Contents

List of Figures

Chapter 3: Search Engine Resource Indexing and Control.

Chapter 4: Metadata: Definition and Overview

Chapter 6: Overview of Metadata Formats.

Chapter 7: Metadata as Markup.

Chapter 7: Metadata as Markup (continued).

Chapter 8: Metadata Initiatives.

Chapter 9: Metadata Tools and Client Software

Chapter 10: Questionnaire Results and Analysis.

Chapter 10: Questionnaire Results and Analysis. (continued).

Appendix A. Comparison Table of Metadata Format Elements.

Appendix C. Sample HTML Metadata Program.

Appendix E. Additional Metadata Formats.

Appendix F. Controlled Qualifiers and Vocabularies.

Glossary of Terms and Acronyms.

Occasionally, it is necessary to refer to some acronyms and technical terms.

It is not necessary to read these unless a term/ acronym definition is required,

however, this section does provide a useful introduction to metadata-related

terminology. These are a few frequently used throughout this project, with

accompanying explanations:

Applet.
Application written using Java code, resident with parent HTML files, and executable during HTML processing via a *hyperlink* tag. Applets may be used for a variety of functions in supporting and enhancing *browser* features, (including special effects, keyword search facilities and forms for user-input.)

Browser.
Client software used to receive data transmitted by *http* protocol, and interpret HTML script to display text, multimedia, *hyperlinks* and other features within the client window.

CGI (Script/ Protocol.)
Common Gateway Interface; a method used to activate applications or access data on a network, often through the use of *hyperlinks* in a HTML *browser* document. When the *hyperlink* is activated, a client-server communications script built into the HTML, such as *Perl*, sends instructions to the server via CGI protocol. A CGI script on the server is activated, and the appropriate client response is delivered as a result of the CGI request. Examples of CGI include: accessing a server database for data searching and retrieval from the client *browser*, or starting a networked application from an HTML *hyperlink.*

CLI.
Command Language Interface; text-based interface usually accompanied by a formal command language. CLI is largely necessary to communicate with the computer at the lower machine-code level. CLI-based systems are usually more efficient than *GUI* counterparts because of their greater proximity to machine code.

CORC.
Cooperative Online Resource Catalogue; 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*.

DC.
The Dublin Core; a set of 15 resource description elements, used to classify Web pages in HTML v.4. The Dublin Core resource elements use the standard HTML *META* tag. The following DC element describes a resource *creator*:

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

GUI.
Graphical User Interface; increasingly popular interactive graphical display format supported by many operating systems and software. GUI is usually characterised by the *WIMP* standard, (Windows, Icons, Menus, and Pointers.)

HTML
Hyper-text Mark-up Language; script used to format a Web-*browser* interpretable document, using *forms*, *hyperlinks* and other interactive features. On request by the client *browser*, the HTML code is transferred from a remote server to a local Web server using *http* protocol. The HTML formatting code is interpreted by the Web-*browser* software on the client PC, displaying a Web document as the result. HTML script is based on the concept of tag fields; each field defines a function, and each function must exist within the parameters of the field. The following tags are used to print text in the *browser* window:

This is a line of text!

HTTP
Hyper-text Transfer Protocol; the communications protocol used to transfer data via remote and local Web servers to a client PC for interpretation via Web *browser* software.

Hyperlink.
Abbreviation for *Hyper-text Link*; the hyperlink, encoded using HTML tags is a route to another Web page or resource on the Internet. Additionally, hyperlinks are also used to activate and access networked applications, databases and other digital resources. The following hyperlink allows the user to open and run a *MIDI* (Musical Instrument Digital Interface) music file by clicking on the words *the anthem* on the Web page:

the anthem...

