Automatic RDF Metadata Generation for Resource Discovery

Charlotte Jenkins, Mike Jackson, Peter Burden, Jon Wallis
School of Computing & IT
University of Wolverhampton
Wulfruna Street, Wolverhampton
WV1 1SB, UK

Abstract

Automatic metadata generation may provide a solution to the problem of inconsistent, unreliable metadata describing resources on the Web. The Resource Description Framework (RDF [10]) provides a domain-neutral foundation on which extensible element sets can be defined and expressed in a standard notation. This paper describes how an automatic classifier, that classifies HTML documents according to Dewey Decimal Classification (DDC [8]), can be used to extract context sensitive metadata which is then represented using RDF. The process of automatic classification is described and an appropriate metadata element set is identified comprising those elements that can be extracted during classification. An RDF data model and an RDF schema are defined representing the element set and the classifier is configured to output the elements in RDF syntax according to the defined schema.

Keywords: RDF, metadata, classification.

---

1. Introduction

A major problem facing tools for information resource discovery on the Web is the lack of a mechanism for resource description within the Web's architecture. There are now said to be in excess of 320 million individually accessible objects on the Web [5]. There is no one accurate, reliable, up-to-date, comprehensive method of finding out what each one of these objects is, what type of resource it is, what the subject matter is and so on, without accessing and analysing each one individually. This is a problem, not only for resource discovery, but also for content rating where illicit material is concerned. The World Wide Web Consortium (W3C [13]) has introduced the Resource Description Framework (RDF), in an attempt to produce a standard language for *machine-understandable* descriptions of resources on the Web. RDF is intended to support resource descriptions for resource discovery and also for rights management, privacy preferences, content ratings (PICS [9]), evaluation and classification. RDF is seen as the framework for producing a *Web of trust* where the content of each individually accessible object is well described in a format that is extensible yet universally understood. RDF may enable search engines and other tools for resource discovery to exchange and share metadata. This paper is concerned with the automatic generation of metadata in RDF format for use in describing HTML documents for the purposes of resource discovery.

Various attempts have been made to introduce embedded metadata into HTML documents, most notably through the use of the HTML META tag and embedded Dublin Core [12]. It is also now possible to include an embedded RDF description of a document. The problem with such techniques is that they are not compulsory so many authors still choose not to include meta information. M. Marchiori, in his paper entitled *The Limits of Web Metadata and Beyond* [7], addresses this issue by proposing a scheme that involves *back-propagating* meta information from pages with known metadata to those that are linked from it. An alternative method of automatically generating metadata is to use an automatic classifier. The automatic classifier described in this paper works by comparing terms found within documents with manually defined clusters of terms representing the nodes of a classification hierarchy (DDC). This process results in the identification of other useful metadata such as the document title, keywords, abstract and word count in addition to the classification classmarks. An RDF schema has been defined for representing this metadata and the process by which it is extracted and represented in RDF is described.

---

2. Automatic Classification

The automatic classifier [3] described below has been designed and developed for use as an automated component of a distributed automated search engine. The use of automatic classification within an automated search engine is quite unusual - commonly automated search engines (such as AltaVista) are huge indexes and classified tools (such as Yahoo and Galaxy) require some degree of manual intervention, typically in specifying the classification category and other such meta information. It has been observed [6] that, classified tools, although often hopelessly incomplete and out-of-date because of the lack of automation, are less likely to inundate users with irrelevant information. Automatic resource discovery combined with automatic full text indexing is faster and more comprehensive than manual classification but much less accurate. It is hoped that the use of automatic classification will combine the advantages of both approaches resulting in an accurate, comprehensive, up-to-date, well classified, automated search engine. Documents sharing the same subject matter will be automatically clustered together under the same classification classmarks and therefore will be retrieved together more easily.

The automatic classifier classifies documents according to DDC. DDC is considered appropriate because it is a universal classification scheme covering all subject areas and geographically global information. It is familiar to anyone accustomed to using a library and has multilingual scope. The hierarchical nature enables the users of a search engine to refine their search from rough classifications to increasingly more accurate ones.

The automatic classifier is an object oriented system, written in Java, that retrieves HTML documents from given URLs, analyses the contents and assigns appropriate DDC classmarks. A hierarchy of Java classes is used to model the DDC classification hierarchy. Documents are filtered through this hierarchy according to which *class representatives* (manually defined terms representing each DDC class) best match the document's contents. Each term found within the document is given an associated weight which is greater if the term is found in the title or a heading element. Terms found within the *keywords* or *description* elements of existing META tags are also stored with significant associated weight. Terms also acquire more weight the more often they occur. These weighted terms are then compared with the manually defined terms representing DDC classes. Initially the document is compared with the top ten DDC classes shown in Figure 1.

000 Generalities
100 Philosophy, paranormal phenomena, psychology
200 Religion
300 Social sciences
400 Language
500 Natural sciences and mathematics
600 Technology (Applied sciences)
700 The arts, Fine and decorative arts
800 Literature (Belles-lettres) and rhetoric
900 Geography, history, and auxiliary disciplines

Figure 1. The ten DDC classes

If a significant match is found between the document and a DDC class, the document is then compared with subclasses of that DDC class. This comparison process continues recursively through the hierarchy until significant matches with leaf nodes are found, the classmarks of which are assigned to the document.

