiers (validated by OCLC in July '00), provide guidelines for the compilation of Dublin Core content. Qualifiers include use of the ISO 339 standard abbreviations for languages, MIME types to define multimedia resources, and use of a sub-field *scheme* to more precisely define element contents; the following example demonstrates the *scheme* to specify a Library of Congress Subject Heading:

<META NAME="Subject" SCHEME="LCSH" CONTENT= "General Works">

However, whilst the qualifiers exist, and are gaining widespread use amongst specialist online catalogues, such as CORC, it unlikely that these qualifiers will be widely adopted by the HTML authoring community as a whole, due to their complexity, and extensive accompanying documentation. A detailed overview of the Dublin Core Qualifiers is provided in Appendix F.

For those inexperienced in using HTML, the Dublin core provides an on-line editor to add DC META tags to any WWW page. This is the 'DC DOT Generator.' The advantage of this application is its GUI interface and ease of use, also it is an entirely free service.

Whilst the Dublin Core is not currently supported by popular Search Engines, this format does provide a highly accessible medium for metadata compilation amongst the majority of HTML authors. The practical research element will assess the extent to which HTML authors are familiar with the META tag, and how comfortable they

are compiling Dublin Core data; the practical research element will also assess how easily users are able to define Dublin Core elements.

Ron Chepesuik (1999,) has suggested, that although the Dublin Core is still primitive, it does represent a comprehensive and accessible format for popular use:

> 'The Dublin core has become the predominant candidate for describing electronic resources… It's not a metadata element set that's going to replace MARC, its going to evolve and coexist alongside it…'

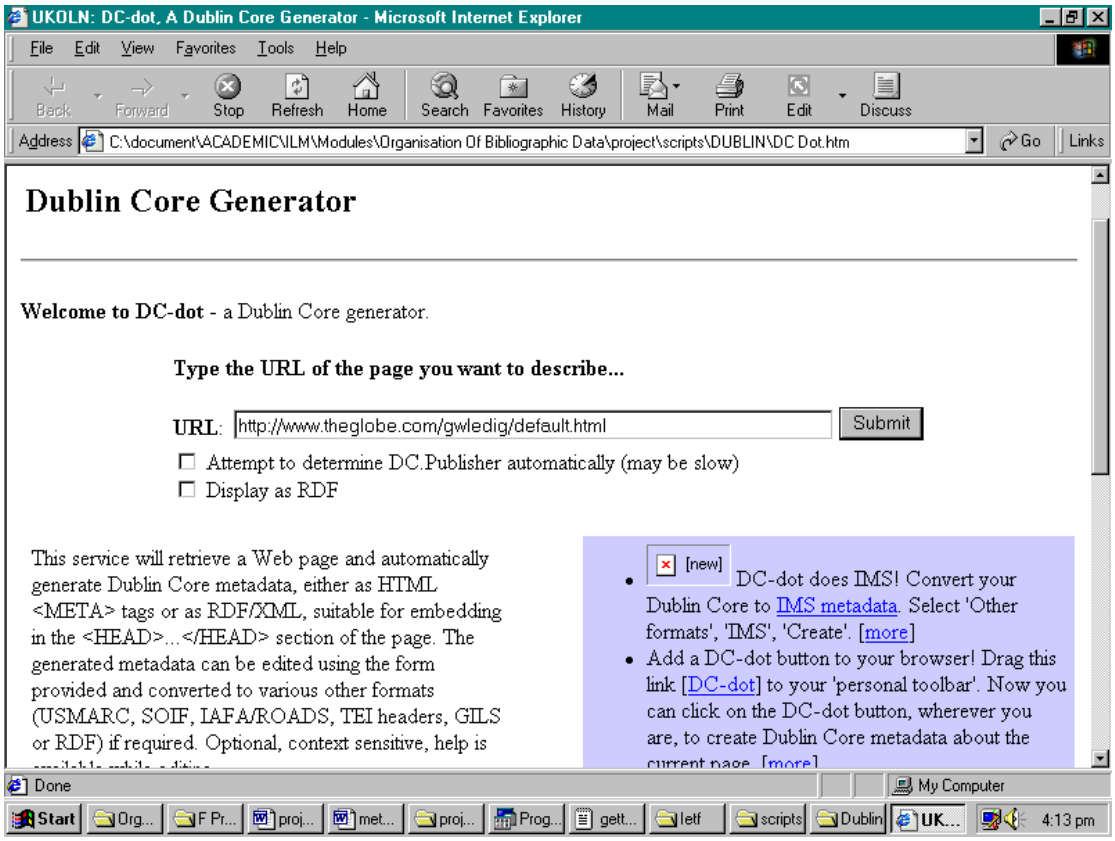


Figure 3. The Dublin Core Generator. (From the Dublin Core Web Page, 2000.)

6.2. The CIMI Schema.

CIMI, the Consortium for the Computer Interchange of Museum Information, is an association of leading cultural heritage institutions and organizations.
The mission of CIMI is to provide standard models for resource description in the conservation industry. The CIMI Schema is intended for use within the heritage industry, and is expressed in SGML (Standard Generalized Markup Language), using an SGML DTD (Document Type Definition) to define allowable content.

The CIMI schema is similar in breadth and scope to the MARC format, containing detailed elements such as *Copyright Restriction* (equivalent to the USMARC 540 field,) *Creator date of birth* and *Creator Nationality*.

6.3. The Getty Standards Programme.

The Getty Standards programme comprises several widely used resource description standards. The categories for the description of works of art (CDWA,) are a Getty Trust standard intended for use within database systems, according to a standard file format. There are eight core categories, including *Materials and Techniques* and *Style/ Period,* with many allowable sub-fields for inclusion within each category.

Formal qualifiers are used throughout CDWA to define element content, e.g.: the *classification* element allows one of nine terms to define the primary media of the object, including *Architecture, Sculpture, Graphic arts*, *Textile,* and *Ceramics*.

The Getty Standards Programme has also implemented the Guidelines for the Description of Architectural Drawings (GDAD).
These comprise a detailed element set intended for specific use in the field of architectural manuscripts; there are 6 core element types, and 19 elements in all.
As in the CDWA specifications, there are controlled qualifiers for the definition of some element contents.

6.4. The CIDOC Guidelines.

CIDOC, the International Committee for Documentation is organised and funded by ICOM, the International Council for Museums.
The CIDOC Guidelines are a description of museum object categories for use in database record structures.

Unlike the CDWA standard, there are presently no suggested qualifiers for elements.
The CIDOC Guidelines include data types specific to works of art and historical items, these include *Acquisition Information, Condition Information, Image Information* and *Subject Depicted Information*.

A comparison table is provided in Appendix A, displaying elements used by metadata formats discussed here, alongside several other widely used standards, including USMARC (a variant of the Machine Readable Catalogue format), and the MESL element set (Museum Educational Licensing Project Data Dictionary) an RLG funded project to develop categories for describing items in the heritage industry.

The comparison table is based on all the official specifications discussed here, and on two metadata comparison tables:

- The Research Libraries Group Metadata Crosswalk, published in Cromwell-Kessler, W. (1998). Crosswalks, Metadata Mapping and Interoperability. *In:* Baca, M. ed. *Introduction to Metadata.* USA, The Getty Information Institute.

- The CDWA Metadata Standards Crosswalk, published in The Getty Standards Programme. (2000). [Online]. Cited 28/06/'00. http://www.getty.edu/gri/standard/fda/index.htm

(Overleaf: an example CDWA record.)

Classification*:	**Drawings**				
Object/Work Type*:	watercolour				
Title or Names*:	Conway Castle, North Wales				
Creation-Creator/Role*:	Joseph Mallord William Turner				
	artist: | Turner, Joseph Mallord William (British painter, 1775-1851)				
Creation-Date*:	1798				
	start: 1798	**end**: 1798			
Subject Matter*:	castle	seascape	Conway Castle (Wales)		
	fishermen	struggle	ocean		
	coast	storm	rocks		
Measurements:	53.6 x 76.7 cm (21 1/8 in. x 30 1/8 in.)				
	height: 53.6 cm	**width:** 76.7 cm			
Materials and Techniques:	Watercolour and gum arabic with graphite underdrawing				
	watercolor	gum arabic	graphite	paper	
Style/Period	Romanticism				
Descriptive Note:	This is the largest of Turner's four extant watercolors of this medieval castle on the northern coast of Wales. Turner portrays the landscape and ocean in a dramatic fashion, using angry clouds, sunshine, and roiling waves to animate the scene and emphasize the struggle of the fishermen...				
Current Location-Repository Name*:	J. Paul Getty Museum				
Current Location-Repository Location	Los Angeles (California, USA)				
Current Location-Repository Numbers	95.GC.10.				

Image Credits: **Conway Castle, North Wales**; Joseph Mallord William Turner; English, 1798; Watercolor and gum arabic with graphite underdrawing; 53.6 x 76.7 cm; 21 1/8 in. x 30 1/8 in.; J. Paul Getty Museum (Los Angeles, CA). 95.GC.10.

Figure 4. An example CDWA record for a Turner watercolour painting.

Chapter 7.

Metadata as Markup.

This chapter explores the relationship between resource description formats and the primary method used on the Internet to carry that content, i.e.: markup languages. A markup language can be thought of as a set of rules that govern the syntax and structure of a digital document.

7.1. SGML and Markup Concepts.

Digital metadata has its origins in the IBM markup language, SGML (Standard Generalized Markup Language), developed in 1986 as a standard formatting language for digital documents; SGML performs the following two functions:

- Content Appearance: i.e.: providing a framework to define how document content appears, using tags to indicate specific aspects of appearance, such as headers, paragraphs and page breaks.

- Content architecture: i.e.: providing a framework to define the nature of document structure and possible content within documents.

Both HTML (*Hypertext Markup Language*,) and XML (*Extensible Markup Language*) are superset forms of SGML, and are the current standards for information markup on the Web today.

HTML, a W3C standard was developed as an accessible and uncomplicated formatting language for the general user community. In selecting only the most fundamental aspects of SGML, i.e.: content display features, HTML lost the capacity to define document structures and handle variable data.

Conversely, the XML markup language (another W3C standard,) is intended to facilitate the definition of data classes and allowable content within the XML document, allowing the XML programmer to create user-defined document structures within a formal DTD (Document Type Definition.)
There are essentially two components to SGML-based documents, the DTD (Document Type Definition,) and the markup document itself; both physically consist of a text file using a standard character set, such as ASCII or DOS.

7.2. The DTD (Document Type Definition).

The DTD essentially defines the laws by which the markup document can function, and may be thought of as a document defining how markup is able to store textual information.

In an SGML-based DTD, there are three essential categories of data: *elements, child elements* and *attributes*. Elements define classes of objects that are being described in

the markup document, while child elements hold data specific to each element. Like child elements, attributes contain data specific to elements; the essential difference between elements and attributes is that elements consist of user-defined content, and should refer to open-ended information, such as names or locations, whilst attributes should contain formal descriptive qualifiers, drawn from a range of allowable terms specified in the DTD itself.

SGML-based languages all have differing levels of dependency on the DTD data, and in some cases, markup documents may be parsed using appropriate software independently of the DTD; the levels of DTD dependency are explained as follows:

- SGML (Standard Generalized Markup Language): The most detailed markup language, with the widest-ranging capabilities. SGML documents must contain a reference to the DTD that their structure adheres to.

- HTML (Hypertext Markup language): This language is used only for displaying content according to the conventions of leading browser parsers, and does not use DTD data.

- XML (Expansible Markup Language): XML does not require a DTD for interpretation by a parser, thus allowing some degree of user-defined structures for inclusion in databases, however, XML may be associated with a DTD, requiring the XML document adhere to that criteria.

The following example displays a DTD specifying allowable elements and attributes for an SGML data file. The markup document adhering to this DTD will contain a list of books, with details on each item in the list.

```
<!--   A Sample DTD for a Bookshop  -->
<!ELEMENT Bookstore (Book+)>
<!ELEMENT Book (Title, Author, Year_Published, ISBN, Price, Review)>
<!ATTLIST Book
        Genre (Fiction | Non-Fiction) "Fiction"
        In_Stock (Yes | No) "Yes">

<!ELEMENT Title (#PCDATA)>
<!ELEMENT Author (#PCDATA)>
<!ELEMENT Year_Published (#PCDATA)>
<!ELEMENT ISBN (#PCDATA)>
<!ELEMENT Price (#PCDATA)>
<!ELEMENT Review (#PCDATA)>
```

Figure 1. A DTD specifying allowable content for a bookshop catalogue.

(This is an example provided with the Microsoft XML Notepad Application, 2000)

The simple conventions and syntax found is this example are common to all SGML based DTDs, including SGML and XML.

The <!-- --!> tag is a standard *remarks* tag used in all SGML-based languages; the first <! ELEMENT > tag specifies the name of the *root element* for inclusion at the beginning and end of the document, in this case, this will be <Bookstore>.

The Second <!ELEMENT > tag defines an allowable *element* that may be included in the markup document, in this case the *element name* 'book' will be used to store details for each book included in the document. The element name is followed by a list, separated by commas specifying the primary data types, or *child elements* for each book, in this case, these include 'Title,' 'Author' and 'Year published.'

The long list of <!ELEMENT> tags at the bottom of the DTD specify how the parser will interpret characters included as content for each child element, in this case, they will be converted into PCDATA, or ASCII standard characters; this is necessary for HTML browsers, since XML content is not automatically converted into ACII text.

The <!Attlist > tag specifies the attributes that complement each element, much in the same way that child elements complement each element. The main difference between an attribute and child element is the capability to specify qualifiers or controlled terms that must be included in attribute content.

The <!Attlist > tag includes the *element name* attributes are associated with, followed by allowable qualifiers that may be used for attribute content. Finally, a default attribute qualifier is defined; this default will be assumed by the parser where attribute content has not been defined in the markup document.

SGML-based documents are most effective when used within a database system, according to the specifications of a formal DTD.

DTD structures, based on SGML and other markup languages, are currently used within online database systems for containing conventional classification and

cataloguing standards, including MARC, the Dublin Core, AACR2 (Anglo American Cataloguing Rules,) and others. Example systems using markup conventions for catalogue functions are described in Chapter 8, *Metadata Initiatives.*

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Bookstore SYSTEM "bookshop.dtd">
<Bookstore>
  <!--J&R Booksellers Database-->
  </Book>
  <Book Genre="Fiction" In_Stock="No">
    <Title>Over The Hills Of Yukon</Title>
    <Author>Bert Colewell</Author>
    <Year_Published>1993</Year_Published>
    <ISBN>5-6524-3054-1</ISBN>
    <Price>$22.00</Price>
    <Review>A Warm Story About A Man And A Moose In Yukon</Review>
  </Book>
  <Book Genre="Fiction" In_Stock="Yes">
    <Title>The Lion's Gold</Title>
    <Author>Daphne Griswald</Author>
    <Year_Published>1989</Year_Published>
    <ISBN>6-7896-2498-2</ISBN>
    <Price>$15.00</Price>
    <Review>One Of The Most Compelling Books Ever Written.</Review>
  </Book>
</Bookstore>
```

Figure 2. An XML document built to the specifications of the bookshop DTD. (Taken from XML Notepad, 2000.)

7.3. XML.

XML (Extensible Markup Language,) was developed by key information science organisations, and standardised by the World Wide Web Consortium (W3C.)
As we have seen in the previous example, XML uses SGML standard features, including *root elements, elements, child elements* and *attributes* to compile data files for describing the metacontent of resources.
In the W3C XML 1.00 (2000) specifications, the aims of XML are as follows:

1. XML shall be straightforwardly usable over the Internet.
2. XML shall support a wide variety of applications.
3. XML shall be compatible with SGML.
4. It shall be easy to write programs which process XML documents.
5. The number of optional features in XML is to be kept to the absolute minimum, ideally zero.
6. XML documents should be human-legible and reasonably clear.
7. The XML design should be prepared quickly.
8. The design of XML shall be formal and concise.
9. XML documents shall be easy to create.
10. Terseness in XML markup is of minimal importance.

Alongside conventional markup functions, XML is envisaged to function as an Application Layer protocol, for data interoperability between applications and database systems, as a communications script, and for use with CGI (Common Gateway Interface) scripting for advanced Web functions.

The main advantage of XML, as its name suggests, is its extensibility, insofar as XML authors may define XML names and allowable attribute qualifiers, and their structural relation to one another within the XML document.

Ken Sall (2000), a writer for the Internet *Web Developer* ezine (2000), describes the benefits of XML extensibility:

> 'XML parsers can infer the structural rules of the language (including ones they have never encountered before) from the context of the elements in the particular document instance…'

The following simple XML script demonstrates the potential to describe two books using user-defined XML tags. The *element* 'book' is used to store details for each record. The book-titles are defined as *attributes,* (*Geirfa Idiomadau* and *Guide to Science;*) associated within each title are two *child elements*, defined as *author* and *publisher*:

```
<?xml version="1.0" ?>
    <PC_Book_Collection>
      <!-- Paul's book collection. --!>
      <Book Title="Geirfa Idiomadau">
        <Author>Rhys, A</Author>
        <Publisher>Lolfa</Publisher>
      </Book>
      <Book Title="Guide To Science">
        <Author>Azimov, Issac</Author>
        <Publisher>Penguin</Publisher>
     </Book>
    </PC_Book_Collection>
```

There are two classes of XML document, *Well Formed* XML and *Valid* XML.
All XML documents must be *Well Formed,* that is they must comply with the basic rules governing interpretation by XML 1.0. standard parsers, such as emerging XML database system software, and the Internet Explorer 5 browser.
These rules include the necessity for SGML hierarchical data, including *root elements, elements* and *attributes* within the XML document.

A Well Formed XML file does not require a DTD, and so may be compiled and used by any individual familiar with basic XML syntax and conventions.
The practical research element in this project aims to demonstrate the potential of Well Formed XML for use amongst the general, non-metadata specialist Web development community. Functions for Well Formed XML could include user-defined resource description documents, or use of a standard format, such as Dublin Core for carrying metadata external to the HTML resource itself.

At present, the Internet Explorer 5 (IE5) browser is capable of parsing Well Formed XML document files with the .xml extension, these may be displayed visually as a hierarchical *data tree*, with expandable and collapsible branches corresponding to the hierarchical structure of XML *root elements, elements, child elements* and *attributes.*
Unlike HTML, Well Formed XML must contain no errors, otherwise the file is invalid and cannot be interpreted by the parser.
It should be noted that an XML file is not intended for displaying resource content, but for containing data describing content, i.e.: metacontent. The visual representation of XML data using IE5 is simply intended to allow users to view XML data associated with either a physical or digital resource, eg: a HTML site.

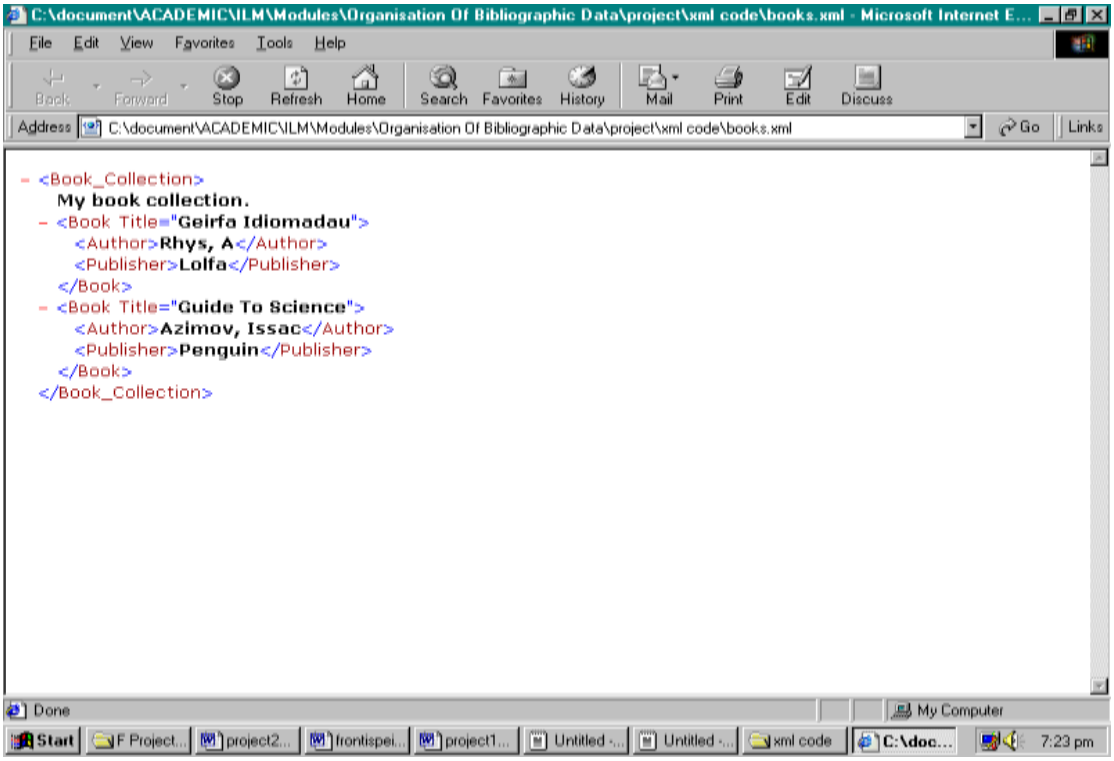


Figure 3. Hierarchical XML script displayed in Internet Explorer 5 Browser. This example describes two books, using the element 'Book-Title,' and two attributes, 'author' and 'publisher.' (The *minus* symbol allows the user to collapse a branch, this then becomes a *plus* symbol for branch expansion.)

Because XML is not DTD dependent, this visual representation of XML using the inbuilt parser and XML style sheet in Internet Explorer 5 provides an ideal way for HTML users to describe Web pages using XML data, either within HTML itself, in HTML frames, or as an external file accessible from the source HTML via a hyperlink.

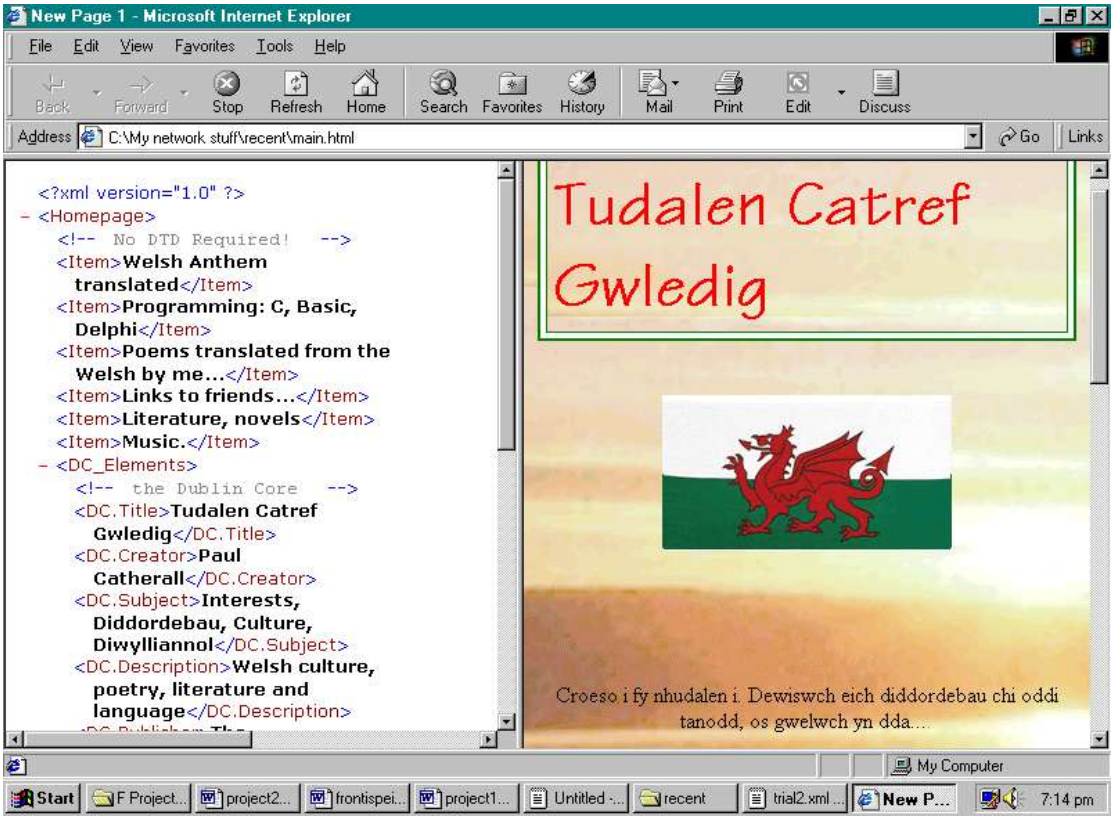


Figure 4. An example XML file displayed within a HTML page frame.

(Taken from *Tudalen Catref Gwledig* [Online], Paul Catherall 2000.)

Unlike Well Formed XML, *Valid* XML refers to an XML file associated with a DTD. The corresponding DTD filename must be specified in the <!DOCTYPE> tag. The following tag instructs the parser to use a DTD called “bookshop.dtd”:

<!DOCTYPE Bookstore SYSTEM " bookshop.dtd">

The DTD specifications are checked against XML document content, and must be compatible in order for successful interpretation by the parser. In more complex database systems, and possibly in an online context, the DTD might be located on a

machine remote form the actual XML document, it is this process which could validate any popularly used XML structure.

The following DTD was designed by the author and placed in the same directory as an XML file compiled to the specifications of the DTD.

The DTD is called ‘Software_List’ and uses the element *software* for software items, and child elements *name, manufacture and price* to contain user-defined content. The attributes associated with each element are *Genre* and *OS* (Operating System,) and qualifiers are specified for each attribute following the *attribute name*; default attributes are also defined:

Genre (i.e.: type of software): **Utility, Application or Game**
The default for Genre is Utility.

OS (i.e.: the operating systems required to run the software):
Windows, Dos or Linux.
The default for OS is Windows.