Information Provider.
An online index of printed, online, multimedia or other resource types, providing facilities for resource discovery, such as keyword searching and subject-specific indexes. Information Providers may focus on particular issues or subjects, or may simply provide access to resources on the Internet as a whole. Additionally, Information Providers fall into one of several classes, including *Information Gateways*, or *Portals* (typically providing manually-compiled indexes on a range of subjects, and *Search Engines*, (using a variety of indexing methods to provide access to resources across the Internet as a whole.)

Internet.
A huge global network of independent networks, using *TCP/IP* protocols and a variety of communications media to transfer, exhibit and exchange information and data.

ISP.
Internet Service Provider, the large routing systems used to provide access to the Internet and World Wide Web.

Java Script.
A formatting script supported by HTML v.4, based on the *Java* programming language, intended to enlarge and enhance the functionality of HTML. Although less powerful than the *Java applet* model, Java Script allows for enhanced HTML functions, such as dynamic *variable* handling and event-trapping (an event triggering a response.)

Markup Document.
Any digital document containing content expressed using a formal script language; example *markup languages* include *XML* and *HTML*.

Markup Language.
Formal syntax and conventions used to express textual content; markup languages are used to define how textual content is presented, and to define classes of data within a document.

Metadata.
Data describing the content of a document, application, data file, multimedia file or other resource.

META tag.
A markup structure available in HTML v. 4. The most common use of this tag is for storing keywords in the *HEAD* of an HTML document. Some *Search Engines*, such as Alta Vista and Yahoo use this tag in indexing and ranking Web documents:

> <META NAME= “KEYWORDS” CONTENT=“Poetry, Welsh Poetry, Barddoniaeth Cymreig, Barddoniaeth.”>

MIME.
Multipurpose Internet Mail Extensions; the agreed standard for defining data types on the Internet, and their accompanying file extension names, e.g.: .txt for a text file, .bmp for a *Bitmap* picture, or .mid for a *MIDI* music file (Musical Instrument Digital Interface).

OCLC.
The Online Computer Library Centre, an international organisation founded and based at Ohio State University. OCLC facilitates a number of public online cataloguing systems, and its members participate widely in the research and development of information industry standards.

OSI.
Open Systems Interconnection; term often referred to in describing standardisation of communications protocols and systems for data transfer. The client-server model, running on *TCP/IP* protocols is an example of Open Systems architecture.

RDBMS.
Relational Database Management System; these databases are designed for interoperability between various different applications and statistical utilities. They often use *SQL* or *Oracle* standard formatting, allowing for system-level and cross-application data interchange.

RDF.
Resource Description Framework, a formatting standard, or schema for the *XML markup language*; RDF is characterised by *nodes* and *values*, these allow for user-defined tag classes, and dynamic *variable* handling within those tags. The following script uses the *description*, *title* and *creator* nodes to describe the author of a Web page at a specific address:

```
<RDF:RDF>
  <RDF:Description RDF:HREF="http://www.users.theglobe/gwledig">
    <Title>My Web Page.</Title>
  </RDF: Description >
    <Creator >Paul Catherall.</Creator>
</RDF:RDF>
```

Schema (external.)
Function used to associate a *markup document* with an external system for validation of some kind; HTML-supported scripts, such as *XML* may be declared for *browser* processing using the script *prolog* tag (<? ?>), but if the script is structured according to a resource description schema, such as *RDF*, then validation by an external interpreter may be required. When a cataloguing or indexing system, such as *CORC* interprets an HTML document, the schema function allows this system to recognise the presence of a compatible script (such as OCLC MARC,) and interpret it accordingly. The following *schema* tag is used to direct the *browser* to an online parser for script validation:

```
<?xml:namespace ns = "http://www.w3.org/RDF/RDF/" prefix = "RDF" ?>
```

Search Engine.
Software/ server technology used to index HTML documents online, and provide a user interface for resource retrieval.

Server.
Powerful, high storage capacity computer used to store networked software, and regulate communications traffic between client workstations, server resources and the Internet. This machine is also used to access the Internet via an *ISP*.