To illustrate this process more clearly, figure 2 shows a document object which comprises:

- an accession number that is used to uniquely identify the document
- a series of keyword objects, each one representing a term found within the document, with an associated weight depending on where it was found within the document and how frequently it occurs (note, ALL found terms are stored in this manner)
- a series of classmark objects, each one comprising the actual classmark together with a textual label e.g. *303.483 Development of science and technology*. Appropriate classmark objects are assigned here when the keywords match significantly with the keywords of DDC objects that have no subclasses (see figure 3) i.e. leaf nodes in the hierarchy.

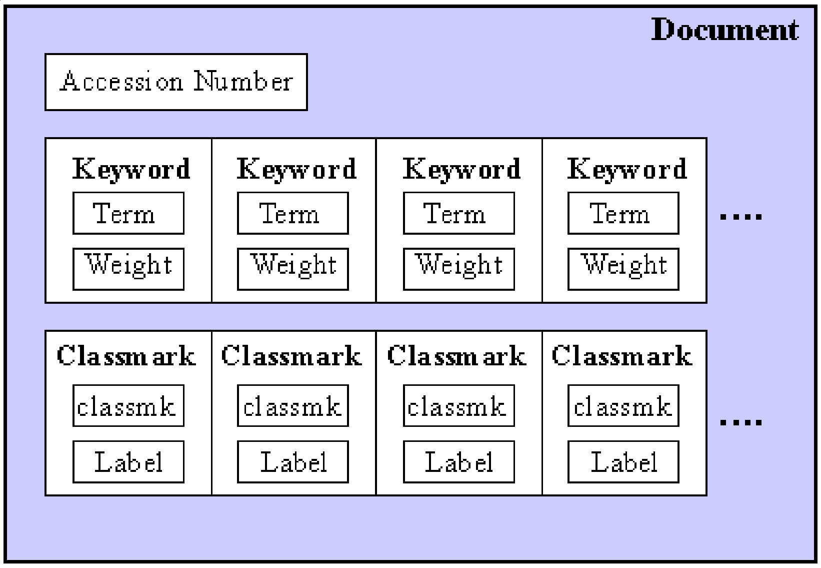


Figure 2. A document object comprising a series of keyword objects and a series of classmark objects

Figure 3 shows a DDC object which comprises:

- a series of keyword objects, identical in structure to the document keywords but comprising manually defined terms representing this particular DDC class
- a series of subclasses - each of which is itself a DDC class representing the next layer of the hierarchy beneath this class. Leaf nodes obviously have no subclasses
- a classmark object defining and uniquely identifying this class. If the keywords of this class match significantly with the keywords of a document object and there are no subclasses, this classmark object is assigned to the document.

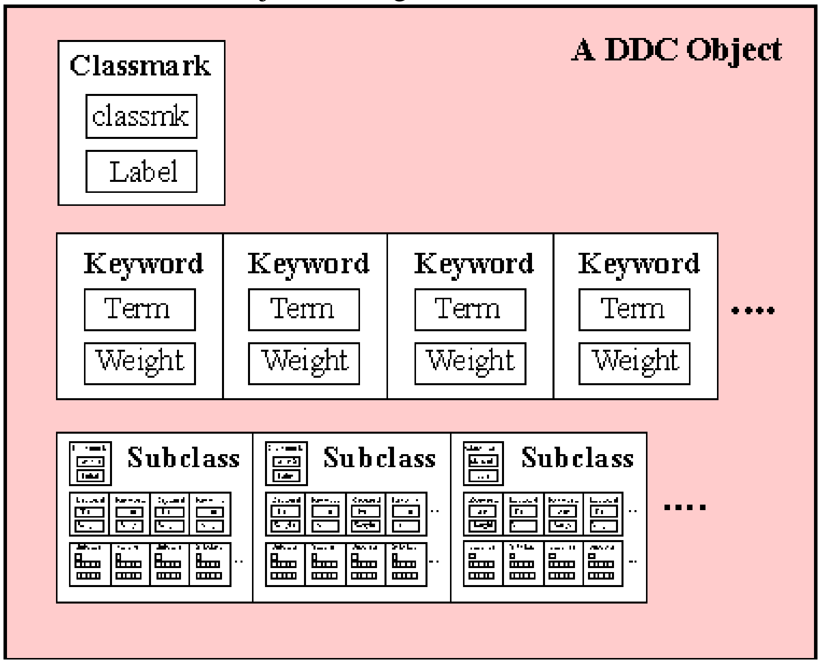


Figure 3. A DDC object comprising a series of keyword objects, a series of DDC objects representing its subclasses and a classmark object

The hierarchical nature of the DDC classes enables ambiguous terms to be concealed and considered in context lower down the class hierarchy. The class representatives at the top of the hierarchy contain a broader range of terms than those nearer the bottom which are more detailed and more specific.