```
<!--   A Sample DTD -->
<!ELEMENT Software_List (Software+)>
<!ELEMENT Software (Name, Manufacturer, Price )>
<!ATTLIST Software
        Genre (Utility | Application | Game ) "Utility"
        OS (Windows | Dos | Linux) "Windows">

<!ELEMENT Name (#PCDATA)>
<!ELEMENT Manufacturer (#PCDATA)>
<!ELEMENT Price  (#PCDATA)>
```

Figure 5. The Software_List DTD.

The following valid XML file conforms to the specifications in the *Software_List* DTD; two element items are included: *Paul's Password Protection Checker*, and *Microsoft Word*. The <? ?> *prolog* tag is used to declare an XML version for processing in the browser.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Software_List SYSTEM "Software.dtd">
<!-- DTD and This List by Paul Catherall 2000 -->
<Software_List>
<Software Genre="Utility" OS="DOS">
<Name>Paul's Password Protection Checker</Name>
<Manufacturer>Paul Catherall</Manufacturer>
<Price>None (Freeware)</Price>
</Software>
<Software Genre="Application" OS="Windows">
<Name>Microsoft Word</Name>
<Manufacturer>Microsoft</Manufacturer>
<Price>£40</Price>
</Software>
</Software_List>
```

Figure 6. A valid XML file conforming to the Software_List DTD.

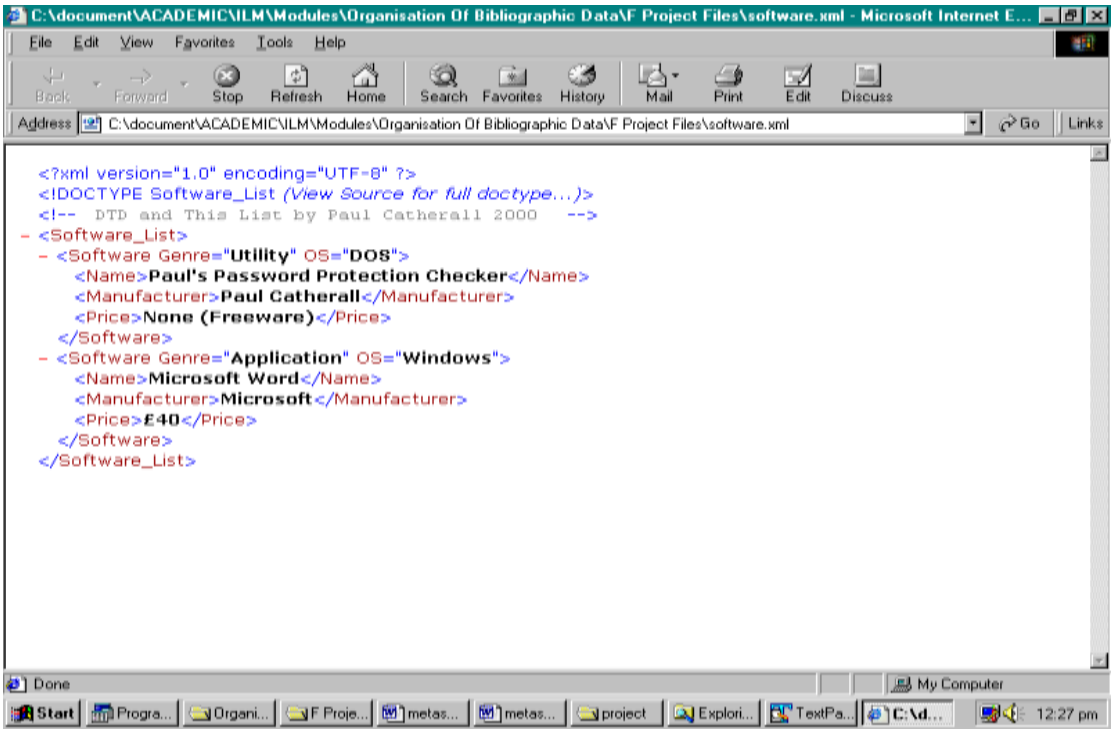


Figure 7. The Software_List XML document, validated by the Internet Explorer 5 XML parser (using the DTD), and displayed using the inbuilt XSL stylesheet.

7.4. XSL (Extensible Style Language) and Related Standards.

XSL (The Extensible Stylesheet Language) was developed from two standards used widely by the HTML authoring community: DSSSL (Document Style Semantics and Specification Language) and CSS (Cascading Style Sheets).

The main difference between XML and XSL, is that where the XML file describes content, the XSL file defines how content should be displayed. A partially implemented feature of Internet Explorer 5, the XSL file will provide a set of markup tags and conventions for allowing the XSL author to define how content is displayed.

The proposed associate standard, XSLT (Extensible Style Language Transformation) is intended as an inbuilt feature of XSL compliant browsers to transform hierarchical content from an XML file into XSL data for display.

Currently, XSL is supported by Internet Explorer 5 using an inbuilt mechanism for displaying XML data as an hierarchical tree within the browser window, but currently, no browsers support user-defined XSL files, and neither is XSLT supported.

The XSL working specifications can be found online at the following address:

http://www.w3.org/TR/WD-xsl/

The XLink (Extensible Link) and XPointer (Extensible Pointer) specifications are also proposed W3C standards for document description and control, and will provide advanced *anchor*, or *hyperlink* features within XML-based languages.

Proposed functions include the capability to merge external hyperlinked documents into the current display (without frames), and to hyperlink to specific parts of a remote document, without arriving at the default location (a useful feature considering the size of some information science specifications.)

The XLink specifications can be found at the following address:

http://www.w3.org/TR/WD-xlink

The XPointer specifications can be found at this address:

http://www.w3.org/TR/WD-xptr

7.5. XHTML (Expansible Hypertext Markup Language.)

XHTML is simply XML expressed within a HTML document. XHTML is supported in Internet Explorer 5 and above, since this browser allows non-HTML XML markup as an integral component of HTML content.

The best application of Well Formed XML within an HTML file, is in defining content elements such as names or phrases using appropriate XML element tags.

The following example demonstrates a few uses of XHTML markup, it should be noted that conventional HTML tags should be avoided for XML tag definitions. XHTML markup within a Web page could potentially be interpreted by a parser; possibilities might include extraction of *valid* XML data for inclusion in an online database, or the structural display of markup content using XSL. XML tags themselves do not appear in the browser with content. (Note, some HTML tags are used here, such as <P> paragraph.)

```
<P>
I have also used free verse in place of the tight rhyming scheme seen in the anthem,
although I have tried to retain the original metre and melody.
</P>
```

Figure 8. A Paragraph in HTML.

(From *Tudalen Catref Gwledig*, [Online] P. Catherall. 2000.)

```
<P>
<!-- XML Markup of personal details --!>
<Anthem_Details>
I have also used <form>free verse</form> in place of the tight rhyming scheme seen
in the <content>anthem</content>, although I have tried to retain the <metre>original
metre</metre> and melody.
</Anthem_Details>
</P>
```

Figure 9. The HTML content with XML markup and Remark Tags; the HTML paragraph has been structured using *root elements* and *attributes* to define tag content.

XML elements could be used to define sections within a document; attributes within each section could provide key descriptive information on content. A plug-in or style-sheet supporting XML within HTML could display this information in table format, providing an overview of lengthy documents, or for keyword searching using an index of tag definitions and content.

(Overleaf: The HTML document with XHTML tags, displayed in Internet Explorer 5.)

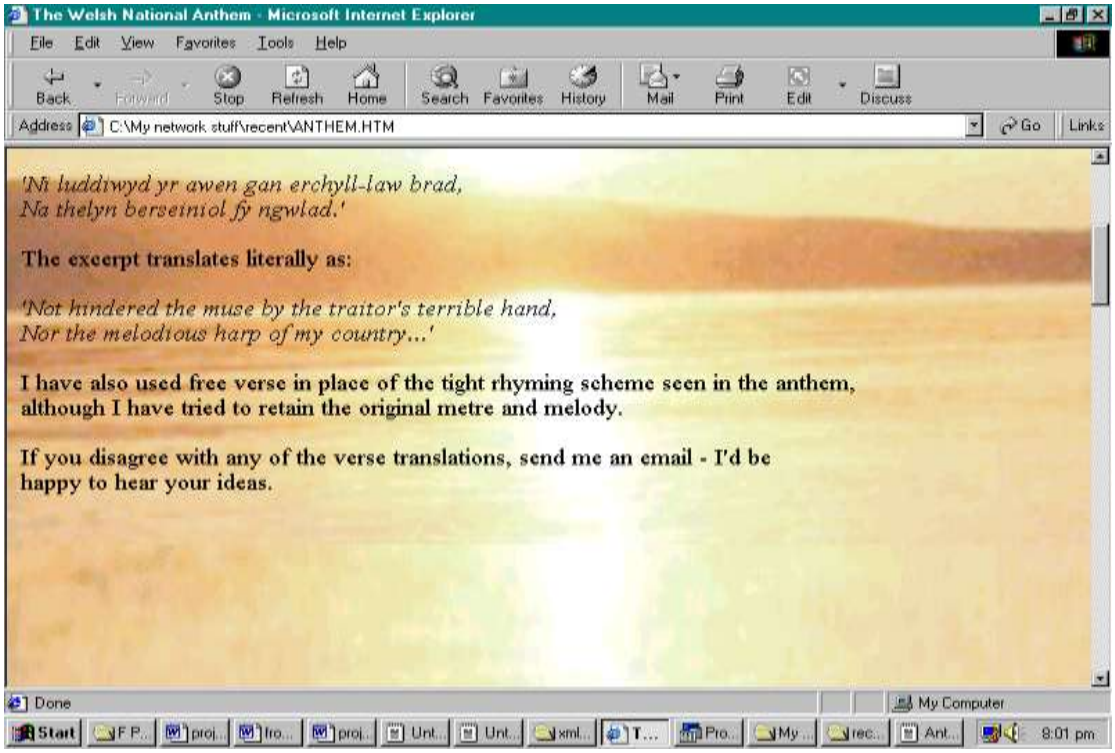


Figure 10. The resulting HTML document with XHTML tags hidden by the browser.

7.6. RDF (The Resource Description Framework.)

The Resource Description Framework was developed by a number of information science organisations, including UKOLN and W3C (the World Wide Web Consortium.)

RDF is essentially a formatting standard for XML; RDF is characterised by *nodes* and *values*, these allow for user-defined tag classes, and content.

RDF is currently used by several cataloguing and indexing systems, such as CORC and the DC. Dot.

On RDF, UKOLN research scientist, Andy Powell (2000) comments:

> 'RDF provides a generic metadata architecture that can be expressed in XML... What can be said is that RDF is likely to become the pervasive metadata architecture, implemented in servers, caches, browsers and other components that make up the Web infrastructure.'

In RDF, *values* (or *atomic values*,) refer to resources on the World Wide Web, and correspond broadly to XML elements. In turn, each *value* may contain *properties*, containing open-ended descriptive data about values; these correspond to XML *child elements*. An RDF hierarchy consisting of a value and associated properties is called a *description*. The properties and values used in a Description are defined and validated by an RDF *schema*, similar in concept to a DTD, and similarly identified by a URI (Uniform Resource Indicator address) which must exist to validate the RDF document.

Like its parent language (XML), RDF is extensible, since the RDF author may define *value* and *property* content; in addition, RDF content must adhere to a defined schema both for validation and parsing. However, unlike XML, RDF contains a wide range of in-built conventions for defining the relationship between resources, whether they are linked by conventional hyperlinks in HTML or via alternative methods, such as CGI, Java applet, Perl or XLink .

The following example demonstrates a schema-defined RDF document; the relationship between the current document and another resource, *resource 2* is indicated in the HREF tag, an element intended for use by RDF-compliant parsers. A *value* tag is also provided to define a Dublin Core 'Creator'.

The *namespace mechanism* seen in the first line, instructs the parser to interrogate a specified online directory of reserved names, that cannot be used for tag definitions. This RDF script is based on an example from the UKOLN RDF site, by OCLC researcher Eric Miller (1998.)

```
<?xml:namespace ns = "http://www.w3.org/RDF/RDF/" prefix="RDF"?>
<RDF:RDF>
  <RDF:Description RDF:HREF ="http://mydomain/document2">
    <DC:Creator>John Smith</DC:Creator>
  </RDF:Description>
</RDF:RDF>
```

The following diagram illustrates the relationship between resources using property tags within an RDF value. (From *An introduction to the Resource Description Framework*, [Online], by E. Miller, 2000):

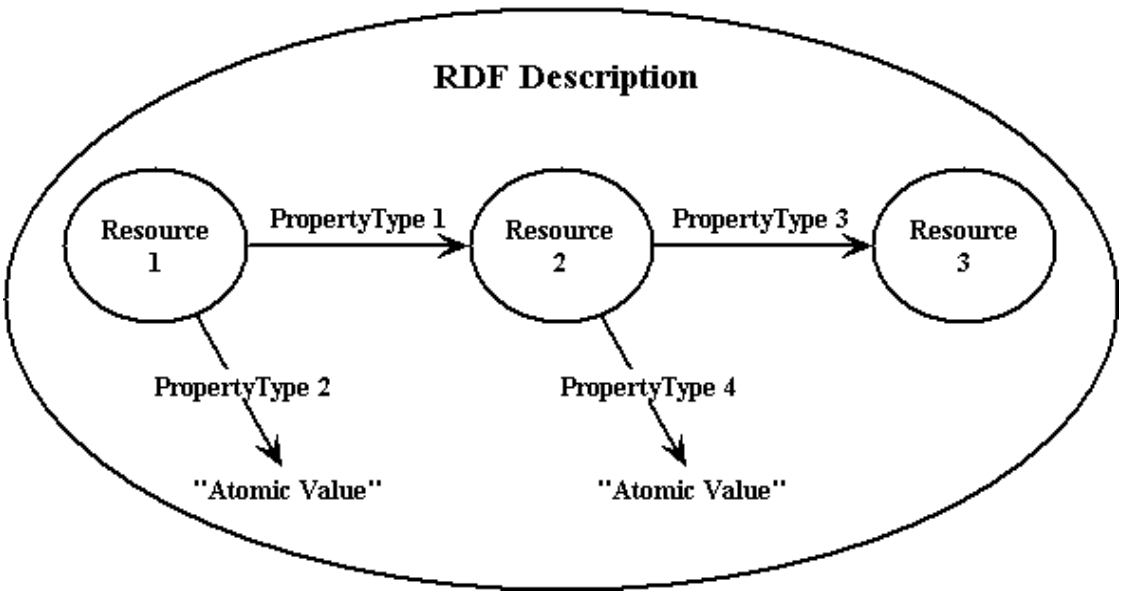


Figure 11. RDF Mapping features, describing the relationship between 3 resources.

The following RDF example from the CORC project (2000,) uses Dublin Core elements. The WC3 RDF *namespace* validation file is defined in the second line (using the xmlns tag), and this is followed by *namespace mechanisms* for both the Dublin Core element set and the official Dublin Core qualifiers. *Namespace* data is used for a variety of validation and indexing functions.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.0/"
xmlns:dcq="http://purl.org/dc/qualifiers/1.0/">
<rdf:Description about="http://prospero.ahds.ac.uk/public/metadata/discovery.html">
<dc:title>Discovering online resources across the humanities : &#183; a practical
implementation of the Dublin Core /</dc:title>
<dc:title>Dublin Core</dc:title>
<dc:contributor>Miller, Paul.</dc:contributor>
<dc:contributor>Greenstein, Daniel.</dc:contributor>
<dc:contributor>Arts and Humanities Data Service.</dc:contributor>
<dc:contributor>Great Britain. &#183; Office for Library and Information
Networking.</dc:contributor>
<dc:publisher>UKOLN,</dc:publisher>
<dc:publisher>[Bath, UK] :</dc:publisher>
<dc:date>1997</dc:date>
<dc:description>95 p. ; &#183; 30 cm.</dc:description>
```

Figure 12. Dublin Core Record in RDF, (from the CORC project, 2000).

Whilst RDF is still in the early stages of development, many compatible online systems and tools are now available to facilitate RDF-based schemas, including the DC. Dot, the CORC project, and the RDF-compatible browser, Mozilla, which can use RDF to create it's hierarchical bookmark display for Web sites containing RDF data.

7.7. The TEI Specifications.

TEI (the Text Encoding Initiative,) is an international RLG (Research Libraries Group) funded project, to develop a markup language for electronic documents possessing both expansible features and the markup functionality of HTML.
TEI is a superset of SGML, and possesses many features found in other SGML derived languages, such as XML and HTML.

One of the key features of TEI is its ability to combine the functions of document display with functions of content description.
In SGML or its subset XML, markup is essentially a means of containing data describing the content of another resource, external to the actual XML or SGML file itself. In TEI markup however, both the document content, i.e.: the document itself, and data describing that content are incorporated into a single markup script.

As in XML, TEI markup is also based on the user-defined DTD (Document Definition) master file, which defines the names and contents of *elements, child elements* and *attributes* allowable in the dependent document.

In addition to user-defined data classes, and element attribute definitions, TEI supports a rich and wide ranging markup syntax to perform standard HTML features such as hyperlinks, frames, style definitions and tables.

The interrelation between markup and resource description is clearly illustrated in the TEI Specifications, (2000.) where the HEADER tag is defined both as a container for plain content and detailed bibliographic information:

> 'The TEI header provides information analogous to that provided by the title page of a printed text. It has up to four parts: a bibliographic description of the machine-readable text, a description of the way it has been encoded, a non-bibliographic description of the text (a text profile), and a revision history.'

Like an HTML document, TEI markup uses start and end tags to define markup syntax and content. The basic format for a TEI document contains a HEADER, a FRONT (Containing bibliographic data), a BODY (Containing content) and a BACK (Containing additional bibliographic data, such as a bibliography.)