SGML.
Standard Generalized Markup Language, an *ISO* standard script (International Organisation for Standardisation), used to define document structures and allowable content. HTML and *XML* are both based on this early *markup language*.

SQL.
Structured Query language; the standard formatting script used in some database systems. SQL allows integration and interchange between databases of different format, architecture and purpose.

String.
Definition of a sequence of text characters.

TCP/IP.
Transmission Control Protocol and Internet Protocol; the *Transmission Control Protocol* breaks data into small *packets* of no more than 1.5 MB (MB=*Megabyte*, 1MB=1000 *Kilobytes*); as it generates each data *packet*, a *checksum* of the precise *packet* byte size is stored within that *packet*. The *Internet Protocol* refers to the specific address of a client PC to which data is being delivered. The TCP *packets*, with data and *checksum* are then stored within an IP data *packet*, containing the *IP* address of the destination client machine. When the combined TCP/IP packet has been dispatched from the sender machine, it is sent via server routers to the correct receiver. *Packets* are then recombined until the original checksum has been restored.

UKOLN.
The UK Office for Library and Information Networking; UKOLN is a Bath based organisation, facilitating the research and development of network information management systems and standards. UKOLN also contributes to current information systems research, and awareness services for the library and information community.

UNIX.
Network *Operating System* used to run and maintain the Internet. UNIX is a widely functional system, useable as a network server, communications interface and as a platform for network software development.

URL.
Universal Resource Locator, used to 'point' a *browser* query to the server/ domain home of a World Wide Web resource. The URL also contains the protocol needed to retrieve the resource, eg: *http:*\\, *telnet:*\\, *ftp:*\\.

URN.
Universal Resource Name, a *URL* linking a user query to a stable database containing *URLs*. The use of URNs was developed by the *PURL* project (Persistent Uniform Resource Locations.)

Variable.
A referable static value containing dynamic or/ and compilation-level defined data. As an example, the integer variable *number%* could contain a value defined in the application code itself, eg: *1*, but might also change according to user-defined actions; e.g.: user presses the '+' key, incrementing the variable by 1, so that the variable *number%* now contains the value *2*.

Web Crawler.
Autonomous detection/ retrieval systems (or 'bots,') used to discover HTML resources on the World Wide Web, and store HTML document content for *Search Engine* indexing. Autonomous detection/ retrieval systems also interrogate the HTML content of discovered resources for *hyperlinks* in the script; these *hyperlinks* are then the subject of further resource retrieval and indexing processes.

World Wide Web, or WWW.
The network of Web compliant servers existing across the Internet, using the *http* protocol and HTML *browser* software to publish and access HTML-based Web pages. The World Wide Web is the most popular entity on the Internet, mainly due to the highly accessible nature of its interactive graphical interface.

W3C.
The World Wide Web Consortium, an organisation founded in 1994 to develop information standards and models for the evolution of the World Wide Web. W3C is an international association of industrial and service companies, research laboratories, educational institutions, and information services.

XML.
Extensible Markup Language; metadata script allowing for the definition of class types and their *variable* content. XML is characterised by *root elements*, *child elements* and *attributes*, which allow for defining hierarchical data structures. XML may be used within an HTML document for embedded resource description, or may be used independently as a fully functional formatting script. XML may also be used to display data as an interactive hierarchical structure via the Web *browser*. *RDF* is an emerging standard for XML script formatting. The following XML script demonstrates the potential to describe a book:

```
<PC_Book >

  Favourite Book.

  <Book Title="Cannary Row">
    <Author>Steinbeck, J</Author>
    <Publisher>Penguin</Publisher>
  </Book>

</PC_Book >
```

Z39.50.
Applications Layer Protocol used to transfer data between database systems on the internet. The standardisation of Z39.50 has provided an easier means of communication between *RDBMS* and proprietary applications than existed previously under integrated network *operating systems*, such as *UNIX*.