Measures of similarity between the document and DDC *class representatives* are calculated using the Dice Coefficient [11]:

$$2 \frac{X \cap Y}{X \cup Y}$$

Figure 4. The Dice Coefficient

Each time a word in the document matches a word in the DDC class representative, the two associated weights are added to a total score (x intersection y). This score is then divided by the sum of the number of keywords in the document and the number of keywords in the class representative (x union y) and the result is multiplied by 2. Any result greater than 0.5 is considered significant and the document will proceed to be compared with any subclasses or be assigned the classmark if there are no subclasses. If the score is not significant, the comparison process will proceed no further through this branch of the hierarchy.

The comparison process may proceed through several unrelated branches of the hierarchy for as long as significant matches are found. In a Web library multiple classifications are appropriate - the same book can appear on several different shelves.

2.1 Metadata Elements

The classification process results in the production of a series of classmarks appropriate to describe a particular document. However, the process can easily be used to pull out various other metadata elements. During the parsing of the document, terms found in the title element are singled out as being important, these can easily be extracted as can those terms which match those found in the class representatives of appropriate DDC leaf nodes i.e. significant keywords. Keywords and descriptions found in existing META tags can be extracted. Other useful metadata that is easily accessible is shown in Figure 5. This element set is based on those metadata elements that can easily be obtained from the process of automatic classification. These elements are particularly suited to the domain of the automated search engine.

	Element	Description	Purpose
1.	Unique accession number	Number assigned by the system.	Uniquely identifies the resource.
2.	Title	Taken from the HTML <TITLE> element.	Usually helps in discerning the subject matter.
3.	URL*	The URL given to the system, used to extract the document for classification.	Indicates the location of the document.
4.	Abstract	Either the first 25 words found in the body of the page, or, if present, taken from the *Description* META tag. (A much more sophisticated abstracting technique could be used here in future implementations).	Provides further clues about the subject matter.
5.	Keywords	Terms found within the document that match terms found within the class representatives of DDC classes found to be appropriate.	Indicate key issues/topics.
6.	Classmarks	DDC classmarks found to be appropriate as a consequence of the classification process.	Indicate subject area(s).
7.	Word count	The number of words found on the page, including the title.	Indicates extent, detail, download time.
8.	Classification date	The system date when the classification took place (GMT or BST)	Indicates currency of the metadata.
9.	Last modified date when classified	Taken from the HTTP Last-modified header. (Gives *Not known* if equal to the "epoch" - 1st January 1970)	Indicates currency of the information.

* The classifier only handles individual HTML documents so the URL, not URI, is appropriate. The URL is not used as an identifier within the search engine because it is possible for the same page to have more than one URL; this is one of the causes of repetitions in automated search engine results.

Figure 5. An appropriate metadata element set - The " Wolverhampton Core "

It is thought that these elements (Wolverhampton Core) are sufficient to uniquely identify the document, state where it can be found, provide a good indication of the subject matter and of how current both the actual information and its metadata are.

The most well known and well used metadata element set for resource discovery is Dublin Core [12]. Compliance with a recognised standard is advisable because it encourages interoperability and consistency between applications. Dublin Core has evolved from the Digital Library community and consequently not all of its elements are as well suited to the automated search engine domain as those defined in figure 5. There is, however a significant overlap and none of the Dublin Core elements are compulsory. RDF enables developers to tailor an element set to suit their application while still reusing appropriate standard elements defined elsewhere (see section 3).

Figure 6 compares the fifteen elements of Dublin Core with the elements defined in figure 5.

	Dublin Core Elements	**Equivalent Wolverhampton Core Elements**
1.	Title	Title
2.	Creator	-
3.	Subject	Keywords + Classmarks
4.	Description	Abstract
5.	Publisher	-
6.	Contributor	-
7.	Date	Last modified when classified
8.	Type	-
9.	Format	-
10.	Identifier	Accession number + URL
11.	Source	-
12.	Language	-
13.	Relation	-
14.	Coverage	-
15.	Rights	-
16.	-	Date Classified
17.	-	Word count

Figure 6. Comparison between Dublin Core and the Wolverhampton Core element sets.

It can be observed that most of the *Wolverhampton Core* elements have a Dublin Core equivalent. The implications of this comparison are discussed again in the later section (3.2) on RDF schema definition. It is thought that the specified Wolverhampton Core elements represent an appropriate subset of Dublin Core (with one or two additions) that is suited to the requirements of an automated search engine.

Once the necessary metadata elements have been identified they can then be represented in RDF.

---

3. Resource Description Framework (RDF)

Three things are required in order to generate RDF statements about a resource: a data model, a schema and the actual representation in XML (eXtensible Markup Language [2]) syntax. Several RDF schemas might actually be involved; schemas are required for the interpretation of RDF statements. The following three subsections explain how the metadata elements shown in figure 5 can be represented by an RDF data model, defined using an RDF schema and, most importantly, automatically generated in RDF/XML syntax.

3.1 RDF Data Model

The RDF data model is expressed using directed labelled graphs (or "nodes and arcs" diagrams) which identify the properties and property values associated with a resource as shown in figure 7. (This notation is taken from the RDF Model and Syntax Specification [4]).