```
<TEI.2>
   <teiHeader> [ TEI Header information ]  </teiHeader>
   <text>
      <front>  [ front matter ... ]   </front>
      <body>  [ body of text ... ]   </body>
      <back>   [ back matter ...  ]   </back>
   </text>
</TEI.2>
```

Figure 13. The basic markup conventions for a TEI document.

The TEI Specifications include many tags for describing the bibliographic history and nature of the document, these tags would normally be inserted in the FRONT section of the TEI document, much in the same way that Dublin Core Meta tags are separated from content in the HTML HEAD section.

<date>
contains a date in any format, with normalized value in the value attribute.

<docAuthor>
contains the name of the author of the document, as given on the title page (often but not always contained in a <byline>).

<docDate>
contains the date of the document, as given (usually) on the title page.

<docEdition>
contains an edition statement as presented on a title page of a document.

<docImprint>
contains the imprint statement (place and date of publication, publisher name), as given (usually) at the foot of a title page.

Figure 14. Bibliographic TEI tags. (From the TEI Specifications page 2000.)

As we have mentioned, TEI markup provides methods of describing content using referable tags; these tags define forms of content such as the literary structure of poetry or drama; in the case of a novel, this might allow us to display all the instances of character speech, or the speech of a particular character.

In the case of a technical manual, they could list element content containing selected terminology or subject terms.

The following tags identify elements of the novel, and could be used in a searching or indexing process:

<lg>
contains a group of verse lines functioning as a formal unit e.g. a stanza, refrain, verse paragraph, etc.

<sp>
contains an individual speech in a performance text, or a passage presented as such in a prose or verse text. Attributes include:
who
identifies the speaker of the part by supplying an ID.

<speaker>
contains a special form of heading or label, giving the name of one or more speakers in a performance text or fragment.

<stage>
contains any kind of stage direction within a performance text or fragment. Attributes include:
type
indicates the kind of stage direction. Suggested values include entrance, exit, setting, delivery, etc.

Figure 15. Tags describing text elements. (From the TEI Specifications Page 2000.)

What follows is a poem excerpt in which verse lines and stanzas are defined using a combination of formatting tags for content display, and meta tags for content-type definition:

```
<lg n=I>
<l>I Sing the progresse of a
  deathlesse soule,</l>
<l>Whom Fate, with God made,
 but doth not controule,</l>
<l>Plac'd in most shapes; all times
 before the law…
```

Figure 16. TEI poem excerpt (John Donne: *The Progresse of the Soule*), with verse-specific tags defining content as stanza elements. (Taken from the TEI Specifications Page, 2000.)

Whilst TEI is supported only by compliant database systems in a few locations around the world, such as the Oxford text digitisation programme, it does have many advantages over similar markup languages, such as HTML and the more recent XML. The only problem in implementing the combined markup features of TEI, as a medium for content description and presentation, is the necessity to provide an equally sophisticated parser for interpreting the TEI document, and the user interface to access and manipulate the features present in TEI markup.

The advantages to the entire HTML authoring community would be significant if markup languages like HTML were able to provide the kind of combined display/ description functionality available in TEI.

Such an application of TEI functions might be envisaged in more advanced and comprehensive versions of HTML, where XML markup might support formal tag definitions for interpretation via compatible browsers.

Chapter 8.

Metadata Initiatives.

This chapter explores two initiatives to develop resource catalogues using a variety of emerging metadata standards; often, new formats, such as XML are supported alongside well-established formats, such as USMARC, allowing data exchange and conversion between formats.

8.1. The CORC Project.

The Cooperative Online Resource Catalogue (or CORC,) is an online public catalogue developed by OCLC in partnership with several hundred volunteer libraries.
The CORC catalogue supports a variety of record formats, including RDF, MARC and the Dublin Core.

Web resource records may be entered via an intuitive GUI interface, in either USMARC or Dublin Core formats, or entered in these formats using a plain text window.

The following interface types are provided for data input and viewing:

Record View.	Description.
MARC	MARC 21 (Machine Readable Cataloguing Format,) displayed in text boxes.
MARC Text Area.	MARC 21 (Machine Readable Cataloguing Format,) displayed in a single text window.
Dublin Core 2.	Dublin Core Elements provided in text boxes using plain language.
DC Text Area.	Dublin Core Elements displayed in a single text window using plain language.
DC HTML.	Dublin Core Elements displayed in a single text window with HTML tags.
DC RDF.	RDF (resource Description Framework format,) displayed in a single text window, for view/ export only.

RDF script may not be entered initially, but conversion from DC or MARC formats to RDF is provided for resource viewing or export.

The CORC project provides a number of standard library management system functions, including an *export* function to download record files, and an *import* function to upload MARC 21 compliant resource data into the server index. The following screenshot illustrates the GUI MARC view:

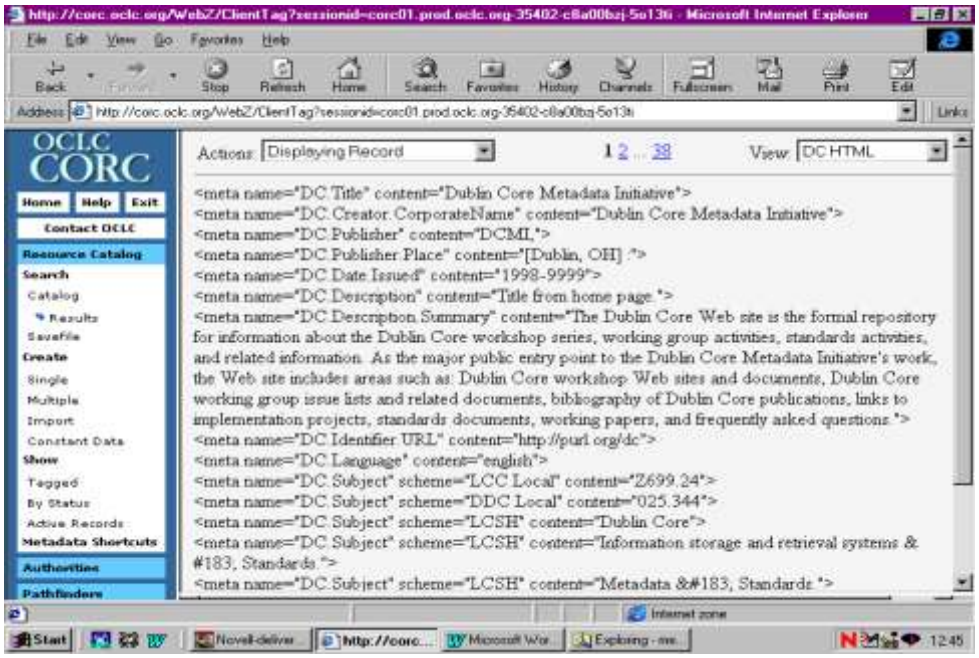


Figure 1. MARC view in CORC. (From the CORC Page, 2000.)

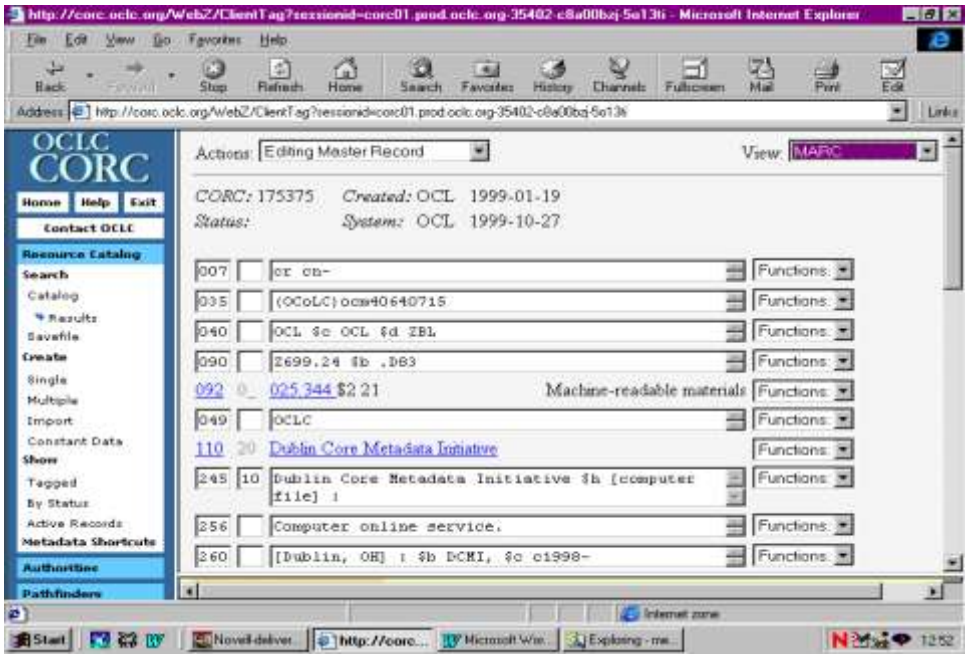


Figure 2. HTML Dublin Core view in CORC. (From the CORC Page, 2000.)

8.2. The Medlane XML Catalogue.

The Medlane XML project, entitled, 'Medlane Experiment - MARC to XML,' is an experiment based at the Chicago Medical Association library to build a library management system capable of converting USMARC records into XML data files, and able to use both formats for standard catalogue functions.

The system is based on a proprietary server using the experimental XML-MARC software; this software reads USMARC records and converts them to XML format according to a DTD. The result of this process is that an XML file is produced to mirror the contents of the original MARC record.

Marc *tags*, *sub-field codes* and *indicators* are converted to XML *elements* and *attributes*. *Element names* are defined according to Marc fields and sub-field codes, e.g. <v035> for the USMARC field 035.

```
<!-- Revision 11/20/99 by Dick Miller and Prisdha Dharma -->
<!ELEMENT work (numcode | pn | cn | cf | to | kw | descr | title |
        relation | location | fg | serprice | staff | v939)*>
<!ATTLIST work type (mono | serial | analytic | component |
        collection | subunit | error) #REQUIRED>
<!ATTLIST work origin (ETC | AUT | LML | SUL) #REQUIRED>
<!ATTLIST work f001 CDATA #REQUIRED>
<!ATTLIST work f005 CDATA #REQUIRED>
<!ATTLIST work f008dc CDATA #REQUIRED>
<!ATTLIST work f009 CDATA #IMPLIED>
<!ATTLIST work f000-21 (Y | N) "N">
<!ATTLIST work status (increased | corrected | deleted | new | prepub | error)
```

Figure 3. The Medlane DTD for a MARC Record, defining initial MARC tags.

The Medlane experiment is located at: http://xmlmarc.stanford.edu/

Cross-domain catalogues could potentially share records using XML, creating a combined database using a common DTD to control record conventions and controlled content.

XML data could provide more flexibility for joint-database architecture, addressing the current lack of standardisation between record formats across the information systems industry; current library systems such as HERITAGE and MILLENIUM use very different record formats, based on differing bibliographic standards.
XML and DTD structures could therefore provide an effective standard for the encoding of records within cross-domain library management system catalogues.

Chapter 9.

Metadata Tools and Client Software.

This chapter provides an overview of a few of the computer programs available for compiling metadata according to markup specifications, or creating DTDs to specify metadata content within a database or browser scenario.

9.1. Internet Explorer 5.

As we have noted earlier, the *Internet Explorer 5* browser contains an inbuilt XSL style sheet for displaying XML data, and an inbuilt parser for validating well-formed XML; in addition, validation against a DTD is possible.

Internet Explorer 5 can be downloaded from the following address:

http://www.microsoft.com/ie5

(Overleaf: XML data displayed in Internet Explorer 5.)

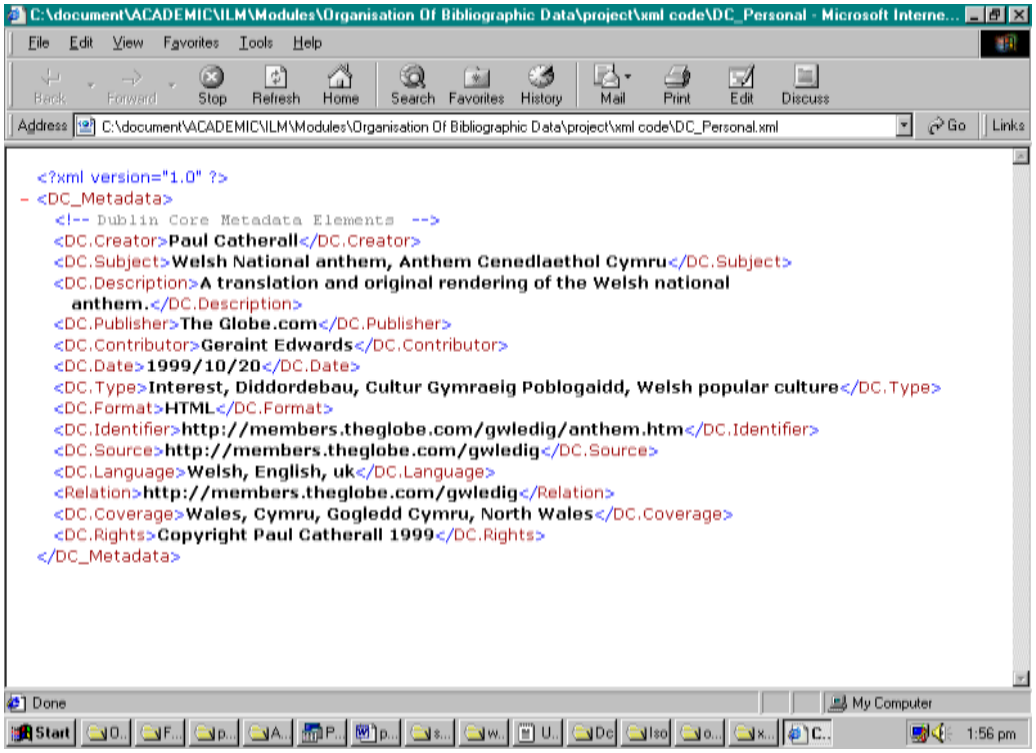


Figure 1. XML data containing Dublin Core elements displayed in Internet Explorer 5.

9.2. The Mozilla Browser.

Mozilla Seamonkey, developed by the Mozilla corporation, is a client browser based on the source code for Netscape's *Netscape Navigator*.

This browser is able to parse and validate XML data for well-formedness and DTD compatibility, but cannot display XML as a hierarchical structure, as seen in Internet Explorer 5. Mozilla can, however interpret RDF data to compile the *history* and *bookmarks* section of its interface, displaying relative links from the current page, as specified by RDF script contained in the resource.

Mozilla is not available commercially, although the source code is available for download and compilation using a Java compiler; source code compilation is not recommended for those without Java experience; the Mozilla source code is available at the following address:

http://www.mozilla.org/downloads

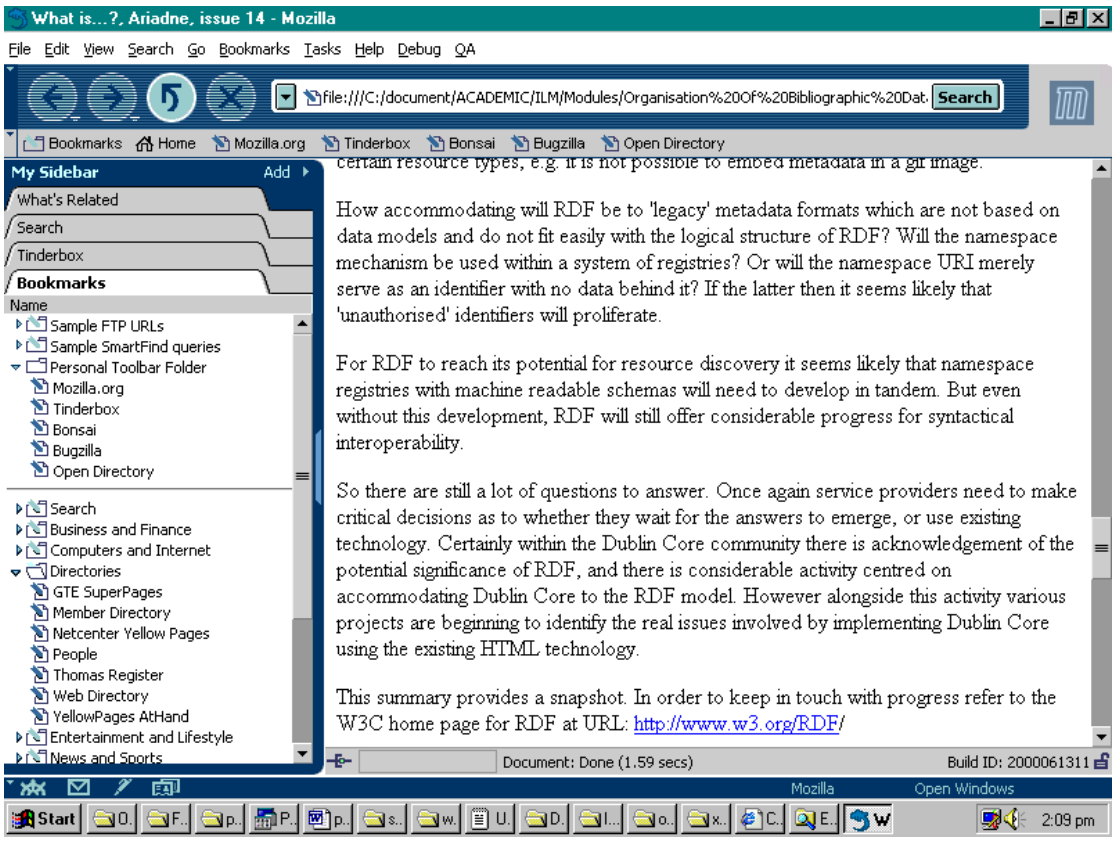


Figure 2. The Mozilla browser, illustrating the use of RDF-based Bookmarks.

9.3. The Near and Far Designer.

The Near and Far designer, by Microstar, is an application for designing DTDs and authoring XML data. Powerful authoring tools are provided, including multiple views of XML data as a text or tree-based structure. XML components, including elements, attributes and other structural components can be inserted into an XML document using macro-style functions, eliminating the need for lengthy typing.

Near and Far Designer is an advanced application, and is unsuitable for most novice XML authors. The trial version, allowing only one inbuilt DTD file to be edited can be downloaded from Microstar at the following address:

http://www.microstar.com

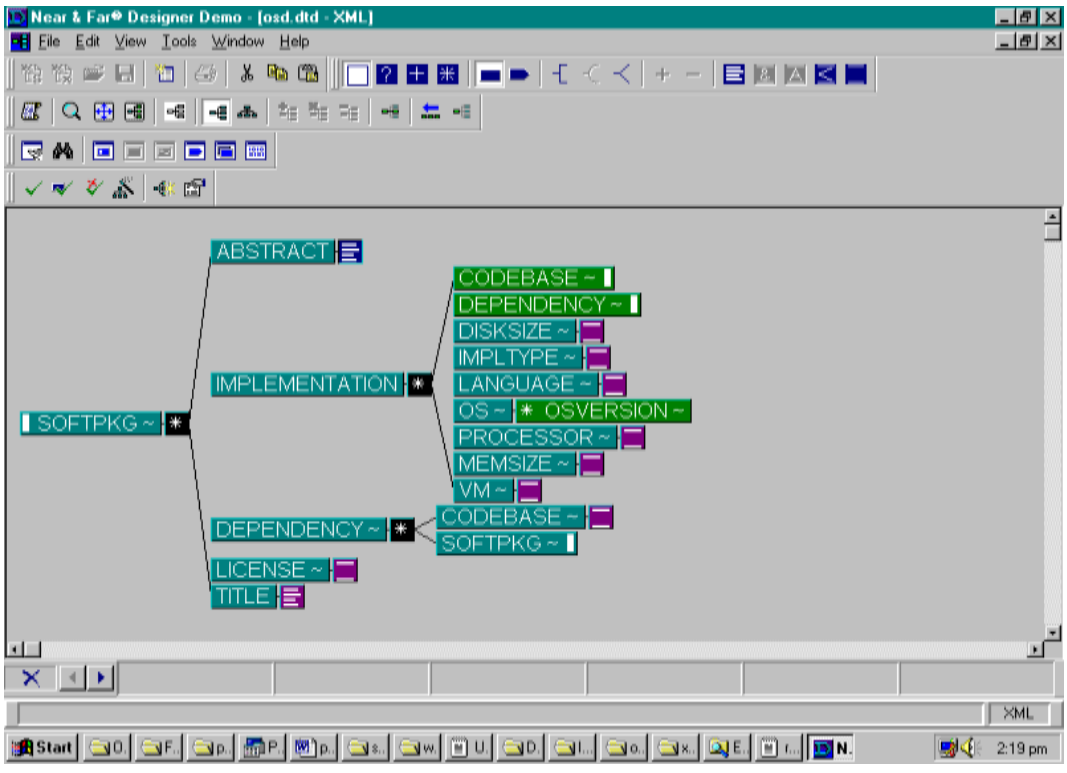


Figure 3. The Near and Far designer.

9.4. XML Notepad.

XML Notepad, by Microsoft is a simple XML document editor, and allows the user to define *root elements, elements, child elements, attributes, comments* and content by inserting XML tags as plain text within a GUI style interface.

Icons are used to differentiate between classes of XML data; In the following example, an exclamation mark, corresponding to <!-- --!> is used to define a comment, while a directory icon (like a little briefcase) defines an element. The red bar represents an attribute.

Files can be saved as XML data, although DTDs cannot be created using this application. The resulting XML file will be validated for well-formedness according to the XML 1.0 specifications.

XML Notepad is an ideal application for the novice XML author, and may be downloaded at the following address:

http://msdn.microsoft.com/xml/notepad

(Overleaf: XML Notepad.)

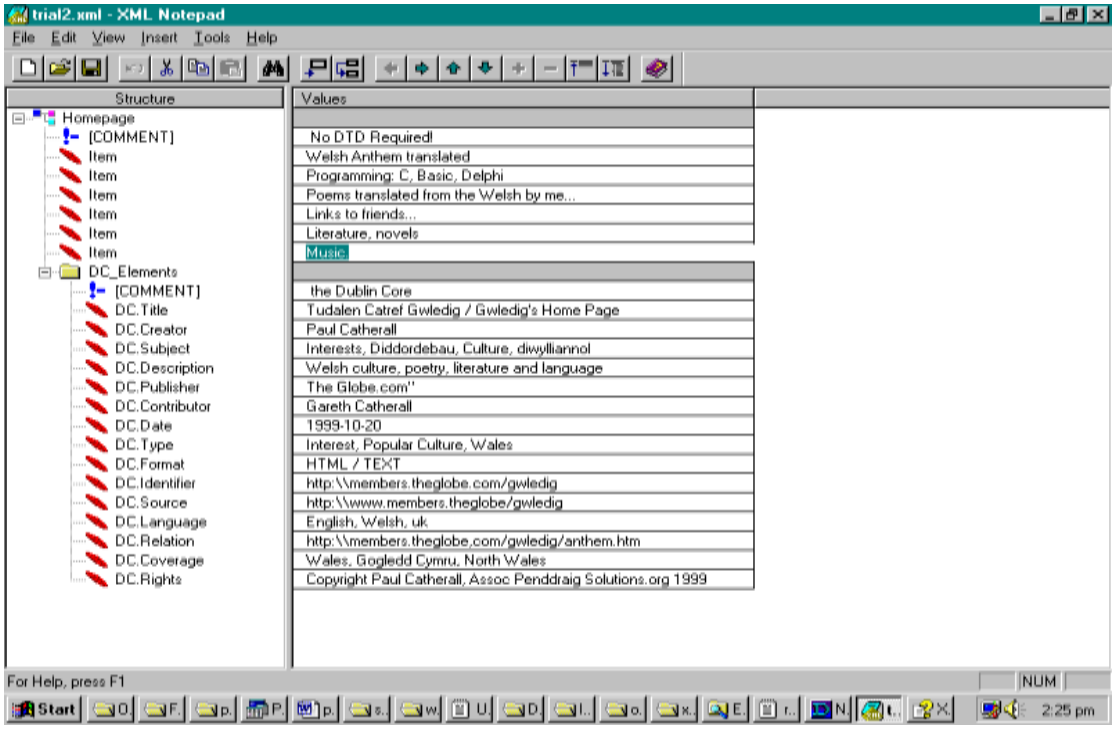


Figure 4. A Well-Formed XML data file loaded in XML Notepad.

9.5. Example HTML Metadata Program.

In the early stages of planning this project, I wrote a small utility that allows the user to embed Dublin Core elements into the HEAD of their Web page, simply by entering plain text for each element when prompted. The most recent version of the program requires a formal RFC MIME standard for the TYPE element, illustrating the potential for a simple metadata tool based on standard qualifiers and controlled vocabularies.

Subsequent versions could require the user select an official Dublin Core *scheme*, and could include any number of qualifier types to ensure standard content is include by the user.

No knowledge of HTML is required to use this application, and it works with any ASCII standard text file (.txt, .htm etc.)

A full description of this program is included in Appendix C.

This program may be requested for email transfer from the following email address:

e9501788@newi.ac.uk

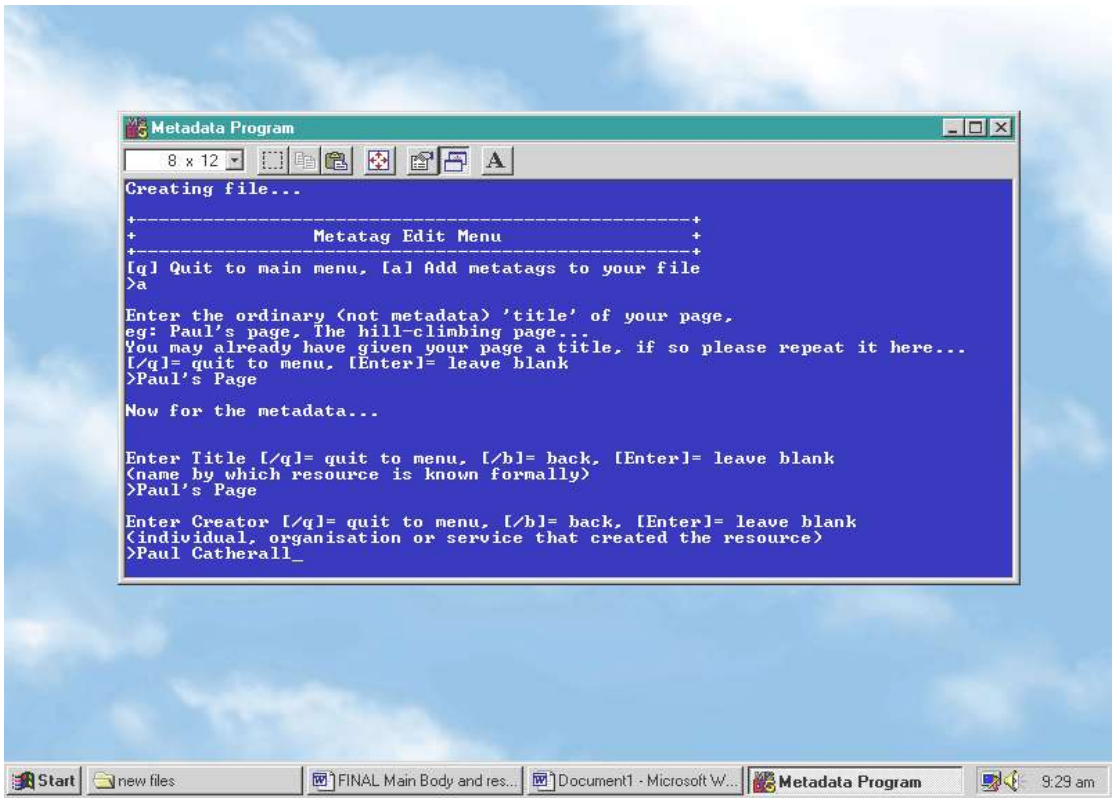


Figure 5. Example HTML Metadata Program.

Chapter 10.

Questionnaire Results and Analysis.

This chapter aims to interpret the primary statistics from the practical research element. Key objectives for research, defined in the project objectives, and in the methodology (2.8) will form the basis of analysis.

The five key areas for investigation in the practical investigation included the following:

- Awareness of metadata amongst HTML users,
- Perceptions of metadata amongst HTML users,
- Metadata accessibility for HTML users,
- Metadata interoperability with networking standards,
- Script protocol transparency with prevalent markup languages.

Note: Standard deviation is used here to indicate dispersion from the mean average; this illustrates the degree of variance across a series of results. In this section, standard deviation will be used to analyse scale-based results.

10.1. Awareness of metadata standards amongst HTML authors.

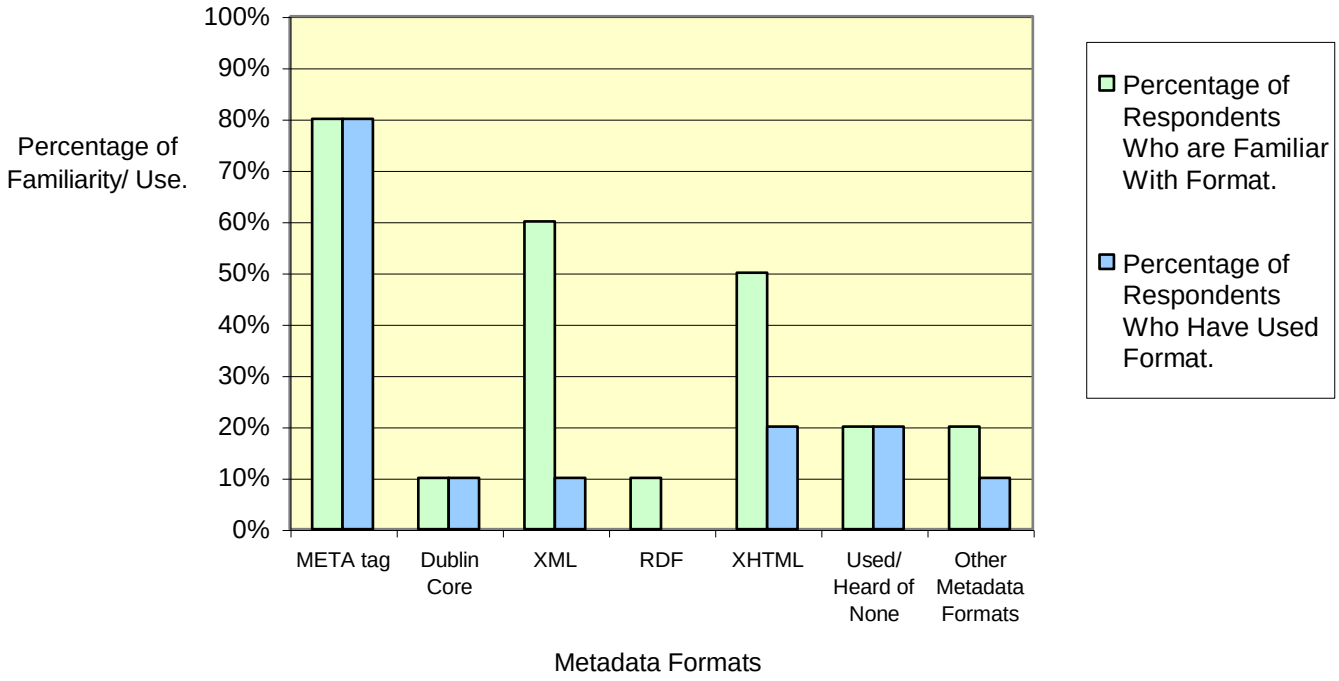
One key area for investigation in the practical research, was an assessment of how aware users were of metadata technology. Metadata awareness was assessed in two multiple choice questions: question 3, which asked the respondent which metadata formats they has heard of, and question 5, which asked which formats they had used. Two distinct forms of metadata were included in these multiple choice questions: well known markup structures (such as the META tag), and more recent advanced structures (like RDF and the Dublin Core.)

Overall, 80% of respondents were familiar with at least one form of metadata, with only 20% who had not heard of any format.

The results demonstrated significant awareness of basic metadata structures, such as the META tag (80%), and XHTML markup (60%), but less awareness of advanced structures, such as RDF (10%).
Similarly, a high proportion had used at least one metadata standard (80% overall), and a large proportion had used the META tag (80%). However, few had used advanced structures such as RDF (10%) or the Dublin Core (10%).

(Overleaf: chart comparing respondent awareness and use of formats.)

Figure 1. Comparison Between Respondent Awareness and Use of Formats.



There were varying levels of correlation between awareness and use of specific metadata structures. Levels of awareness and use were consistent for the META tag, suggesting significant awareness of this format, and its accessibility for HTML authors. However, there were lower levels of correlation between XML awareness and use (50% variance) and XHTML awareness and use (30% variance), indicating that whilst these structures were well known, only a few had used them. Additionally, there were low levels of awareness and use for advanced resource description formats (eg: the Dublin Core and RDF), suggesting that few HTML authors have either heard of or have used these formats.

These results therefore indicate high awareness of the META tag, and markup languages designed as carriers of metacontent, (eg: XHTML), and high use of the META tag.

However, the results also indicated low use of advanced metadata structures and formats for expressing metadata, such as XML, the Dublin Core and RDF.

10.2. Perceptions of metadata standards amongst HTML authors.

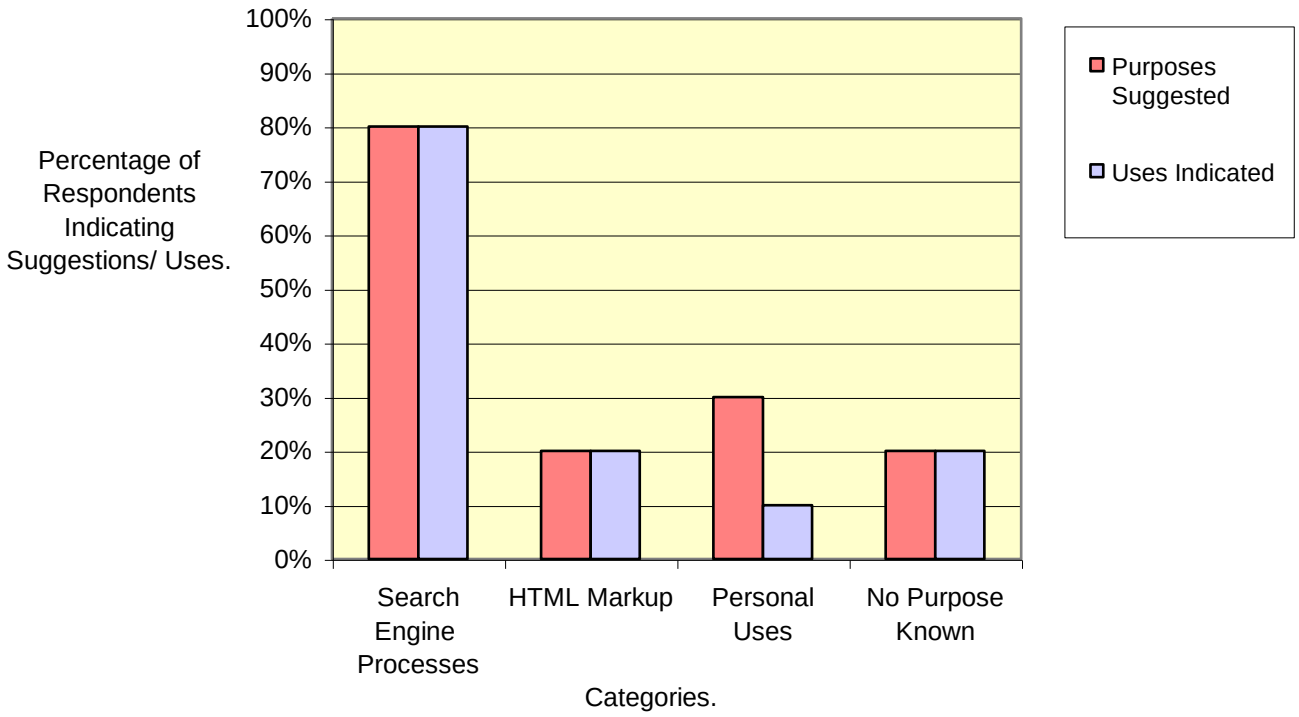
The practical research also attempted to assess how well respondents understood the technology surrounding metadata, and the purpose of metadata compilation.

Two open-ended questions were used to investigate these issues: question 4, which asked respondents what they considered the purpose of metadata to be, and question 6, which asked what reasons they had for creating metadata.
Responses from these questions were categorised, allowing frequency calculations for distinct comment categories.

Overall, 80% had some perceptions on the purpose of metadata, with only 20% who had no perceptions on metadata purpose; a large proportion (80%), thought metadata was used for Search Engine indexing or ranking using the META tag, whilst 30% thought metadata was used for personal reasons, such as describing their Web page using the META tag for source code viewing.
Similarly, 80% had used metadata for Search Engine processes, with fewer instances for other purposes, such as internal markup (20%) and personal reasons (10%).

(Overleaf: chart comparing suggested purposes and reasons for creating metadata.)

Figure 2. Comparison Between Suggested Purposes and Uses of Metadata.



Results indicated significant awareness of metadata functions in Search Engine processes, and personal uses of metadata for search engine indexing (80%). There was an exact correlation between use of XHTML markup, and perceptions of metadata as a method for structuring documents using markup tags (20%), indicating low awareness and use of metadata for this function.

Additionally, 30% suggested that metadata was useful for personal Web authoring functions, such as structural tags to define HTML elements, or HTML document description using META tags.

These results indicate that Search Engine indexing processes are well known, and that this is the primary reason for metadata compilation, although XML-based metadata is also considered a useful tool for a variety of markup functions

10.3. Metadata compilation accessibility for HTML authors.

One of the most essential criteria for metadata structures is their accessibility to users at the compilation level. The practical research element attempted to assess how easily users were able to use metadata formats. Respondents were asked to create brief metadata records for a Web page in two practical experiments, one for well-formed (user-defined) XML and another for the Dublin Core.

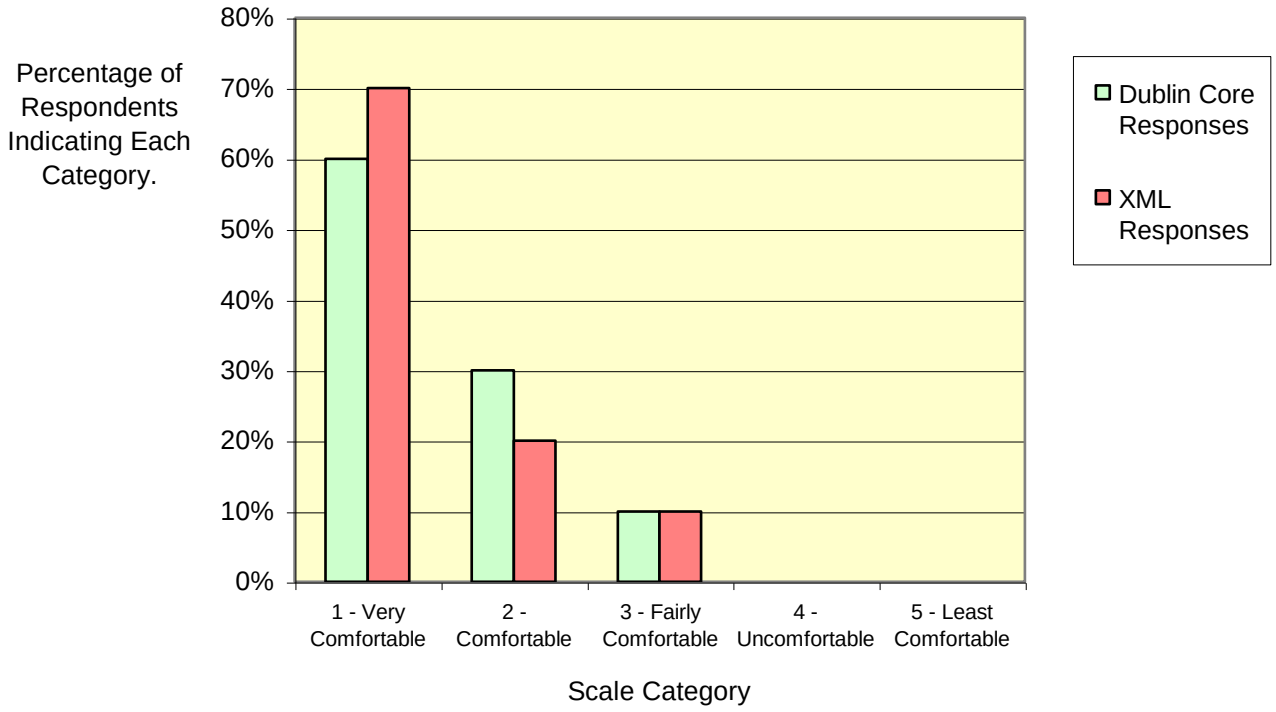
A series of scale-based questions were used to assess format accessibility, where 1 represented the highest score, and 5 represented the lowest, (e.g.: 1 - Very Comfortable, to 5 - Least Comfortable.)

The first series of questions related to format accessibility included questions 7 and 15, which asked 'how comfortable' respondents generally were using the Dublin Core and XML respectively.

A high proportion of respondents selected 1 (Very Comfortable) in both questions, with 60% selecting 1 for the Dublin Core, and 70% selecting 1 for XML.
Overall, 90% of respondents selected either 1 or 2 for both questions, with the remaining 10% selecting 3 (Fairly Comfortable) for both formats.

(Overleaf: chart comparing how comfortable respondents generally were compiling Dublin Core and XML data.)

Figure 3. Comparison Between How Comfortable Respondents Generally Were Compiling DC and XML Data.



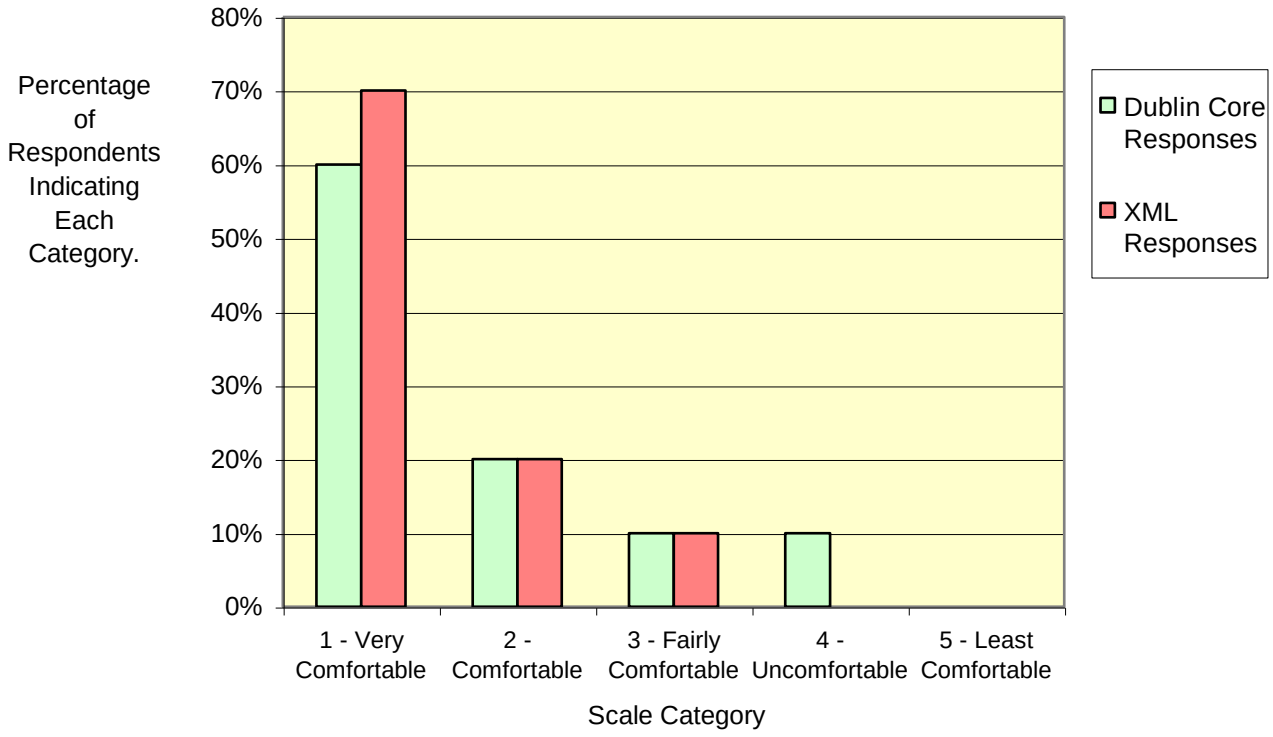
These results demonstrate a high and consistent sense of ease using both formats; this is confirmed by mean averages for both questions, where the average for the Dublin Core was 1.5, and the average for XML was 1.4.

However, averages also indicate that respondents felt more comfortable using XML than they did using the Dublin Core; this fact is confirmed by a lower standard deviation for XML (0.663325) than the Dublin Core (0.670820393), indicating a greater central tendency to option 1 for the XML experiment.

These results therefore indicate that a high number of respondents were generally comfortable using both formats, although respondents were slightly more comfortable using XML than the Dublin Core.

Format accessibility was also assessed in questions 8 and 16, which asked respondents how comfortable they were 'using the conventions and syntax' of each format. Again, there were high frequencies for option 1 in these questions, with 80% selecting 1 (Very Comfortable) for the Dublin Core, and 70% selecting 1 for the XML question. The remainder for each question selected 2 (Comfortable) or 3 (Fairly Comfortable.)

Figure 4. Comparison Between How Comfortable Respondents Generally Were Using DC and XML Conventions and Syntax.



As in the previous question, the majority of respondents were very comfortable using the syntax and conventions of each format; this was confirmed by a high mean average for the XML question (1.4), and the Dublin Core question (1.7).

However, the averages and standard deviation for each range of scores indicate that respondents were slightly more comfortable using XML conventions and syntax than those of the Dublin Core. The standard deviation for the XML question was 0.663324, lower than the Dublin Core result 1.004988; this indicates a greater central tendency to a score of 1 in the XML question.

These results therefore indicate a high and consistent sense of ease using the conventions and syntax of both formats. However, slightly more respondents felt more comfortable using the conventions and syntax of XML than those of the Dublin Core.

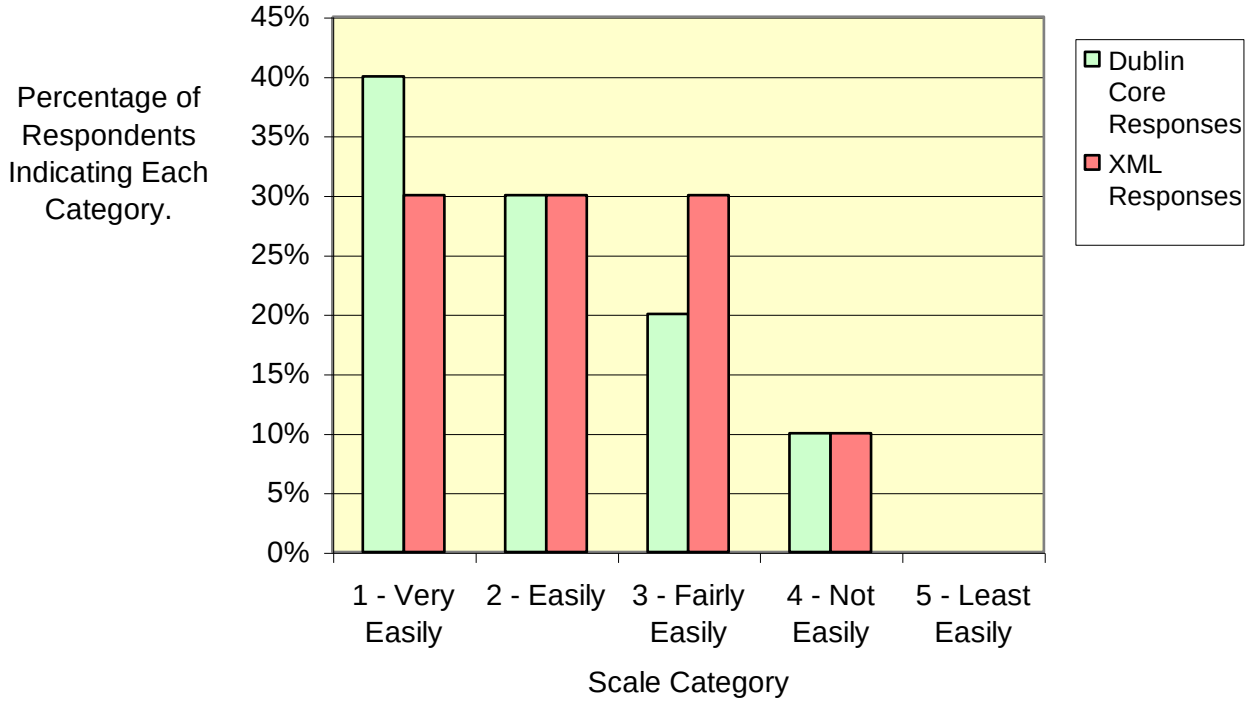
Another aspect for investigation in the practical research was how easily users were able to decide key terms, phrases and free text for inclusion in metadata content. This investigation reflected the current problem of open-ended definitions for META tag elements, where users ordinarily choose terms or phrases themselves, rather than use a standard vocabulary or subject headings.
Currently, there are no standard terms for the conventional *keywords* or *description* META tag, although the Dublin Core has introduced a range of standard qualifiers (see Appendix F.) Consequently, the practical research element attempts to discover how comfortable respondents were deciding key terms without a standard vocabulary structure.

The first content-based questions were questions 11 and 19, which asked how easily respondents were ‘able to decide the kind of content’ (i.e.: the nature of the content) to include in Dublin Core *keywords* and XML elements respectively

There was a balance of scores across each question, indicating that a significant proportion found it difficult to decide types of content for both formats.

In the Dublin Core Question, only 40% selected 1 (Very Easily), with 30% selecting 3 (Fairly Easily). Similarly, in the XML question, only 30% selected 1, whilst remaining scores were spread across options 2 to 4.

Figure 5. Comparison Between How Easily Respondents Were Able to Decide Types of Content for XML and DC Elements.



Proportional averages and standard deviation confirm varying levels of ease deciding kinds of format content, this is seen in the wide dispersion from the mean average for each series of results; the mean average for the XML question was 2.2, with a standard deviation of 0.979796, whilst the mean average for the Dublin Core was 2, with a standard deviation of 1.

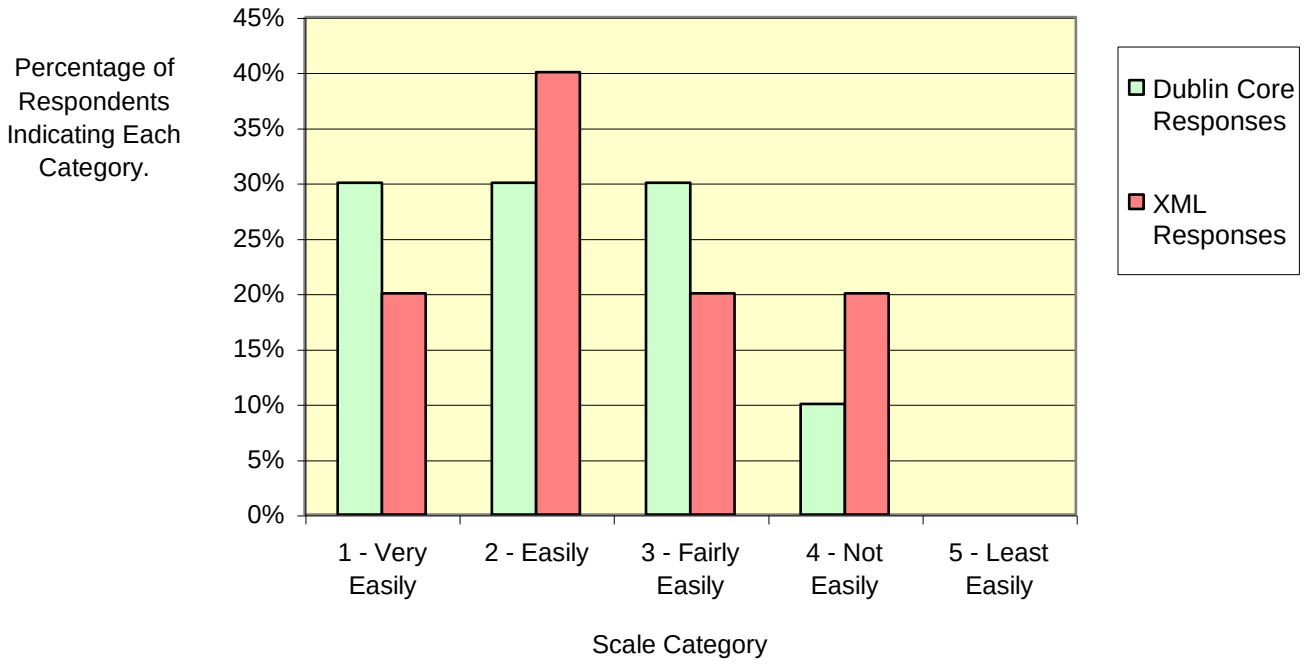
Results therefore show that a significant proportion of respondents thought that choosing types of content was not entirely easy for either format, and that it was slightly harder to choose XML content than DC content.

This trend may be due to the open-ended nature of XML, whilst DC element names provide some indication of content form, eg: the *Keywords* and *Description* tags.

Another content-related comparison may be seen between question 12, which asked how easily respondents were able to decide Dublin Core keywords, and question 20, which asked how easily respondents were able to define XML tags.

As in the previous example, scores were widely spread, with only 20% selecting option 1 for the XML question, and 30% selecting 1 for the Dublin Core, whilst a significant proportion selected options 2 to 4 across both questions (75% overall.)

Figure 6. Comparison Between How Easily Respondents Were Able to Decide Terms to Define XML Tags and DC Keywords.



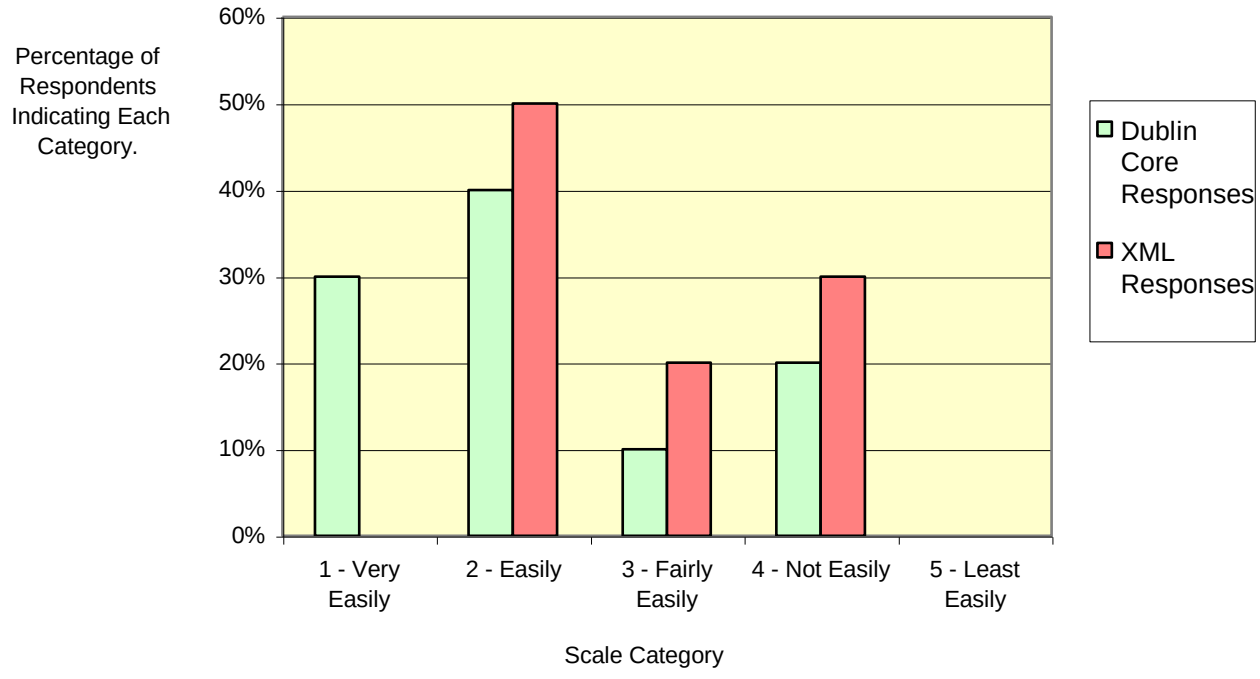
These results indicated that a significant number of respondents found it difficult choosing terms for inclusion in both formats.

The final content-related questions included question 13, which asked how easily respondents were able to compile free text for inclusion in the Dublin Core *description* field, and question 21, which asked how easily respondents were able to decide key terms or free text for inclusion in XML containers.

As in previous questions of this kind, there was a balance across all scores.

In the Dublin Core question, only 30% selected 1 (Very Easily), whilst 20% selected 4 (Not Easily.) In the XML question, none selected 1, whilst 20% selected option 3 (Fairly Easily), and 30% selected 4 (Not Easily.)

Figure 7. Comparison Between How Easily Respondents Were Able to Produce Terms or Free Text for Inclusion in XML Tags and the DC Description.



These results indicate that a significant proportion found it difficult to define terms in each format; this is confirmed by a mean average of 2.2 for the Dublin Core question, and average of 2.4 for the XML experiment.