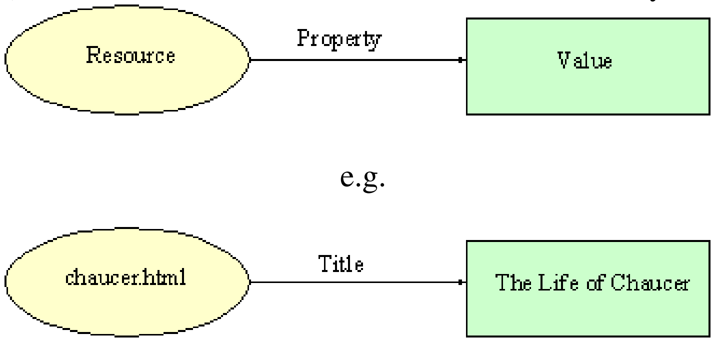


Figure 7. Data Model notation showing an RDF statement; a resource, a named property and the value of that property.

In RDF a resource may be a simple Web page, part of a simple Web page, a collection of pages or a whole Web site. The automatic metadata generator described in this paper generates descriptions of individual HTML pages.

Figure 8 shows how the element set in Figure 5 would be represented for the HTML page at http://www.scit.wlv.ac.uk/index.html

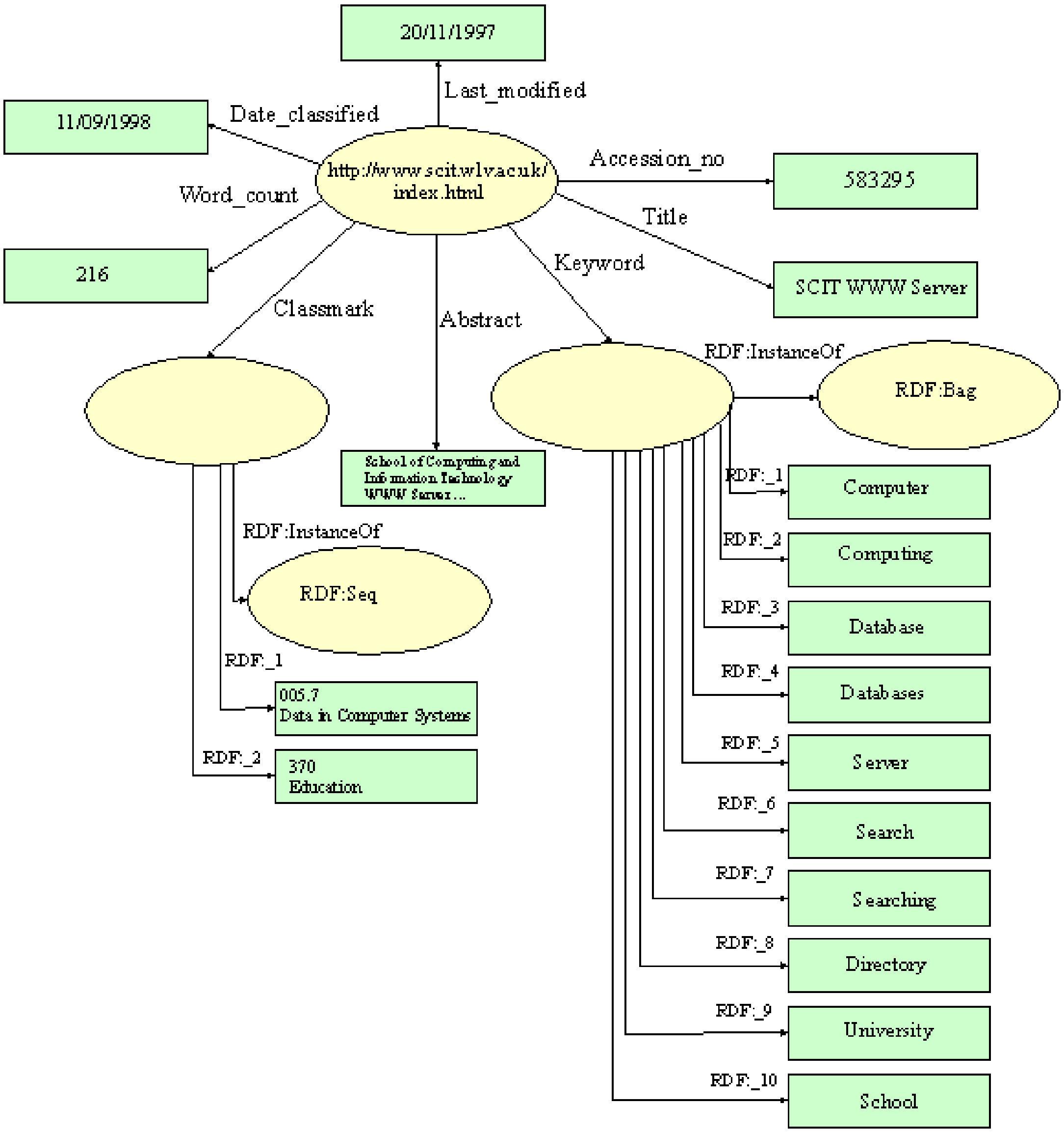


Figure 8. An RDF Data Model using the elements from Figure 5.