Additionally, these averages also indicate that choosing terms was harder in XML than the Dublin Core, indicated by a greater central tendency to a score of 1 across the Dublin Core responses.

These compilation accessibility questions allow valuable comparisons between user responses to each format; averages and standard deviation indicate high levels of ease using the conventions and syntax of XML and the Dublin Core; however, user responses to the compilation of actual content was widely spread across the scale of 1 to 5 in both formats, with a significant proportion who found aspects of content compilation difficult.

Respondents were slightly more comfortable using XML syntax and conventions than those of the Dublin Core (questions 8 and 16); however, a significant proportion found aspects of content compilation easier in the Dublin Core than XML, possibly due to the open-ended nature of XML.

In conclusion, well-formed XML, based on user-defined data was shown to be less accessible for compilation than the Dublin Core, due to the difficulty of users in deciding XML tag names and kinds of content.

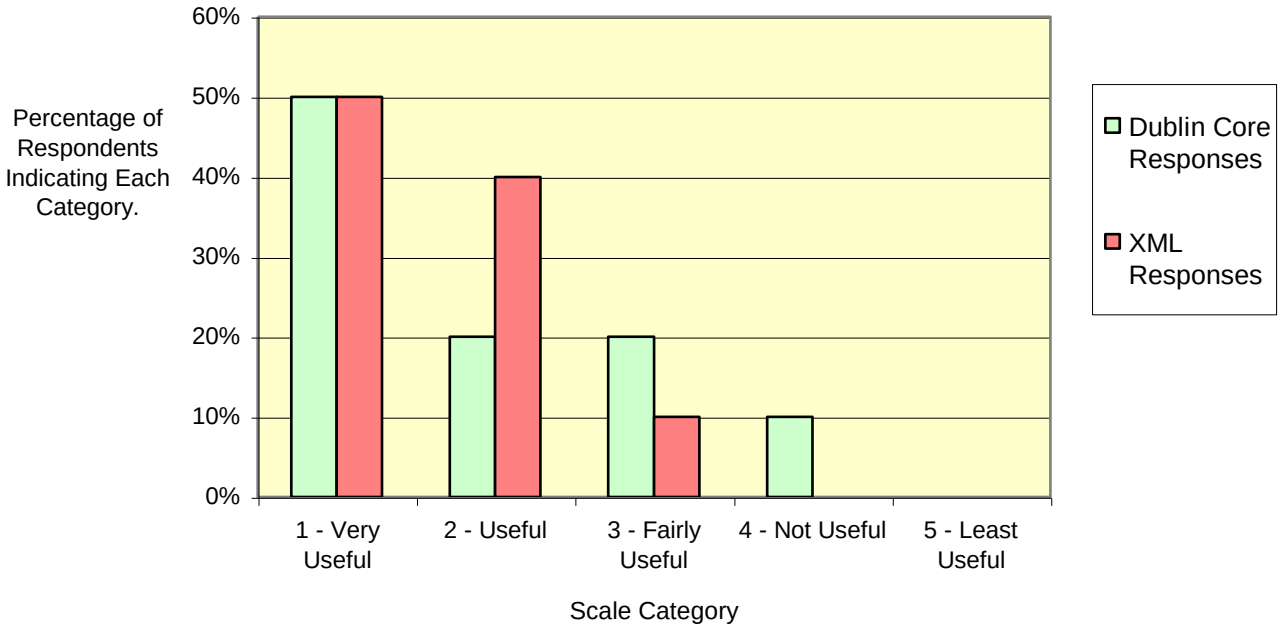
10.4. Script protocol transparency with prevalent markup language syntax, conventions and structures.

Another key area for investigation in the practical research element was the degree of similarity, or *transparency* between emerging metadata standards and well established standards, such as HTML script. Script transparency was investigated to assess the extent that HTML authors could apply existing skills to more recent metadata structures

Markup language transparency was assessed across a series of related questions; the first of these included questions 10 and 18, which asked how useful the respondent's HTML experience was in 'understanding and using' each format.

50% selected option 1 (Very Useful) for both questions, although a significant proportion selected lower scores.

Figure 8. Comparison Between How Useful HTML Experience was In Compiling XML and DC Metadata.



Averages and standard deviation demonstrated that HTML experience was more useful in using XML than the Dublin Core, with a mean average of 1.9 and standard deviation of 1.0440307 for the Dublin Core question, and mean average of 1.6 and standard deviation of 0.663325 for the XML question, (proving that there was a greater central tendency to 1 in the XML question.)

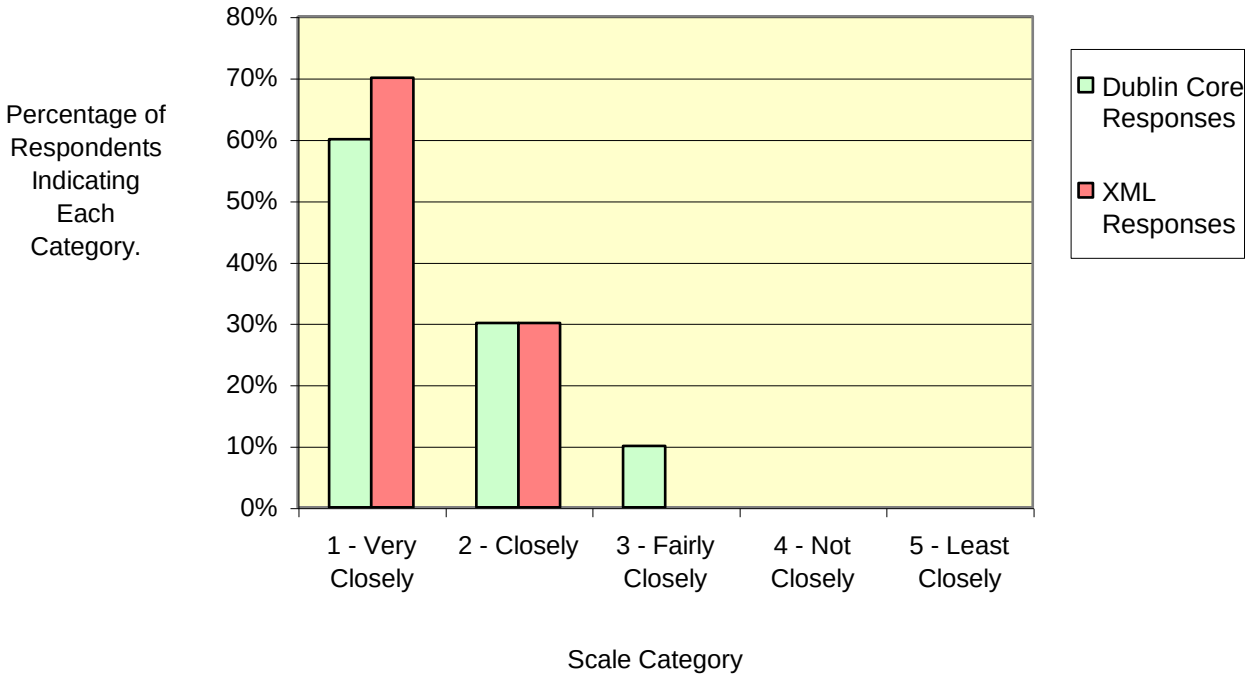
Results therefore indicate that the majority of respondents thought HTML experience was important in using and understanding both formats, but that HTML experience was most important in using XML.

These results were closely reflected in questions 9 and 17, which asked how closely each format 'followed conventional HTML conventions and syntax.'

A high proportion selected option 1 (Very Closely) in each question, with 60% selecting 1 for the Dublin Core, and 70% selecting 1 for XML. Additionally, a significant proportion selected option 2 (Closely) in each question.

(Overleaf: chart comparing how closely each format followed HTML syntax and conventions.)

Figure 9. Comparison Between How Closely DC and XML Followed HTML Syntax and Conventions.



Mean averages across all the responses for each question also indicated the closer resemblance of XML to HTML, with a mean average of 1.5 for the Dublin Core question, and mean average of 1.3 for the XML question.

These results indicate that the majority of respondents considered each format very similar to HTML conventions and syntax, although XML was considered slightly more similar to HTML than the Dublin Core.

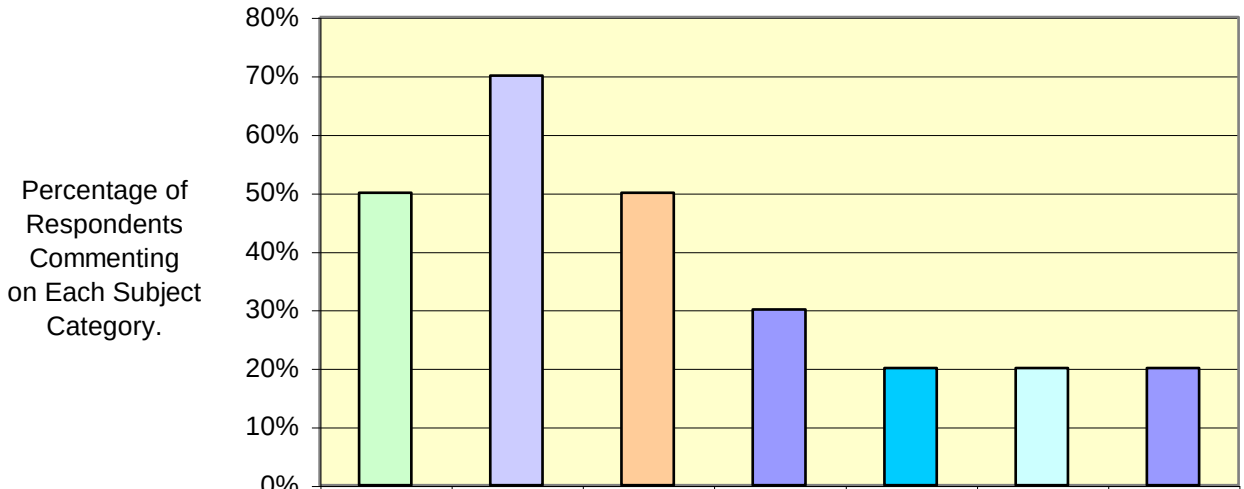
Significantly, these results suggest that Dublin Core metadata is less dependent on HTML knowledge than XML, perhaps due to the differing structures used within each format. In Dublin Core data, a single statement is used to define a data class (or element) and its content, whilst XML requires a clear understanding of HTML start and end tags to define XML elements classes and contents.

Results therefore indicate that well-formed XML requires a sound understanding of markup conventions, whilst the Dublin Core is less markup dependent, and possibly more accessible to the non HTML programmer.

10.5. Open-Ended Responses.

A wide number of comments were made on a variety of metadata-related issues in the open-ended questions, 14, 22 and 23.

Figure 10. Question 23: Further Comments or Opinions Suggested by Respondents.



One of the key comments to emerge from these questions was the importance of Search Engine compatibility with metadata standards. In question 23, 50% recommended increased support for the META tag by Search Engines for more

efficient indexing, and 70% complained that user awareness of the META tag was too low to ensure its effective contribution to Search Engine processes.

There were several comments on approaches to metadata, and related standards.
In question 23, a significant proportion (20%) commented on the need for a standard vocabulary for the selection of terms. Several respondents also queried HTML-based metadata in this question, with 20% suggesting that an external XML file accompanying the HTML document would be preferable to the HTML Meta tag.

Additionally, 30% suggested that metadata was most useful for personal reasons (such as document markup), rather than online indexing.
Browser compatibility with metadata models was also widely queried; in question 23, 50% commented on limited XML support by browsers.
These open-ended responses indicate a high awareness of the technology surrounding metadata, and an awareness of the problems faced by widespread implementation of metadata standards (such as user awareness and Search Engine compatibility.)

10.6. Analysis Conclusions.

These results indicated that users were very comfortable using both formats, although respondents found it easier compiling content for the Dublin Core than XML, suggesting that Dublin Core metadata was the most accessible format.

The results demonstrated high levels of participation in basic metadata standards, such as the META tag, and high levels of understanding on the processes surrounding

Search Engine indexing and ranking. Additionally, many respondents were aware of advanced markup structures, used to carry metadata such as XML, although few had used these standards.

Similarly, the results demonstrate low awareness and use of standard formats for expressing metadata, such as RDF and the Dublin Core, suggesting that whilst users are more familiar with markup structures, they are largely unfamiliar with abstract standards for expressing metadata within markup scripts.

Breakdown of Analysis:

- High awareness of the META tag standard, and markup languages but low awareness of abstract metadata standards expressed using markup, such as RDF and the Dublin Core.

- Search Engine indexing processes are well known, and that this is the primary reason for metadata compilation.

- Respondents felt very comfortable using XML and Dublin Core metadata, but many felt more comfortable using XML than they did using the Dublin Core.

- Respondents felt very comfortable using XML and Dublin Core conventions and syntax, but felt more comfortable using the conventions and syntax of XML than the Dublin Core.

Continued Overleaf.

- Many respondents found choosing types of content for both XML and the Dublin Core was easy, but others found this difficult. On average, respondents found it harder to choose XML content than Dublin Core content.

- Some respondents would have liked to use a standard vocabulary to inform choice of content type and choice of words to define content.

- Many respondents found choosing terms for both XML and the Dublin Core was easy, but others found this difficult. On average, respondents found it harder to choose XML terms than DC terms.

- Many respondents found compiling content for both XML and the Dublin Core was easy, but others found this difficult. On average, respondents found it harder to compile XML content than Dublin Core content.

- HTML experience was important in using and understanding both formats, and HTML experience was most important in using XML than in the Dublin Core.

- The majority of respondents considered that each format followed HTML conventions and syntax closely, although XML was considered slightly more markup-dependent than the Dublin Core.

- Most respondents were aware of the limited metadata support provided by prevalent resource indexing systems, such as Search Engines.

Chapter 11.

Conclusion.

This project has demonstrated that metadata serves a twofold process: as a facilitator of resource description and control in the academic and research community, and as an aide to effective resource control amongst the general majority of HTML authors.

Specialist cataloguing projects, such as CORC are role models against which Search Engine standards should be judged, although it is doubtful that these catalogues (CORC, RSLP, Medlane, etc.) will ever become widely used by the HTML authoring community.

Whilst online catalogues like CORC, will serve as an invaluable gateway to the Internet for the academic community, their lengthy submission and resource cataloguing processes are practical only for information professionals, rather than the general user.

Foreseeably, the only realistic context for specialist catalogues are as a central hub, or Information Gateway for academic resources, with the submission process administered by information professionals within organisations on the behalf of academics, researchers etc. compiling online resources.

This process would closely reflect the current role of information services within academic organisations, as facilitators of information classification and management. It is foreseeable that in the future, information services could require departments specialising in online resource submission and cataloguing in cooperation with projects such as CORC.

The role of metadata standards is therefore well established amongst specialist projects intended to serve the academic or research community, and it is foreseeable that these projects will gain more widespread use amongst this industry in the near future.

However, the greatest problem facing the World Wide Web, is the effective control of resources compiled by the remaining majority of HTML authors, for personal, commercial and other uses.
This project has demonstrated that the role of Search Engines is crucial to an effective solution in resource description and retrieval for Web users and developers as a whole.

Importantly, this project has shown that the Dublin Core represents the most accessible form of metadata for HTML authors, whilst XML is more suitable for proprietary online databases and as a standard record format for shared use across specialist Information Providers.

Whilst XML could become an optional feature of Search Engine indexing, perhaps based on a standard DTD, there would probably not be enough widespread use to enable Search Sngines to rely on this format alone.

If the Internet is to become an effective and efficient source of information, Search Engines must implement enhanced metadata strategies, as outlined in the Recommendations of this project (Chapter 12). Also crucial to the success of metadata use by prevalent Information Providers, is the implementation of standard formats across all leading participants, thus allowing automated systems to effectively index resources without requiring proprietary data.

The practical research element revealed that HTML authors are aware of metadata concepts and some metadata standards; in addition, they are comfortable using these standards, since both Dublin Core and XML were shown to closely resemble conventional HTML.

The practical research also demonstrated significant user awareness of the status of metadata on the Internet; many respondents were aware of the limited support for metadata amongst current Search Engines and browsers, contradicting current Search Engine claims that use of metadata is untenable, due to limited user awareness, and competence using formats.

In addition to metadata-compliant online catalogues and Search Engines, metadata standards should use appropriate qualifiers and standard terms, allowing for meaningful classification and element-specific searching.
In the case of specialist online catalogues, such as CORC, this is already possible, where the user may search within data categories, such as Dublin Core elements, or element sub-fields, such as the Library of Congress Subject Headings.

In the case of a widely used format, perhaps based on the Dublin Core, standard terms, date formats or subject headings would seem impractical, since compilation would largely rely on user discretion; however, were software used to regulate metadata compilation, either using online or offline programs, at least some degree of content control could be effected. (A detailed overview of the possibilities for application-based metadata compilation is described in chapter 12, *Recommendations*.

The long term agenda, is surely the development of consensual approaches for the interpretation, indexing and retrieval of metadata by prevalent Information Providers.

Additionally, profit-making organisations (such as software companies and Search Engines) and official bodies (such as the W3C) must agree practical metadata standards for use within existing technology; these standards should also be user-friendly enough to gain widespread use amongst the HTML authoring community.

The World Wide Web is an inclusive and ever growing society, demanding user-friendly solutions to the problems of structure and volume. If online indexing structures are to become effective, meeting the needs of our growing information society, this large user base must not be ignored, and must be encouraged to participate in the uses and evolution of metadata.

Chapter 12.

Recommendations.

This chapter outlines recommendations for improved resource description structures on the World Wide Web; these recommendations are based on observations from the literature search, the practical research element, and conclusions in chapter 11. (Note: some of these recommendations represent long term solutions, involving extensive modifications to Information Provider systems.)

12.1. Recommendations for Search Engines.

1. The adoption of the 15 element Dublin Core standard by Search Engines, using a standard element definition set and standard qualifiers. The practical research indicated that many thought metadata was not widely supported enough by Search Engines. Common specifications should be adopted following mutual consultation between all leading Search Engine administrations and the W3C. In the practical research element, respondents were comfortable using the Dublin Core, and found it easier compiling DC content than XML, possibly because the process of defining XML tags is less intuitive than using pre-defined DC tags.

2. Search engines should implement support for the Dublin Core across a range of indexing, ranking and parser interface functions, including the following:

- Interrogation of Dublin Core tags, if available as the default procedure for ranking following a natural language user query, rather than entire HTML content.

- User-defined functions to search within a specified DC element or range of elements, possibly by using pull-down menus or check boxes to select elements.

- Functions to search by specifying DC element content within DC categories, possibly using metastatements such as the Lycos 'applet:' statement or pull-down menus; examples might include metastatements for 'creator:', 'type:' or 'location:'; search terms could include standard qualifiers, possibly selected using a pull-down menu of qualifier terms, such as MIME standards for the TYPE element.

  In the research element, respondents commented on the importance of controlled terms. Metastatements for Dublin Core data would be invaluable in effectively retrieving classes of resource, such as the MIME type. Whilst Alta Vista and Lycos use many similar metastatements, these would be far more effective at identifying specific classes of content, since conventional systems interrogate the entire HTML content, as opposed to these content-specific tags.

Continued Overleaf.

- Adoption of official Dublin Core schemes and scheme-qualifiers for use within advanced search functions. Controlled content, such as ISO country names, language codes, and scheme qualifiers, such as LCSH (Library of Congress Subject Headings) could be supported

- Search Engines could provide an online application for inserting Dublin Core tags in Web pages; this application could be made available via the Search Engine Web page, generating publicity and increased use of DC tags. Applications of this kind have already been developed, including Sirpac (UKOLN. 2000), and the DC Dot. An application of this kind should support DC schemes and qualifiers via pull-down menus for inserting controlled content. Search engines could make this process mandatory for Web page submissions. The practical research indicated that HTML experience is important in compiling metadata, so that a GUI approach would seem essential for non-HTML authors who use HTML editors, rather than plain HTML script.

3. There should be at least one standard and one advanced metadata format supported by Information Providers; the Dublin Core provides various levels of scope for metacontent, where individuals may simply include standard elements in their page, or include element schemes for particular kinds of content such as LCSH terms. Advanced structures, using a wider number of elements could also be supported by Search Engines; advanced structures intended for popular use across the Internet would require the same level of user accessibility seen in the Dublin Core, possibly using HTML META tags, rather than an external SGML-based format. Search Engine users could search for data by using metastatements

specific to a particular format. An advanced metadata format could be used to classify resources requiring greater detail than is currently offered by the Dublin Core, such as academic, scientific or research-based resources.

12.2. Recommendations for Specialist and Research-based Online Catalogues.