The model shows two RDF containers - one a bag of keywords and the other a sequence of classmarks. The classification process will usually result in the identification of several keywords within the document but the order in which they are presented is insignificant so a bag is appropriate. A better method of representing the keywords would be to use a *Set* where no duplicates would be permitted, however, RDF does not define a *Set* because there is no defined enforcement mechanism in the event of violation. The classmarks are ordered by the classifier according to which scored the highest measure of similarity and so these are represented as an ordered sequence. The classmarks would be better represented by an ordered collection class where no duplicates would be allowed. Further work layered on the RDF core may define such enforcement mechanisms.

3.2 RDF Schema Definition

Once the appropriate properties have been identified a schema must be created, or existing schemas must be identified, where these properties are defined. Schemas provide the RDF type system and enable applications to interpret RDF statements (see next subsection). The properties could be expressed using appropriate existing vocabularies, in which case it is not necessary to define a new schema - existing schemas can be referenced from within the RDF/XML syntax. It is possible to reference as many different schemas as required, mixing and merging different vocabularies. Schemas are referenced using the namespace mechanism from within the RDF syntax (see next subsection).

The definition of new schemas enables developers to specify properties best suited to their particular application. Schemas can define properties that are sub-properties of those defined elsewhere in existing schemas. This feature has been utilised in the Wolverhampton Core schema definition shown in appendix A. A property has been defined for each element identified in figure 5. Those elements that Wolverhampton Core has in common with Dublin Core (as shown in figure 6) have been defined as sub-properties of the appropriate Dublin Core properties. (The Dublin Core properties are based on those shown in the example Dublin Core schema in the RDF Schema Specification [1]. Note, this is not the authoritative Dublin Core schema which will be made available by the Dublin Core Initiative). This approach has been adopted so that Wolverhampton Core properties have *specialisation relationships* with Dublin Core properties and retain some implementation freedom. It would have been possible to use the Dublin Core properties directly in the automatically generated RDF syntax (see next subsection) but it is very important that the automated search engine is clear about the particular implementation of these properties. For example, both keywords and classmarks could be expressed as Dublin Core *Subject* properties (see figure 6) but the search engine needs to be able to differentiate between keywords and classmarks. Two Wolverhampton Core properties are defined, representing the keywords and classmarks independently, both of which are defined as sub-properties of the Dublin Core *Subject* property. Creating specialisation relationships in this manner will enable applications capable of processing Dublin Core to, at least partially, interpret Wolverhampton Core thereby encouraging both interoperability and extensibility.

The schema for defining the Wolverhampton Core element set can be found in Appendix A. (The schema specification language in which this schema is written is defined in the RDF Schema Specification [1]). This is a very simple definition of the properties required to represent the elements identified in figure 5. Future implementations could define new classes and declare constraints on the properties.

3.3 RDF Syntax

The following shows the RDF representation of the data model shown in figure 8. Appendix B shows automatically generated RDF for a series of test URLs. (The RDF/XML syntax used here is described in The RDF Model and Syntax Specification [4].)

```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
 xmlns:wc="http://scit.wlv.ac.uk/~ex1253/wc/schema/">
     <rdf:Description about="http://www.scit.wlv.ac.uk/">
          <wc:Accession_no>583295</wc:Accession_no>
```

```
        &lt;wc:Title&gt;SCIT WWW Server&lt;/wc:Title&gt;
        &lt;wc:Abstract&gt;
             School of Computing and Information Technology
             WWW server General Information University of
             Wolverhampton School of Computing and Information
             Technology home page Wolverhampton and surrounding
             areas
        &lt;/wc:Abstract&gt;
        &lt;wc:Keyword&gt;
             &lt;rdf:Bag&gt;
                  &lt;rdf:li&gt;computer&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;computing&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;database&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;databases&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;server&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;search&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;searching&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;directory&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;university&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;school&lt;/rdf:li&gt;
             &lt;/rdf:Bag&gt;
        &lt;/wc:Keyword&gt;
        &lt;wc:Classmark&gt;
             &lt;rdf:Bag&gt;
                  &lt;rdf:li&gt;005.7 Data in computer systems&lt;/rdf:li&gt;
                  &lt;rdf:li&gt;370   Education&lt;/rdf:li&gt;
             &lt;/rdf:Bag&gt;
        &lt;/wc:Classmark&gt;
        &lt;wc:Word_count&gt;216&lt;/wc:Word_count&gt;
        &lt;wc:Last_modified&gt;20/11/1997&lt;/wc:Last_modified&gt;
        &lt;wc:Classification_date&gt;11/09/1998&lt;/wc:Classification_date&gt;
     &lt;/rdf:Description&gt;
&lt;/rdf:RDF&gt;
```

Note that there are two XML namespace definitions (xmlns) at the top of this piece of RDF. The first one identifies the location of the RDF syntax specification and the second one identifies the location of the Wolverhampton Core (wc) schema where the property types specified within this RDF description are defined. This wc schema is shown in Appendix A.

W3C and the Dublin Core Initiative recommend the use of ISO 8601 Date format. This has not been implemented in this instance because the automatic metadata generator is to be deployed as part of a UK search engine where dates will be required in UK format.

If the classification process should fail, i.e. no significant measures of similarity are found, other elements such as the title, abstract, word count and dates should still be identified.

---

4. Conclusions

Although it is envisaged that the editing tools of the future will encourage the inclusion of RDF meta information, the current situation, where some authors choose not to include any metadata, is likely to continue to some extent. It is very difficult to automate resource description but it would be impossible to describe everything on the Web manually. Automatic metadata generation would appear to be an essential pre-requisite for widespread deployment of RDF based applications. The *Web of trust* must attempt to be comprehensive because a Web that is partially trust worthy offers little advantage over one that cannot be trusted at all, especially where content rating is concerned.

The automatic metadata generator described in this paper enables an RDF description to be associated with any HTML page, regardless of when it was created and by which editing tool. RDF has enabled the specification of a metadata element set that is tailored to suit an automated search engine but strongly related to a standard, digital library element set, Dublin Core. The ability to create specialisation relationships with appropriate Dublin Core properties increases the potential for interoperability - any application capable of processing Dublin Core will be capable of processing most of the defined Wolverhampton Core properties because they are defined as sub-properties of Dublin Core properties. Such interoperability will encourage information sharing which will improve comprehensive Web coverage; if search engines can process the same standard syntax, they will be able to exchange metadata and integrate their results. Some subject-specific classified directories are known to be attempting to share information through the use of RDF already; information sharing between automated search engines has even greater potential.

References

1. D. Brickley, R. Guha, A. Layman, *Resource Description Framework (RDF) Schema Specification*, http://www.w3.org/TR/WD-rdf-schema (Working Draft - October 1998)
2. D. Connolly, J. Bosak, *Extensible Markup Language (XML)*, http://www.w3.org/XML/, (October 1998)
3. C. Jenkins, M. Jackson, P. Burden, J. Wallis, *Automatic Classification of Web Resources using Java and Dewey Decimal Classification*, Computer Networks and ISDN Systems, Volume 30 646-648 (1998)
4. O. Lassila, R. Swick, *Resource Description Framework (RDF) Model and Syntax Specification*, http://www.w3.org/TR/WD-rdf-syntax (Working Draft - October 1998)
5. S. Lawrence, C. L. Giles, *Searching the World Wide Web*, SCIENCE, Volume 280 (April 1998)
6. L. Lindop, M. Sriskandarajah, M. Williams, M. Bracken, M. Cadden, A. Dabbs, W. Gallagher, *Catching Sites*, PC Magazine 6 (2) (February 1997)
7. M. Marchiori, *The Limits of Web metadata and beyond*, Computer Networks and ISDN Systems, Volume 30 1-9, (1998)
8. OCLC Forest Press, *Dewey Decimal System Home Page*, http://www.oclc.org/oclc/fp/index.htm (October 1998)
9. P. Resnick, *Platform for Internet Content Selection (PICS)*, http://www.w3.org/PICS/ (January 1998)
10. R. Swick, E. Miller, B. Schloss, D. Singer, *Resource Description Framework (RDF)*, http://www.w3.org/RDF/ (October 1998)
11. R. C. J. van Rijsbergen, *Information Retrieval: Second Edition*, Chapter 3, http://www.dcs.glasgow.ac.uk/Keith/Chapter.3/Ch.3.html Butterworths, ISBN 0-408-10775-8, 1981
12. S. Weibel, E. Miller, *Dublin Core Metadata*, http://purl.oclc.org/metadata/dublin_core/ (November 1998)

13. The World Wide Web Consortium, http://www.w3.org (October 1998)

Appendix A

Below is the RDF schema for the Wolverhampton Core (wc) element set referred to in figures 5 and 8. Note that the URL is not specified as a separate property because it is always noted in the `<rdf:Description about="http://...">` statement.

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
   xmlns:rdfs="http://www.w3.org/TR/WD-rdf-schema#">

<rdf:Description ID="Accession_no">
  <rdf:type rdf:resource="http://www.w3.org/TR/WD-rdf-syntax#Property"/>
  <rdfs:subPropertyOf resource="http://purl.org/metadata/dublin_core#Identifier"/>
  <rdfs:label>Accession_no</rdfs:label>
  <rdfs:comment>A unique number assigned by the automatic classifier
  that uniquely identifies this resource.</rdfs:comment>
</rdf:Description>

<rdf:Description ID="Title">
  <rdf:type rdf:resource="http://www.w3.org/TR/WD-rdf-syntax#Property"/>
  <rdfs:subPropertyOf resource="http://purl.org/metadata/dublin_core#Title"/>
  <rdfs:label>Title</rdfs:label>
  <rdfs:comment>The title of the resource taken from the HTML TITLE element.
  </rdfs:comment>
</rdf:Description>