1. As has been demonstrated in the CORC project, major metadata compliant systems should provide compatibility and conversion features between metadata formats, allowing backwards compatibility between more detailed metadata models, such as MARC and less detailed formats, such as Dublin Core.

2. XML could be effectively used by a larger range of specialist Information Providers such as BUBL and OCLC Firstsearch. Library and Information Services could also use XML as a transport layer, and for expressing a standard record format (such as MARC). A common DTD structure would allow the sharing of record files (although the profit-based practices of some organisations would prevent this.)

Continued Overleaf.

12.3. Recommendations for the IT Industry.

1. Metadata should be taken seriously by all leading software companies, such as Microsoft and Sun Microsystems; a publicity campaign by these organisations, possibly including Web advertising and greater coverage of metadata in Web development software manuals might increase user awareness of standard formats.

2. Metadata standards described above, such as an industry-standard Dublin Core model should be supported via leading Web development applications; the Microsoft Frontpage application supports plain text input for the META keywords tag using a graphical interface; this could easily be amended to include all 15 Dublin Core META tags.

12.4. Recommendations for the HTML author.

I recommend the following short term solutions to increasing metadata effectiveness for individuals:

1. Web authors could construct Information Gateways, using a manually selective approach to provide access to useful online resources, thereby contributing to information dissemination and filtering on the World Wide Web.

Continued Overleaf.

2. Web authors may enhance the metadata functionality of their Web published material, through thoughtful and economic use of the *keywords and description* META tags; additionally, they should submit material to a Search Engine that uses these tags for indexing and ranking (such as Hotbot). Since 80% of respondents had heard of and had used the META tag, this does seem a viable option for at least the majority of competent HTML authors.

3. Information professionals may also contribute to online catalogue projects, such as CORC, thereby developing and encouraging the evolution of enhanced metadata standards on the World Wide Web.

4. Finally, Information professionals or any HTML author with interests in this area may attempt to join a research and development foundation, such as the World Wide Web Consortium, allowing them to contribute towards policy development, systems criteria, or suggest new functional tasks for metadata models.

Bibliography and References.

The Alta Vista Information Page. (2000). [Online]. Cited 01/5/'00.
http://www.altavista.com/av/content/help.htm#simple

Armstrong, C. (1997). *Metadata, PICS and Quality*. Ariandne 17. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/17

Berry, R. (1994). *The Research Project: How to write it.* London, Routledge.

Boyatzis, R.E. (1998). *Transforming Qualitative Information.* London, Sage.

Bradley, P. (1999). *The Altavista Relaunch / Personalised Search Engines*. Ariandne 22. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/22

Bradley, P. (2000). *The relevance of underpants to searching the Web*. Ariandne 24. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/24

The British Library Home Page. (2000). [Online]. Cited 01/5/'00.
http://portico.bl.uk

Brunard, L. (2000). *Text Encoding for Interchange: a new Consortium*. Ariandne 24. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/24

Catherall, P. (2000.) *Tudalen Catref Gwledig*. [Online]. Cited 01/5/'00.
http://members.theglobe.com/gwledig/default.htm

Chapman, A. et al. (1998). *Cataloguing practice and Internet subject-based information Gateways*. Ariandne 18. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/18

Chepesuik, R. (1999). Organising the Internet – the core of the challenge. *American Libraries*, 30 (1) January, 60-63.

The CIDOC Home Page. (2000). [Online]. Cited 28/06/'00. http://www.cidoc.icom.org/cidoc0.htm

The CIMI Home Page. (2000). [Online]. Cited 28/06/'00. http://www.cimi.org/

Comer, D. et al. (1993). *Internetworking with TCP/IP, Vol. 2: Design, Implementation and Internals*. New Jersey, USA, Prentice Hall Publishing.

Comer, D. et al. (1993). *Internetworking with TCP/IP, Vol. 3: Client-Server Programming and Applications.* New Jersey, USA, Prentice Hall Publishing.

Cromwell-Kessler, W. (1998). Crosswalks, Metadata Mapping and Interoperability. *In:* Baca, M. ed. *Introduction to Metadata.* USA, The Getty Information Institute.

Day, M. (1998). *At the Event: Metadata and biodiversity information: a report from a US symposium on "Metadiversity"*. Ariandne 18. [Online]. Cited 28/06/'00 . http://www.ariadne.ac.uk/issue1/elib/18

Day, M. (1998). *Image retrieval: combining content-based and metadata-based approaches*. Ariandne 17. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/17

Day, M. (1999). *Metadata for digital preservation: an update*. Ariandne 22. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/22

Day, M. (1997). *Working Meeting on Electronic Records Research*. Ariandne 8. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/8

Dawson, A. et al. (1997). *How BUBL benefits academic librarians*. Ariandne 10. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/10

Dempsey, L. and Heery, R. (1998). Metadata: a current view of practice and issues. *The Journal of Documentation*, 54 (2) March, 145-172.

Donnellan, C. (1997). *The Internet - Marvel or Menace?* Cambridge, UK, Independence Educational Publishers.

Dovey, M. (1998). *Meta-Objects - An Object Oriented approach to metadata.* Ariandne 19. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/19

The Dublin Core Home Page. (2000). [Online]. Cited 01/5/'00.
http://purl.org/DC/

Dunning, A. (1999). *Do We Still Need Search Engines?* Ariandne 22. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/22

Edwards, J. (1998). *The good, the bad and the useless: evaluating Internet resources*. Ariandne 17. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/17

Emmott, S. (1998). *At the Event: SGML, XML, and Databases*. Ariandne 18. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/18

Feeney, M. ed. (1999). *Digital Culture: maximising the Nation's Investment.* London, The National Preservation Office.

Fink, A. (1995). *How to analyse survey data.* London, Sage.

Fink, A. et al. (1985). *How to conduct surveys.* London, Sage.

Fink, A. (1995). *How to design surveys.* London, Sage.

Fink, A. (1995). *How to sample in surveys.* London, Sage.

Fowler, Floyd. J. (1995). *Improving Survey Questionnaires.* London, Sage.

Gardner, T. (1999). *MODELS 9 and MODELS 10.* Ariandne 22. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/22

The Getty Institute Homepage. (2000). [Online]. Cited 28/06/'00.
http://www.getty.edu/gri/standard/

The Getty Standards Programme. (2000). [Online]. Cited 28/06/'00.
http://www.getty.edu/gri/standard/fda/index.htm

Gill, Tony. (1998). Metadata and the World Wide Web. *In:* Baca, M. ed. *Introduction to Metadata.* USA, The Getty Information Institute.

Gilliland-Swetland, A.J. (1998). Defining Metadata. *In:* Baca, M. ed. *Introduction to Metadata.* USA, The Getty Information Institute.

The Google Home Page (2000). [Online]. Cited 01/5/'00.
http://www.google.com

Heery, R. et al. (1998). *CrossROADS and Interoperability*. Ariandne 15. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/15

Heery, R. (1999) *Rachel Heery's Introduction to RDF*. [Online]. Cited 01/5/'00.
http://www.ariadne.ac.uk/issue14/what-is/

Heery, R. (1998). *What is... RDF?* Ariandne 15. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/15

Hodson, Peter. (1995). *Local Area Networks.* London, Guernsey Press.

The Hotbot Home Page. (2000). [Online]. Cited 01/5/'00.
http://www. hotbot.com

Houghton, D. (1996). *Displaying SGML documents on the World Wide Web*. Ariandne 6. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/6

The Internet Engineering Taskforce Home Page. (2000). [Online]. Cited 28/06/'00. http://www.ietf.org

The ISO (International Organization for Standardization) Homepage. (2000). [Online]. Cited 28/06/'00. http://www.iso.ch/

Judge, P. (1988). *Open Systems: The basic guide to OSI and its implementation.* London, Reed Business Publishing.

Kelly, B. (2000). *Reflections On WWW9*. Ariandne 24. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/24

Kelly, B. (1999). *Using The Web To Promote Your Web Site*. Ariandne 22. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/22

Kelly, B. (1997). *WebWatch: A Survey Of Numbers of UK University Web Servers*. Ariandne 8. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/8

Kelly, B. (1998). *What are...Document Management Systems?* Ariandne 18. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/18

Kelly, B. (1998). *What Is... XML?* Ariandne 15. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/15

Kerr, L. (1998). *Subject-based Information Gateways*. Ariandne 18. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/18

Kling, R. (1999). Can the "Next Generation Internet" Effectively Support "Ordinary Citizens"? *Information Society,* 15 (1) 57-63.

Knight, J. (1997). *Handling MARC with Perl*. Ariandne 8. [Online]. Cited 28/06/'00. http://www.ariadne.ac.uk/issue1/elib/8

Knight, J. (1997). *Making a MARC with the Dublin Core*. Ariandne 8. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/8

Leventhal, M. et al. (2000). *Designing XML Internet Applications*. London, Prentice Hall International.

The Library of Congress Homepage. (2000). [Online]. Cited 01/5/'00.
http://lcweb.loc.gov

The Lycos Home Page. (2000). [Online]. Cited 01/5/'00.
http://www.lycos.com

Marshall, C. and Rossman, G.B. (1995). *Designing Qualitative Re*search. London, Sage Publications.

Marshall, P. (1997). *Research Methods: how to design and conduct a successful project.* London, How To Books.

Mason, J. (1996). *Qualitative Researching*. London, Sage Publications.

Martin, W.J. (1995). *The Global Information Society.* Hampshire, Aslib Gower Press.

McNab, A. et al. (1997). *Never mind the quality, check the badge-width!* Ariandne 9. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/9

The Medlane Project Home Page. (2000.) [Online]. Cited 01/5/'00.
http://xmlmarc.stanford.edu/

The MDA Home Page. (2000). [Online]. Cited 28/06/'00.
http://www.open.gov.uk/mdocassn/mda_st01.htm#

The MDA SPECTRUM Specifications. (2000). [Online]. Cited 28/06/'00.
http://www.open.gov.uk/mdocassn/workshop.htm

The Microsoft XML Notepad Page. (2000). [Online]. Cited 28/06/'00.
http://msdn.microsoft.com/xml/notepad

The Microstar Home Page. (2000.) [Online]. Cited 28/06/'00.
http://www.microstar.com

Miller, E. (2000.) *An introduction to the Resource Description Framework.* [Online]. Cited 01/5/'00.
http://www.dlib.org/dlib/may98/miller/05miller.html

Miller, P. ed. (1998). *Discovering Online Resources Across the Humanities: A Practical Implementation of the Dublin Core.* Bath, The UK Office for Library and Information Networking.

Miller, P. (1998). *Dublin comes to Europe.* Ariandne 15. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/15

Miller, P. (2000). *Interoperability What is it and Why should I want it?* Ariandne 24. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/24

Miller, P. (1999*). I say what I mean, but do I mean what I say?*
Paul Miller reports on outcomes from January's MODELS 11 workshop. Ariandne 23. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/23

Miller, P. et al. (1997). DC5: *The search for Santa.* Ariandne 12. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/12

The MODELS Home Page. (2000). [Online]. Cited 28/06/'00.
http://www.ukoln.ac.uk/dlis/models/

Morris, A. et al. (1996). *Overview of EDD Research and Services.* London, South Bank University.

The Mozilla Organisation. (2000). [Online]. Cited 01/5/'00.
http://www.mozilla.org

The Mozilla Organisation RDF page. (2000). [Online]. Cited 01/5/'00.
http://www.mozilla.org/rdf/50-words.html

The National Preservation Office. (2000). [Online]. Cited 28/06/'00.
http://www.bl.uk/services/preservation/

North West INNOPAC Users Conference, NEWI, Wrexham.
[Attended 20 June, 2000.)

The OCLC CORC Home Page. (2000). [Online]. Cited 01/5/'00.
http://purl.oclc.org/corc

The OCLC CORC Training Site. (2000). [Online]. Cited 01/5/'00.
http://purl.oclc.org/corc/practice

The OCLC Homepage. (2000). [Online]. Cited 01/5/'00.
http://www.oclc.org/oclc/research/projects/core

The OCLC PURL Homepage. (1999). [Online]. Cited 01/5/'00.
http://purl.oclc.org/

Oliver, D. and Holzschlag, M. (1998). *HTML 4 in 24 hours.* USA, Sams.net Publishing.

The Oxford Text Initiative. (2000). [Online]. Cited 28/06/'00.
http://ota.ahds.ac.uk/

Powell, A. (1998). BIBLINK. *Checksum - an MD5 message digest for Web pages.* Ian Peacock and Andy Powell. Ariandne 17. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/17

Powell, A. (1997). *Dublin Core Management*. Ariandne 10. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/10

Powel, A. (1999). *Introduction to RDF*. [Online]. Cited 01/5/'00.
http://www.ukoln.ac.uk/metadata/presentations/ukolug98/paper/intro.html

The REACH Element Set. (2000). [Online]. Cited 28/06/'00.
http://www.rlg.org/reach.html

The Research Support Libraries Project. (2000). [Online]. Cited 28/06/'00.
http://www.ukoln.ac.uk/metadata/rslp/

The ROADS Specifications Home Page. (2000). [Online]. Cited 28/06/'00.
http://www.ukoln.ac.uk/metadata/roads/what/intro.html

Rowley, J. (1998). *The Electronic Library.* Emeryville, USA, Library Association Publishing.

Reid, S. (1987). *Working with statistics.* Oxford, Basil Blackwell.

Russell, R. (1998). *At The Event: A Distributed National Electronic Resource?* Ariandne 15. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/15

Russell, R. (1997). *MODELS: Moving to Distributed Environments for Library Services.* Ariandne 8. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/8

Shafer, K. (1999). *Keith Shafer's PURL Homepage*. [Online]. Cited 01/5/'00.
http://www.oclc.org/~shafer/

Shipley, C and Fish, M. (1996). *How the World Wide Web Works.* Emeryville, USA, Macmillan Computer Publishing.

Stanley, T. (1997). *Ask Jeeves: the Knowledge Management Search Engine*. Ariandne 17. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/17

Stanley, T. (1997). *Alta Vista LiveTopics*. Ariandne 9. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/9

Stanley, T. (1996). *Alta Vista vs. Lycos. The Great Search Engine Debate*. Ariandne 2. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/2

Stanley, T. (1997). *Keyword Spamming: Cheat Your Way To The Top*. Ariandne 10. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/10

Stanley, T. (1998). *Meta-Searching on the Web*. Ariandne 16. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/16

Stanley, T. (1997). *Moving Up The Ranks*. Ariandne 12. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/12

Stanley, T. (1998). *Northern Light: A Leading Search Light?* Ariandne 15. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/15

St. Laurent. S, (2000). *Simeon St. Laurent's XML Homepage*. [Online]. Cited 01/5/'00.
http://www.simonstl.com

Stobart, S. et al. (1996). *An investigation into World Wide Web Search Engine use from within the UK - preliminary findings*. Ariandne 6. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/6

The TEI Homepage. (2000). [Online]. Cited 28/06/'00.
http://www.uic.edu/orgs/tei/

The TEI Specifications. (2000). [Online]. Cited 28/06/'00.
http://www.hcu.ox.ac.uk/TEI/Lite/

Trickey, K. (1998). Information Organisation on the Web? It is basically about respect and trust. *Library Review,* 47 (2) 135-137.

Tseng, G. et al. (1996). *The Library and Information Professional's Guide to the Internet.* Bath, Library Association Publishing.

Turner, T.P. and Brackbill, L. (1998). Rising to the top: evaluating the use of the HTML META tag to improve retrieval of World Wide Web documents through Internet Search Engines. *Library Resources and Technical Services,* 42 (4) 259-265.

The UKOLN Homepage. (2000). [Online]. Cited 01/5/'00.
http://www.ukoln.ac.uk/

The VRA Home Page. (2000). [Online]. Cited 28/06/'00.
http://www.oberlin.edu/~art/vra/vra.html

Wang, H. et al. (1999). Service Quality of Internet Search Engines. *The Journal of Information Science,* 25 (6) 499-507.

The Web Developer.com Home Page. (2000). [Online]. Cited 28/06/'00.
http://www.stars.com

Welsh, S. (1997). *OMNI Corner: Harvesting.* Ariandne 7. [Online]. Cited 28/06/'00.
http://www.ariadne.ac.uk/issue1/elib/7

The W3C XML Homepage. (1999). [Online]. Cited 01/5/'00.
http://www.w3.org/XML/

The W3C RDF Specifications. (1999). [Online]. Cited 01/5/'00.
http://www.w3.org/TR/REC-rdf-syntax/

The WWW Conference Consortium Report on XML. (1999). [Online]. Cited 01/5/'00.
http://www.w3.org/TR/1998/NOTE-XML-data

The WWW International Consortium. (1999). [Online]. Cited 01/5/'00.
http://www.w3.org/Consortium/Process/#RecsW3C

The WWW International Consortium SIRPAC workgroup. (1999). [Online]. Cited 01/5/'00.
http://www.w3.org/RDF/Implementations/SiRPAC/

The WWW International Consortium workgroup on XML. (1999). [Online].
Cited 01/5/'00.
http://www.w3.org/TR/1998/WD-xml-names-19980916

The Yahoo Home Page. (2000).). [Online]. Cited 01/5/'00. http://www.yahoo.com

Younger, J. A. (1997). Resource description in the digital age, *Library Trends*, 45 (3) 462-87.

Appendices

Appendix A.

Comparison Table of Metadata Format Elements.

Comparison of Metadata Standards Table 1.

	CDWA *** Bold** = Core Standard. ***** *Italics* = Sub-Category	**CIMI**	**GDAD** *** Bold** = Core Standard.	**MESL**
1.	**Object/ Work**		**Document Classification Catalogue Level**	
2.	*Object/Work-Type*	Object Type	**Document Classification Document Type**	Object Type/ Object Class/ Object Name.
3.	*Object/Work Components*		Document Classification - Extent	Parts/ Pieces
4.	Classification			
5.	Orientation/ Arrangement			
6.	Titles or Names	Title	Group/ Item Identification - Repository Title.	Object Title/ Caption
7.	State			Edition/ State
8.	Edition			Edition/ State
9.	Measurements			Dimension/ Unit
10.	*Measurements - Dimensions*	Measurements	Physical Characteristics Dimension Description.	
12.	*Measurements Dimensions - Type and Value*		Physical Characteristics - Height, Width and Depth.	
13.	*Measurements Dimensions - Unit*		Physical Characteristics - Unit of Measurement.	
14.	*Measurements - Scale*		Physical Characteristics - Scale Description	
15.	Materials and Techniques			