<rdf:Description ID="Abstract">
  <rdf:type rdf:resource="http://www.w3.org/TR/WD-rdf-syntax#Property"/>
  <rdfs:subPropertyOf resource="http://purl.org/metadata/dublin_core#Description"/>
  <rdfs:label>Abstract</rdfs:label>
  <rdfs:comment>This is the first 25 words taken from the BODY of the HTML
  page, or, if present, text taken from the description HTML META tag.
  </rdfs:comment>
</rdf:Description>

<rdf:Description ID="Keyword">
  <rdf:type rdf:resource="http://www.w3.org/TR/WD-rdf-syntax#Property"/>
  <rdfs:subPropertyOf resource="http://purl.org/metadata/dublin_core#Subject"/>
  <rdfs:label>Keyword</rdfs:label>
  <rdfs:comment> This is a keyword from the document that matched a keyword
  in an appropriate DDC class representative. A number of keywords will
  normally appear in an RDF Bag container.</rdfs:comment>
</rdf:Description>

<rdf:Description ID="Classmark">
  <rdf:type rdf:resource="http://www.w3.org/TR/WD-rdf-syntax#Property"/>
  <rdfs:subPropertyOf resource="http://purl.org/metadata/dublin_core#Subject"/>
  <rdfs:label>Classmark</rdfs:label>
  <rdfs:comment>This is a DDC classmark that has been assigned to the document
  as a result of the automatic classification process. Often two appropriate
  classmarks will be shown in an RDF sequence - the highest scoring one
  appearing first.</rdfs:comment>
</rdf:Description>
```

```
<rdf:Description ID="Word_count">
  <rdf:type rdf:resource="http://www.w3.org/TR/WD-rdf-syntax#Property"/>
  <rdfs:label>Word_count</rdfs:label>
  <rdfs:comment>This is the number of individual words found in the
  resource.</rdfs:comment>
</rdf:Description>

<rdf:Description ID="Classification_date">
  <rdf:type rdf:resource="http://www.w3.org/TR/WD-rdf-syntax#Property"/>
  <rdfs:label>Classification_date</rdfs:label>
  <rdfs:comment>The date on which the resource was classified.</rdfs:comment>
</rdf:Description>

<rdf:Description ID="Last_modified">
  <rdf:type rdf:resource="http://www.w3.org/TR/WD-rdf-syntax#Property"/>
  <rdfs:subPropertyOf resource="http://purl.org/metadata/dublin_core#Date"/>
  <rdfs:label>Last_modified</rdfs:label>
  <rdfs:comment>The date on which the resource was last modified
  when it was classified.</rdfs:comment>
</rdf:Description>

</rdf:RDF>
```

Appendix B

The following RDF descriptions have been automatically generated. The automatic metadata generator is a Java program that retrieves HTML pages from given URLs and automatically analyses and classifies them according to DDC (see section 2). The DDC classmarks along with other accessible metadata elements (see figure 5) are then represented in RDF using the Wolverhampton Core (wc) schema (see Appendix A). The example pages have been selected from the top of a random range of Yahoo categories as indicated. Note that the accession number is not set in the following examples because the program is running as a stand alone application and not within the context of the search engine.

Yahoo - Home : Social Science : Psychology

http://dir.yahoo.com/Social_Science/Psychology/Education


```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
 xmlns:wc="http://scit.wlv.ac.uk/~ex1253/wc/schema/">
        <rdf:Description about ="http://www-nmcp.med.navy.mil/psychology/I1.htm">
                <wc:Accession_no>0</wc:Accession_no>
                <wc:Title>I1</wc:Title>
                <wc:Abstract>Psychology Department Home Page Clinical Psychology
                Internship Since 1990 the Psychology Department has offered a predoc
                clinical psychology internship fully accredited by the American Psyc
                <wc:Keyword>
                        <rdf:Bag>
```

```
                        <rdf:li>psychology</rdf:li>
                        <rdf:li>psychological</rdf:li>
                        <rdf:li>association</rdf:li>
                        <rdf:li>adult</rdf:li>
                        <rdf:li>training</rdf:li>
                        <rdf:li>leadership</rdf:li>
                        <rdf:li>American</rdf:li>
                        <rdf:li>navy</rdf:li>
                        <rdf:li>naval</rdf:li>
                </rdf:Bag>
        </wc:Keyword>
        <wc:Classmark>
                <rdf:Seq>
                        <rdf:li>350     Public Administration and Military S
                        <rdf:li>158     Applied psychology</rdf:li>
                </rdf:Seq>
        </wc:Classmark>
        <wc:Word_count>109</wc:Word_count>
        <wc:Classification_date>11-Nov-98 14:53:32</wc:Classification_date>
        <wc:Last_modified>07-Aug-98 14:55:04</wc:Last_modified>
    </rdf:Description>
</rdf:RDF>
```