16.	*Materials and Techniques - Description*	Materials and Techniques	Physical Characteristics - Technique, Medium & Support Description.	
17.	*Materials and Techniques - Processes or Techniques*			Creation Technique/ Method & Process.

Comparison of Metadata Standards Table 1.Continued.

	The VRA Core.	**REACH**	**USMARC**	**The Dublin Core**
1.				
2.	W1. Work Type.	Field #1: Type of Object	655 Genre Form	Type or Source Type
3.			300a Physical Description-Extent	
4.				Subject or Source Subject
5.				Description or Source Description
6.	W2. Title	Field #4: Object Name/ Title	24Xa Title and Title Related Information	Title or Source Title
7.			562c Copy and Version	Description or Source Description
8.			250 Edition Statement.	
9.				
10.	W3. Measurements	Field #7: Dimensions	340b Physical Medium Dimensions or 300c Physical Description Dimensions	
12.				
13.				
14.				
15.				
16.				
17.	W5. Technique	Field #5: Techniques/ Processes	340d Physical Medium-Information Recording Technique.	

Comparison of Metadata Standards Table 2.

	CDWA *** Bold** = Core Standard. ***** *Italics* = Sub-Category	**CIMI**	**GDAD** *** Bold** = Core Standard.	**MESL**
18.	Materials and Techniques-Processes		Physical Characteristics-Technique	
19.	Materials and Techniques-Implementation		Physical Characteristics-Technique	
20.	Materials and Techniques-Materials	Material Medium	Physical Characteristics-Medium	Materials/ Medium
21.	Fracture			
22.	Physical Description	Physical Description		
23.	Inscription/ Marks	Inscription/ Mark	Physical Characteristics-Inscription- Description	Marks/ Inscriptions
24.	Condition/ Examination History.	Condition		
25.	Conservation/ Treatment History			
26.	**Creation**			
27.	**Creation- Creator**	Creator General	**Responsibility**	
28.	**Creation-Creator-Identity-Names**	Creator Name	**Origin/ Maker Name**	Creator/ Maker Name
29.	Creation-Creator-Identity-Dates			
30.	**Creation-Birth-Identity-Dates/Location-Birth**	Creator date of birth		
31.	**Creation-Creator-Identity-Dates/Location-Death**	Creator date of death		
32.	**Creation-Creator-Identity-Nationality/ Culture/ Race**	Creator Nationality-Culture/ Race		
33.	**Creation-Creator-Role**	Creator Role	**Origin/ Maker- Role**	Creator/ Maker Role
34	**Creation Date**	Date of Origin	**Date of Execution Descriptive Date**	Creation Begin & End Date

Comparison of Metadata Standards Table 2. Continued.

	The VRA Core.	REACH	USMARC	The Dublin Core
18.				
19.				
20.	W4. Material	Field #6: Medium/ Materials	340a Physical medium- Material Base and Configuration.	
21.				
22.				Format or Source Format
23.			562a Copy and Version Identification- Identifying Markings	
24.			5831 Action Note Status	
25.			583x Action Notes- Non-public or 583x Public Note	
26.				
27.				
28.	W6. Creator	Field #10: Creator/ Maker	1XX Main Entry and 7XX Added Entry	Creator or Source Creator
29.		Field #11: Dates of Creator/ Maker	1XX Main Entry Associated Dates	
30.				
31.				
32.	W15. Nationality/ Culture	Field #12: Nationality/ Culture of Creator/ Maker	65X Subject Index term	
33.	W7. Role		1XXe Main Entry Relator Term	
34.	W6. Date	Field #2: Date of Creation/ Date range	260c Imprint- Date of Publication	Date or Source Date

Comparison of Metadata Standards Table 3.

	CDWA *** Bold** = Core Standard. ***** *Italics* = Sub-Category	**CIMI**	**GDAD** *** Bold** = Core Standard.	**MESL**
35.	Creation Place	Place of Origin		Creation Place
36.	*Creation- Commission- Commissioner*	Association Name	Related people/ Corporate Bodies	
37.	*Ownership/ Collecting History*			
38.	*Ownership/ Collecting History/ Description*	Owner Provenance	Provenance- Provenance Description	
39.	*Ownership/ Collecting History/ Owner*	Owner Provenance	Provenance- Former Owner Name	
40.	*Ownership/ Collecting History/ Transfer Mode*			Accession Method
41.	*Ownership/ Collecting History/ Credit Line*	Credit Line		Credit line
42.	Copyright Restrictions	Copyright Restriction	Internal Documentation Restrictions	
43.	*Styles/ Periods/ Groups/ Movements*	Style Period/ Period Name		Concepts/ Style Period
44.	**Subject Matter**			
45.	*Subject Matter- Description*	Content General	Method of Representation/ Point of View- Method/ View Description	Concept/ Subject
46.	**Subject Matter- Description Indexing Terms**	Subject	Method of Representation/ Point of View- (Broad)	
47.	**Subject Matter- Identification Indexing terms**	Subject	Subject/ Built Work Identification- Subject/ Built Work Name	
48.	Context	Association/ General		Associated Events, People.
49.	*Context-Architectural*			
50.	*Context-Historical/ Cultural*	Context Historical		
51.	Context- Archaeological	Context Archaeological		
52.	Exhibition/ Loan History		Exhibition History	

Comparison of Metadata Standards Table 3. Continued.

	The VRA Core.	REACH	USMARC	The Dublin Core
35.		Field #3: place of origin/ Discovery	651 Subject Term- Geographical Name	
36.			7XX Added Entry 536 Funding Information Note	
37.				
38.			561a ownership and Custodial History- History	
39.		Field #13: Current Owner	6XX Subject Access Fields	
40.			541c Immediate Source of acquisition- Method of Acquisition	
41.			541c Immediate Source of acquisition- Method of Acquisition	
42.			540a Terms Governing Use and Reproduction	Rights or Source Rights
43.	W14. Style/ Period/ Group/	Field #9: Style/ Period/ Group/	65X Subject Index Term	Description or Source Description
44.				Description or Source Description
45.		Field #8: Subject Matter	520 Summary, etc.	Subject or Source Subject
46.	W16. Subject		65X Subject Index term	Subject or Source Subject
47.	W16. Subject		65X Subject Index term	
48.				Coverage or Source Coverage
49.	W13. Original Site W14. Current Site		651 Subject Access- Geographic Name	
50.		Field #3: Place of origin/ Discovery		Coverage or Source Coverage
51.	W13. Original Site W14. Current Site	Field #3: Place of Origin/ Discovery	651 Subject Access- Geographical Name	
52.			585a Exhibitions Note	

Comparison of Metadata Standards Table 4.

	CDWA *** Bold** = Core Standard. ***** *Italics* = Sub-Category	**CIMI**	**GDAD** *** Bold** = Core Standard.	**MESL**
53.	Related Works	Related Objects	Related Items.	
54.	*Related Works Relationship Type*			
55.	*Related Works- identification*			
56.	Related Visual Documentation	mr Object		
57.	*Related visual Documentation Relationship Type*	mr Object- Resource Type		
58.	*Related visual Documentation- Image Type*			
59.	*Related visual Documentation- Image Measurements*	mr Object Description		
60.	*Related visual Documentation- Image Ownership- Owner Name*	mr Object- Publisher		
61.	*Owner's Numbers*			
62.	*Related visual Doc. View*	mr Object Coverage		
63	*Related Visual Documentation View- Indexing Terms*	mr Object- Subject Keyword		
64	*Related visual Documentation- Image Source Name*			
65	*Related visual Documentation- Image Source Number*	Rendition Resource Identifier		Image File Name
66.	*Remarks*			Image Caption
67.	Related Textual References	Textual References	Citation	
68	*Related Textual References- ID*			
69.	*Related Textual References- Type*			Document Type
70.	*Related Textual References- Work- Illustrated*		Published Reproductions	
71	Critical Responses			
72	Cataloguing History		Document Note	
73	*Cataloguing History Language*			

Comparison of Metadata Standards Table 4. Continued.

	The VRA Core.	REACH	USMARC	The Dublin Core
53.		Field #20: Related Objects	580 Linking Entry Complexity Note	
54.	W18. Relationship Type		787g Non-specific Relationship Entry- Relationship Information	
55.	W17. Related Works		787n Non-specific Relationship Entry Note	
56.		Field #19: Electronic Location & Access		
57.			533a Reproduction Note- Type of Reproduction	
58.	V1. Visual Document Type		533e Reproduction Note- Physical Description	
59.	V3. Visual Document Measurements		533e Reproduction Note- Physical Description	
60.	V1. Visual Document Owner		533c Reproduction Note- Agency Responsible for Reproduction	
61.	V1. Visual Document Owner Number		533n Reproduction Note- Note about Reproduction	
62.	V1. Visual Document View Description		245p Title-Name of Part or Section of Work	
63.	V1. Visual Document Subject		65X Subject Index Tem	
64.	V1. Visual Document Source		533n Reproduction Note- Note about Reproduction	
65.			035 System Control Number	
66.				
67.			581 Publications about Described Materials	
68.				
69				
70				
71			581 Publications about Described Materials	
72			510 Citation/ Reference Note	
73.		Field #18: Language		

Comparison of Metadata Standards Table 5.

	CDWA *** Bold** = Core Standard. ***** *Italics* = Sub-Category	**CIMI**	**GDAD** *** Bold** = Core Standard.	**MESL**
74.	Cataloguing History- Remarks		Internal Documentation- Sources	
75.	**Current Location**			
76.	**Current Location- Repository Name**	Repository Name	**Group/ Item Identification- Repository Name**	Holding Institution
77.	**Current Location- Geographic Location**	Repository Place	**Group/ Item Identification- Repository Geographic Location**	
78.	**Current Location- Repository Numbers**		**Group/ Item Identification- Group/ Item ID Code**	Accession Number
79.	**Descriptive Note**		**Descriptive Note**	Description

Comparison of Metadata Standards Table 5. Continued.

	The VRA Core.	**REACH**	**USMARC**	**The Dublin Core**
74.				
75.				
76.	W9. Repository Name	Field #14: Current Repository Name	535a Location of Originals/ Duplicate Custodians	
77.	W10. Repository Place	Field #15: Current Repository Place	535bc location of Originals- Address and Country	
78.	W11. Repository Number	Field #6: Current Object ID Number	035 System Control Number	
79.	W19. Notes	Field #21: Notes	5XX General Notes	

Appendix B.

The Questionnaire and Practical Experiments.

Questionnaire on Web Resource Description.

Introduction:

This questionnaire was designed by Paul Catherall, and comprises part of an Information and Library Management MA dissertation entitled *Resource Description and Control on the World Wide Web.*

The purpose of the dissertation is to assess the viability of current and emerging standards for the description, indexing and retrieval of Web resources.

The popular term for data describing the content of Web resources, is *Metadata*.

Metadata can consist of a basic *keywords* tag in a HTML document, allowing some search engines to more efficiently index that Web page.
Other *metadata* models include many classes of descriptive data, such as *title*, *author*, *publisher* and *contributor*.

This questionnaire introduces you to several *metadata* standards and asks you to describe a Web page using simple metadata syntax; brief instructions and examples are provided for your reference.

It is hoped that in return for your time and effort, this questionnaire may provide an interesting introduction to the emerging concept and uses of *metadata*.

Paul Catherall.

MA Information & Library Management student.
John Moores University, Liverpool.

Email: e9501788@newi.ac.uk

Date: Participant No:

Experiment 1. Describing a web site using Dublin Core tags.

In this experiment, I would like you to describe a web page of your choice using simple Dublin Core META tags.
Each tag has two parts, META NAME and CONTENT (indicated here in bold).

The META NAME tells us what is being described, eg: a title, author or publisher.
The CONTENT contains the information for each specific tag.

The following tags will be used in this experiment:

Keywords: Words that describe the subject or content of the resource.

Description: A free text description of the resource.

The following example demonstrates how a web page is described using the above Dublin Core tags:

<META NAME="DC.Keywords" CONTENT="**Literature, Fiction, Music, Poetry**">
<META NAME="DC.Description" CONTENT="**This page is devoted to my favourite novels, music and poetry.**">

All you need to do to describe a web page of your choice is copy the demonstration tags, inserting your own CONTENT in the place of the examples (additional sheets are provided if required.)

Please type or write an entry for **Keywords**.

_____

_____

_____

Please type or write an entry for **Description**.

_____

_____

_____

_____

_____

Experiment 2. Describing a Web page in XML (Extensible Markup Language.)

In this section, I would like you to describe a Web Page of your choice using XML. XML allows us to create tags containing information about web pages.

An XML document must be enclosed by *root* start and end tags (indicated in bold.)

eg: **<Books>**
← (XML Content)
</Books>

Similarly, XML data items must be enclosed by start and end tags (called *element tags*, or *containers*.*)*

eg: <Favourite>Cannery Row</Favourite>

Please note that XML tags and the data they hold are all user-defined.
The following example demonstrates the use of XML tags:

<Homepage>

<Heading>About my amazing Web page.</Heading>

<Author>Paul Catherall</Author>

<Date>21/05/00</Date>

<Content>poetry, literature, language, culture and music</Content>

<Interests>programming, literature, badminton</Interests>

</Homepage>

Please choose key words or phrases to describe the Web page of your choice using XML tags, as shown above.

It is entirely up to you what you call your tags, (eg: *Heading, Author, Date*,) what information they contain, how many you use, and in what order you use them.

Please use the area below to type or write your XML. (Please use attached sheets if required.)

_____

_____

_____

_____

_____

_____

_____

Questions about you and your use of HTML.

1.) Please describe your occupation by selecting from the categories below.
(You may tick as many boxes as you like.)

Student of IT, Programming or Computer Systems.	☐	Information Management or Information Science related post.	☐
IT Technician or IT Assistant.	☐	IT Coordinator, Systems Manager, or other senior IT related post.	☐
Web Developer.	☐	Programmer or Systems Developer.	☐
Student of other subject (Please describe area of study.)		Other occupation (Please describe.)	
_____		_____	
_____		_____	

2.) How long have you been using HTML? Please indicate approximately in years and months in the boxes below.

Years ☐ **Months** ☐

Questions about your contact with Metadata.

3.) Please indicate which of the following *metadata* models you have heard of by ticking the corresponding boxes below.

HTML <META> tag	☐	RDF	☐
The Dublin Core	☐	XHTML	☐
XML	☐	None of these.	☐
		Others (Please specify below)	

_____

4.) If you have heard of Metadata before, or any of the above models, what did you think the purpose of metadata was?

If no, please go to question 5.

_____

_____

_____

5.) Please indicate which of the following *metadata* scripts you have used by ticking the corresponding boxes below.

HTML <META> tag	☐	RDF	☐
The Dublin Core	☐	XHTML	☐
XML	☐	None of these.	☐
		Others (Please specify)	

_____

6.) If you have used any metadata models before, please briefly describe the purpose of the metadata you used or created.

If no, please go to question 7.

_____

_____

_____

Questions about Experiment 1 (Dublin Core.)

7.) Generally, how comfortable were you creating Dublin Core data?
Please circle one of the following.

Very Comfortable 1 2 3 4 5 Least Comfortable

8.) How comfortable were you using the conventions and syntax of this script?
Please circle one of the following.

Very Comfortable 1 2 3 4 5 Least Comfortable

9.) How closely did this script follow HTML conventions and syntax?
Please circle one of the following.

Very Closely 1 2 3 4 5 Least Closely

10.) How useful were your knowledge of HTML conventions and syntax in understanding and using this script?
Please circle one of the following.

Very Useful 1 2 3 4 5 Least Useful

11.) How easily were you able to decide the kind of content for the *keywords* field?
eg: subject or interest areas, disciplines, place names or individual's names.
Please circle one of the following.

Very Easily 1 2 3 4 5 Least Easily

12.) How easily were you able to decide terms describing the page content in the *keywords* field?
Please circle one of the following.

Very Easily 1 2 3 4 5 Least Easily

13.) How easily were you able to decide the kind of information for inclusion in the *description* field?
eg: Page summary or précis, breakdown of page contents or subject area outline.
Please circle one of the following.

Very Easily 1 2 3 4 5 Least Easily

14.) Please describe any issues of interest or difficulties encountered during this experiment.
If no, please go to question 15.

_____

_____

_____

Questions about Experiment 2 (XML).

15.) Generally, how comfortable were you creating XML data?
Please circle one of the following.

Very Comfortable 1 2 3 4 5 Least Comfortable

16.) How comfortable were you using the conventions and syntax of this script?
Please circle one of the following.

Very Comfortable 1 2 3 4 5 Least Comfortable

17.) How closely did this script follow HTML conventions and syntax?
Please circle one of the following.

Very Closely 1 2 3 4 5 Least Closely

18.) How useful were your knowledge of HTML conventions and syntax in understanding and using this script?
Please circle one of the following.

Very Useful 1 2 3 4 5 Least Useful

19.) How easily were you able to decide the kind of content for XML containers?
eg: *Personal Interests and hobbies*: **<Interests> </Interests>**.
Please circle one of the following.

Very Easily 1 2 3 4 5 Least Easily

20.) How easily were you able to decide words or phrases to define XML containers?
eg: *Interests*: **<Interests> </Interests>**.
Please circle one of the following.

Very Easily 1 2 3 4 5 Least Easily

21.) How easily were you able to decide words, phrases or free text for inclusion in XML containers?
eg: <Interests>**Music, literature and art**</Interests>.
Please circle one of the following.

Very Easily 1 2 3 4 5 Least Easily

22.) Please describe any issues of interest or difficulties encountered during this experiment.
If no, please go to question 23.

_____

_____

_____

23.) Please include any further comments or opinions on metadata or any issues raised in this questionnaire below.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Thank you for your time and co-operation with this questionnaire.

Appendix C.

Example HTML Metadata Program.

Purpose and Application.

This program allows the user to append Dublin Core META tags into an existing document, or into a new document as a template for HTML coding/ editing.

The user may specify any ASCII text file format (e.g.: HTM, HTML, doc, txt…)

The program is available from myself on request, and is accompanied by a 'readme' file, which should be read first.

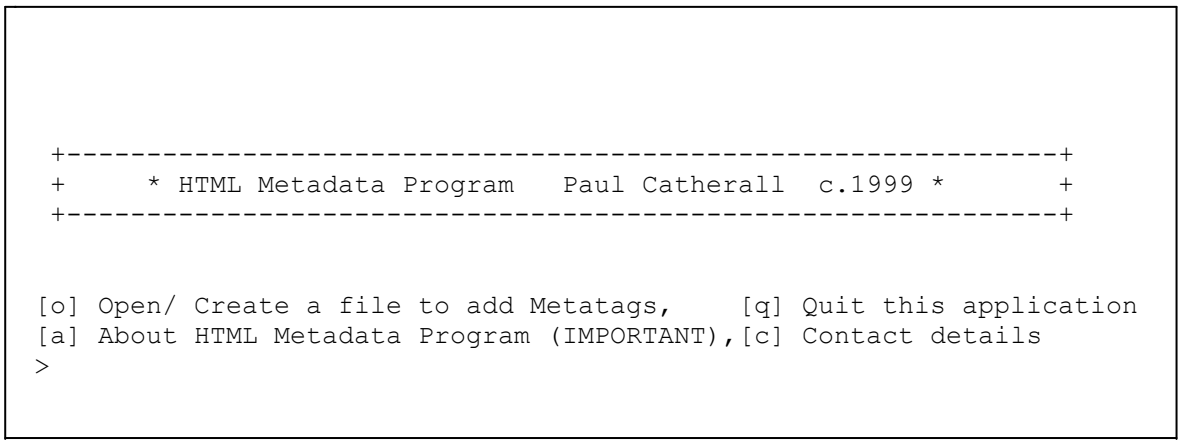


Figure 1. Opening Screen.

```
******************  About HTML Metadata Program  ************************

This program appends meta data into an HTML (htm, html etc.) or other text
file (eg: txt, doc.)   The meta-data tags into which values are converted
are based on the specifications of the Dublin core.
The program is simple, you are asked for the values corresponding to DC
meta variables, these are converted into HTML tags and inserted into the
head of your HTML document. (Overwriting any existing head data.)
Also, you can create a new metadata html template/ document from the menu.
If you try to open a file to append metadata and the specified file does
not exist, you can create a new document under that name.
When specifying files, you MUST include file extensions, eg: .htm, .txt...
At any time you may go back one stage, or cancel any modification.
The primary purpose of the program is to enable those inexperienced in
HTML to include metadata in their HTML documents, a function HTML editors,
such as Frontpage, do not contain.
Remember that once these values have been added, they cannot be edited in
your document using this application, to change meta data after using
this program you will need to use a text or HTML editor such as Notepad,
DOS Edit or Frontpage.

Type [q] to quit this help, or [d] to view DC meta specifications.
>
```

Figure 2. The Information Screen.

```
******************  Summary of Dublin Core Meta Terms *********************

Title        - name of the resource
Creator      - primary Author (organisation possible, some entity)
Subject      - keywords and phrases, should use a standard/ suggested
               vocabulary
Description  - may be abstract,  contents table, free-text account of
               content
Publisher    - entity like author/creator
Contributor  - a contributor
Date         - creation of the resource - yyyy-mm-dd
Type         - genre/ nature of content - DC recommend using  DC types
Format       - physical or digital definition - DC recommend MIME format
               types
Identifier   - a formal code identifier: URI, URL, DOI, ISBN etc.
Source       - a resource from which the present resource is derived
Language     - language of content - DC recommend  RFC 1766 codes
Relation     - a related resource
Coverage     - administrative / geographical location
Rights       - intellectual property and copyright declaration etc.

Type [e] to exit to menu, or [a] to return to help.
>
```

Figure 3. Dublin Core Metadata Specifications Screen.

```
*********************  Contact details  ***********************


     This program was written by Paul Catherall, 1/11/99.

I am always open to suggestions and comments. I can be contacted
at the following email:

                        gwledig@hobbiton.org

     I hope you find this program useful...

*************************************************************
```

Figure 4. Contact Details Screen.

```
 +-------------------------------------------------------------+
 +     * HTML Metadata Program   Paul Catherall  c.1999 *       +
 +-------------------------------------------------------------+


[o] Open/ Create a file to add Metatags,    [q] Quit this application
[a] About HTML Metadata Program (IMPORTANT),[c] Contact details
>o

Enter path and filename of html file to add metatags...

Note ++  Please use DOS directory names, no spaces, 8 characters only,
     ++  eg: C:\My network stuff\...  becomes  C:\Mynetw~1\...
     ++  Please include file extension, eg: txt, .htm, .html, .doc etc.

Example: C:\windows\internet\homepage.htm
Other commands [/q] =Quit to menu
>c:\windows\default.htm
Opening File...
```

Figure 5. User selects ‘Open/ Create file to add meta-tags’.

```
+-------------------------------------------------+
+                Metatag Edit Menu                +
+-------------------------------------------------+
[q] Quit to main menu, [a] Add metatags to your file
>a
```

Figure 6. Adding Metadata to the file.

```
Enter the ordinary (not metadata) 'title' of your page,
eg: Paul's page, The hill-climbing page...
You may already have given your page a title, if so please repeat it here...
[/q]= quit to menu, [Enter]= leave blank
>Tudalen Catref Gwledig

Now for the metadata...


Enter Title [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(name by which resource is known formally)
>Tudalen Catref Gwledig

Enter Creator [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(individual, organisation or service that created the resource)
>Paul Catherall

/q]= quit to menu, [/b]= back, [Enter]= leave blank
(key phrases, keywords (x, x,...), classification codes describing resource)
>Wales, Cymru, Welsh, Cymreig, Interests, Diddordebau, Culture, diwylliant


Enter Description [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(account of contents - eg: abstract, contents listing)
>A page devoted to Welsh culture, poetry, literature and language.


Enter Publisher [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(individual, organisation or service that made resource available)
>The Globe.com
```

Figure 7. User inputs values that will be written to file.

```
Enter Contributor [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(individual, organisation or service contributing to resource)
>

Enter Date [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(date resource created or made available in yyyy-mm--dd format.)
>1999-10-20

Enter Type [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(catagories, functions, genres description - should use controlled
vocabulary.)
>Interest, Popular Culture, Wales

Enter Format [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(media -mime- type or dimensions of resource, eg: size and duration)
>HTML / TEXT

Enter Identifier [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(eg: FTP, URL, URI, GOPHER or TELNET address or ISBN for books.)
>http:\\members.theglobe.com/gwledig

Enter Source [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(source from which present source is derived.)
>

Enter Language [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(language of content - should use RFC 1766 language/country codes)
>English, Welsh, uk

Enter Relation [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(a related resource, eg: URL, ISBN formal identification system.)
>http:\\members.theglobe,com/gwledig/anthem.htm

Enter Coverage [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(geographic coordinates/ location, time period/range, juristiction entity)
>Wales, Gogledd Cymru, North Wales

Enter Rights [/q]= quit to menu, [/b]= back, [Enter]= leave blank
(copyright/ intellectual property rights etc held by entity over resource)
>Copyright Paul Catherall, Assoc Penddraig Solutions.org 1999.
```

Figure 7. (Continued.)

```
Save these meta tags to your file?
[n]- quit to main menu, [y] - save meta tags to html.
>y
```

Figure 8. User prompted to save information to file.

```
c:\..\METAD.TXT
c:\..\OLDH.TXT
        1 file(s) copied
HTML updated and saved in original directory: c:\windows\default.htm
```

Figure 9. File is saved and User returned to main menu.

```
<HTML>
<HEAD>
<TITLE>Tudalen Catref Gwledig</TITLE>

<META NAME="DC.Title" CONTENT="Tudalen Catref Gwledig">
<META NAME="DC.Creator" CONTENT="Paul Catherall">
<META NAME="DC.Subject" CONTENT="Wales, Cymru, Welsh, Cymreig,
Interests, Diddordebau, Culture, diwylliant">
<META NAME="DC.Description" CONTENT="A page devoted to Welsh culture,
poetry, literature and language.">
<META NAME="DC.Publisher" CONTENT="The Globe.com">
<META NAME="DC.Contributor" CONTENT="">
<META NAME="DC.Date" CONTENT="1999-10-20">
<META NAME="DC.Type" CONTENT="Interest, Popular Culture, Wales">
<META NAME="DC.Format" CONTENT="HTML / TEXT">
<META NAME="DC.Identifier"
CONTENT="http:\\members.theglobe.com/gwledig">
<META NAME="DC.Source" CONTENT="">
<META NAME="DC.Language" CONTENT="English, Welsh, uk">
<META NAME="DC.Relation"
CONTENT="http:\\members.theglobe,com/gwledig/anthem.htm">
<META NAME="DC.Coverage" CONTENT="Wales, Gogledd Cymru, North Wales">
<META NAME="DC.Rights" CONTENT="Copyright Paul Catherall, Assoc.
Penddraig Solutions.org 1999">
</HEAD>
```

Rest of page HTML omitted…

Figure 10. The finished file with new Meta-tags embedded by the application.

Possibilities – A controlled Vocabulary.

The example program I have constructed is quite simple (it took about 1 ½ hours to write, and uses 2 simple subroutines.) This kind of application could easily be extended to support a controlled vocabulary, where defined parameters for each metadata class would be selected from established values stored in the code. Some meta-statements would have to remain open-ended, such as 'Creator' and possibly the 'Subject' keywords. Many meta tags could however conform to existing ISO standards, eg: options for country/ location could include abbreviations, such as 'fr' for France,'uk' for the UK.

The program is available by request from the following email address:

e9501788@newi.ac.uk

Appendix D.

XML-Based Schemas.

XML Schemas.

An XML schema is a standard format for expressing XML, unlike an XML DTD, which constitutes a template, or form for the compilation of XML documents according to standard XML conventions.

An XML schema usually contains formal syntax or tag content, of particular significance for interpretation only by parser systems supporting the schema. Whilst XML schemas may use agreed syntax or tag definitions, schema script still constitutes XML, and may be associated with DTDs in the same way that ordinary well-formed XML is used according to a DTD standard.

XML-Data

The schema for XML called XML-Data emerged following its suggestion by Microsoft, and other leading software developers.
XML-Data is mainly intended as a database standard, rather than for online use in resource description.

XML-Data is still not officially endorsed by the W3C, but has status as a W3C 'Note.' The main feature of XML-Data would be standard element definitions. The XML-Data Note is available online at the following address:

http://www.w3.org/TR/1998/NOTE-XML-data-0105/

Mathematical Markup Language Specification (MathML)

The Mathematical Markup Language Specification (MathML) was the first W3C approved XML *schema*. The schema supports mathematical notation, containing over 100 mathematical markup elements; MathML is intended for use in building mathematical and scientific authoring tools.

The MathML Specifications are available online at the following address:

http://www.w3.org/TR/REC-MathML

Appendix E

Additional Metadata Formats.

The REACH Element set.

The REACH elements (Record Export for Art and Cultural Heritage,) were developed in 1997 by the Research Libraries Group (RLG).

Like the CDWA and CIDOC guidelines, REACH is intended to provide a set of guidelines for the compilation of data elements in the conservation industry.

The REACH elements include a detailed range of data types describing artistic movements, the cultural background or category of artefacts and nationality or culture of the creator. REACH qualifiers include some standard content types.

The REACH elements include *Techniques/Process*, *Medium/Materials*, *Style/Period/Group/Movement/School* and *Nationality/Culture of Creator/Maker.* (The REACH Homepage. 2000.) See Appendix A for the REACH elements.

The VRA Core.

The Visual Resources Association (VRA) developed a standard element set in 1993, to manage complex visual collections in the arts and heritage industries.

The VRA Core is intended to serve as guidelines for describing works of art, architecture, and material from popular and folk culture.

The Core element set contains two groupings of elements, the *Work Description Categories* (19 elements), and the *Visual Document Description Categories* (9 elements) There are currently no formal qualifiers associated with the VRA standard.

The VRA Core includes elements such as *Record Type*, *Technique*, *Style/Period* and *Culture*.

Record Type = image
Type = black and white slide
Title = full view
Measurements.Format = 35 mm
Material.Support = LPD4 film
Creator = William Staffeld
Creator.Role = staff photographer
Date.Creation = ca. 1990
Location.Current Repository = Ithaca (NY, USA), Knight Visual Resources Collection, AAP, Cornell University
ID Number.Current Repository = B-J3 Nur 1.32 Cha 4b-2
Source = Geisberg, Max. "German Single-leaf Woodcut: 1500-1550." p. 787
Rights = © Geisberg, 1974

Figure 1. A VRA data file describing a black and white slide.

Appendix F.

Controlled Qualifiers and Vocabularies.

The Dublin Core Qualifiers.

There are 15 sets of qualifiers for use with the Dublin Core element set; these are now formal specifications approved by the Dublin Core Working Group.
The full qualifier specifications can be found at the following URL:

http://purl.org/metadata/dublin_core_elements

Several basic kinds of controlled terms are used to define languages and countries in standard Dublin Core content, using the ISO abbreviations for languages (ISO 639), the Country Code List (ISO 3166,) and the RFC MIME types (RFC2045).

Essentially, a Dublin Core qualifier comprises an extension to the standard Dublin Core element structure, providing an additional two features: the *scheme* and *value*. A *scheme* resembles a MARC sub-field, and defines the context of an element, whilst a *value* contains the content for the element, drawn from terms approved in the Dublin Core qualifier specifications.

When using Dublin Core qualifiers, some indexing systems, such as CORC repeat particular elements, adding different types of *scheme* for each instance of that element. An example of this in CORC is seen in the use of the SUBJECT tag, where this element is used to define both Library of Congress Subject Headings for the resource, and a valid Dewey Decimal Classification code:

<meta name="DC.Subject" scheme="DDC Local" content="025.344">
<meta name="DC.Subject" scheme="LCSH" content="Cataloguing of computer network resources.">

Although it is beyond the breadth of this project to discuss the possible *schemes* for each Dublin Core element (TITLE, SUBJECT, RIGHTS etc.) the following table illustrates approved qualifiers for the DESCRIPTION tag:

Element: Description (DESCRIPTION)	The Description element contains a textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources.
Scheme : Internal	The information provided is not part of an external coding system and the coding should be qualified by an accompanying TYPE from the list below. This is the default scheme value if the scheme is not explicitly stated.
Scheme : URL	The value of this element is a Uniform Resource Locator pointing at an external representation of the description of the resource.
Scheme : URN	The value of this element is a Uniform Resource Name identifying an external representation of the description of the resource.
Type: Freetext	A brief, freetext description of the resource provided by a third party (such as a cataloguer). This is a the default if a TYPE is not specified.
Type: Abstract	The abstract as provided by the creator/publisher of the resource.

Figure 1. The DC.DESCRIPTION Qualifiers. (Text from the Dublin Core Qualifiers Page, 2000 at: http://purl.org/metadata/dublin_core_elements.)

```
<meta name="DC.Title" content="Dublin Core Initiative">
<meta name="DC.Creator.CorporateName" content="Dublin Core Initiative">
<meta name="DC.Publisher" content="DCMI,">
<meta name="DC.Publisher.Place" content="[Dublin, OH] :">
<meta name="DC.Date.Issued" content="1998-9999">
<meta name="DC.Description" content="Title from home page.">
<meta name="DC.Identifier.URL" content="http://purl.org/dc">
<meta name="DC.Language" content="english">
<meta name="DC.Subject" scheme="LCC Local" content="Z699.24">
<meta name="DC.Subject" scheme="DDC Local" content="025.344">
<meta name="DC.Subject" scheme="LCSH" content="Dublin Core">
<meta name="DC.Subject" scheme="LCSH" content="Information storage and
retrieval systems &#183; Standards.">
<meta name="DC.Subject" scheme="LCSH" content="Metadata &#183; Standard.">
<meta name="DC.Subject" scheme="LCSH" content="Cataloging of computer
network resources.">
<meta name="DC.Relation.Requires" content="Mode of access: World Wide Web.">
```

Figure 2. Dublin Core Record from the CORC Project (2000,) illustrating the use of multiple elements using qualifiers.

ISO Standards.

The ISO standard defining language types (ISO 639) uses three letter abbreviations in lower case to define languages, and is widely used amongst many metadata standards, such as TEI, Dublin Core and CDWA.

The full ISO 639 specification scan be found at the following address:

http://adm5.byu.edu/Images/photoid/aacraoig/segments/dataelmt/nisolang.html

A large number of modern international languages are supported:

- ENG English
- WEL Welsh
- FRE French
- JPN Japanese

Broad classes of languages are also included:

- INE Indo-European (Other)
- GEM Germanic (Other)

Additionally, obscure, artificial and extinct languages are also supported:

- ESP Esperanto
- GRC Greek, Ancient (to 1453)
- NAH Aztec

Similarly, the Country Code list (ISO 3166) defines countries by a two-letter code, the specifications for this standard can be found at the following address:

http://adm5.byu.edu/Images/photoid/aacraoig/segments/dataelmt/country.html

Examples include obsolete country names, such as:

- SU Union of Soviet Socialist Republics (no longer exists)
- YD Democratic Yemen (no longer exists)

The list mainly supports modern European country names, including the following examples:

- AF Afghanistan
- BM Bermuda
- IL Israel

W3C Standards.

The W3C has reviewed suggestions to use the ISO 8601 standard for describing date and time information within digital resource description standards.

ISO 8601 describes many date/time formats, such as the definition for a year, month day or time, or any combination of these.
The full W3C date and time specifications may be found at the following address:

http://www.w3.org/TR/NOTE-datetime

The standard formats are as follows; Note: The ‘T’ in the string indicates the beginning of the time element, as specified in ISO 8601.

Year:
YYYY (eg 1997)

Year and month:
YYYY-MM (eg 1997-07)

Complete date:
YYYY-MM-DD (eg 1997-07-16)

Complete date plus hours and minutes:
YYYY-MM-DDThh:mmTZD (eg 1997-07-16T19:20+01:00)

The following example corresponds to November 5, 1994, 8:15am.

```
1994-11-05T08:15
```

These standards could be used within any metadata model, and would provide a valuable constant for displaying times, however they currently lack an indicator for time zones, such as GMT or SET; this addition would make these specifications a useful standard for resource description.

The Library of Congress Subject Headings.

The Library of Congress Subject Headings, (LCSH) are an internationally recognised standard for grouping kinds of literature into categories.
The system also supports a classification code (Library of Congress Call Numbers or LCCN,) indicated by a two letter code for each subject heading.

There are 26 categories of subject headings, corresponding to each letter of the alphabet, examples include the following:

A: General Works
B: Philosophy, Psychology, Religion N: Fine Arts
C: Auxiliary Sciences of History
D: History--General and Eastern Hemisphere
E & F: History--Western Hemisphere
G: Geography, Anthropology, Recreation
H: Social Sciences; Business

LCSH examples from section D, 'History - General and Western Hemisphere' include the following:

D History (General)
DA Great Britain
DB Austria, Czechoslovakia, Hungary
DC France
DD Germany
DE Mediterranean, Greco-Roman world
DF Greece

At present, the Dublin Core recommends the use of LCSH for inclusion as a *scheme* defined SUBJECT tag; this *scheme* is currently used by the CORC project.

A full list of LCSH categories can be found at the following address:

http://alice.library.ohiou.edu/screens/libinfo.33.html#M

Appendix G.

Excel Formulae Used In the Practical Research Element.

Excel Formulae.

The following Excel formulae were used to calculate equations for the practical research element:

(These are purely examples.)

=SUM(AM3:AM12) Note: Generates a total within a specified range.
=COUNTIF(AN3:AN12,"2") Note: Generates a frequency of instances of a given variable (in this case, '2'.)
=STDEVP(AQ3:AQ12) Note: Generates a standard deviation within a specified range.
=AVERAGE(AN3:AN12) Note: Generates a mean average within a specified range.
=MEDIAN(AN3:AN12) Note: Generates a median average within a specified range.
=MODE(AN3:AN12) Note: Generates a mode average within a specified range.
=(AN22/10) =(10-B17)/10 Note: Examples demonstrate uses of logical operators and parenthesis for sub routines in Excel formulae; these allow addition, subtraction, multiplication and division, allowing percentage, proportion and other calculations.