```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
 xmlns:wc="http://scit.wlv.ac.uk/~ex1253/wc/schema/">
    <rdf:Description about ="http://spsp.clarion.edu/mm/RDE3/start/">
        <wc:Accession_no>0</wc:Accession_no>
        <wc:Title>Research Design Explained 3rd ed</wc:Title>
        <wc:Abstract>Aids for teaching research methods in psychology</wc:Ab
        <wc:Keyword>
                <rdf:Bag>
                        <rdf:li>computer</rdf:li>
                        <rdf:li>psychology</rdf:li>
                        <rdf:li>psychological</rdf:li>
                        <rdf:li>measure</rdf:li>
                        <rdf:li>experiment</rdf:li>
                        <rdf:li>experiments</rdf:li>
                        <rdf:li>research</rdf:li>
                        <rdf:li>learning</rdf:li>
                        <rdf:li>single</rdf:li>
                        <rdf:li>teaching</rdf:li>
                        <rdf:li>rights</rdf:li>
                        <rdf:li>writing</rdf:li>
                        <rdf:li>science</rdf:li>
                </rdf:Bag>
        </wc:Keyword>
        <wc:Classmark>
                <rdf:Seq>
                        <rdf:li>158     Applied psychology</rdf:li>
                        <rdf:li>150.724 Experimental research (Psychology)</
                </rdf:Seq>
        </wc:Classmark>
        <wc:Word_count>205</wc:Word_count>
        <wc:Classification_date>11-Nov-98 14:57:27</wc:Classification_date>
```

```
                <wc:Last_modified>31-Aug-98 12:53:37</wc:Last_modified>
        </rdf:Description>
</rdf:RDF>
```


Yahoo - Home : Reference : Libraries : Library and Information Science : Institutes

http://dir.yahoo.com/Reference/Libraries/Library_and_information_Science/Institutes


```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
 xmlns:wc="http://scit.wlv.ac.uk/~ex1253/wc/schema/">
        <rdf:Description about ="http://www.new-zealand.edu/infostudies/">
                <wc:Accession_no>0</wc:Accession_no>
                <wc:Title>Centre for Information Studies</wc:Title>
                <wc:Abstract>Courses on Information Literacy Staff Administration
                Current Courses Diplomas Certificates School based Courses Regional
                Courses Holiday Courses Courses for non teaching staff Newsletter We
                Tutorial</wc:Abstract>
                <wc:Keyword>
                        <rdf:Bag>
                                <rdf:li>librarianship</rdf:li>
                                <rdf:li>newsletter</rdf:li>
                                <rdf:li>education</rdf:li>
                                <rdf:li>school</rdf:li>
                                <rdf:li>administration</rdf:li>
                                <rdf:li>teacher</rdf:li>
                                <rdf:li>teaching</rdf:li>
                        </rdf:Bag>
                </wc:Keyword>
                <wc:Classmark>
                        <rdf:Seq>
                                <rdf:li>021     Relationships of libraries, archives
                                <rdf:li>370     Education</rdf:li>
                        </rdf:Seq>
                </wc:Classmark>
                <wc:Word_count>34</wc:Word_count>
                <wc:Classification_date>11-Nov-98 15:08:20</wc:Classification_date>
                <wc:Last_modified>15-Sep-98 22:25:51</wc:Last_modified>
        </rdf:Description>
</rdf:RDF>
```

```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
 xmlns:wc="http://scit.wlv.ac.uk/~ex1253/wc/schema/">
        <rdf:Description about ="http://www.mmu.ac.uk/h-ss/dic/">
                <wc:Accession_no>0</wc:Accession_no>
                <wc:Title>Department of Information and Communications MMU UK</wc:Ti
                <wc:Abstract>The Department of Information and Communications at
                the Manchester Metropolitan University UK Includes course research a
                contact details</wc:Abstract>
```

```
            &lt;wc:Keyword&gt;
                    &lt;rdf:Bag&gt;
                            &lt;rdf:li&gt;library&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;communications&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;school&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;university&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;science&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;management&lt;/rdf:li&gt;
                    &lt;/rdf:Bag&gt;
            &lt;/wc:Keyword&gt;
            &lt;wc:Classmark&gt;
                    &lt;rdf:Seq&gt;
                            &lt;rdf:li&gt;380     Commerce, Communications, Transporta
                            &lt;rdf:li&gt;027     General libraries, archives, informa
                    &lt;/rdf:Seq&gt;
            &lt;/wc:Classmark&gt;
            &lt;wc:Word_count&gt;7&lt;/wc:Word_count&gt;
            &lt;wc:Classification_date&gt;11-Nov-98 15:16:18&lt;/wc:Classification_date&gt;
            &lt;wc:Last_modified&gt;30-Jun-98 12:02:23&lt;/wc:Last_modified&gt;
    &lt;/rdf:Description&gt;
&lt;/rdf:RDF&gt;
```

Yahoo - Home : Computers and Internet : Programming Languages

http://dir.yahoo.com/Computers_and_Internet/Programming_Languages/

```
&lt;?xml version="1.0"?&gt;
<rdf:RDF
 xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
 xmlns:wc="http://scit.wlv.ac.uk/~ex1253/wc/schema/">
    &lt;rdf:Description about ="http://www.ampl.com/cm/cs/what/ampl/"&gt;
            &lt;wc:Accession_no&gt;0&lt;/wc:Accession_no&gt;
            &lt;wc:Title&gt;AMPL Modeling Language for Mathematical Programming</wc:Ti
            &lt;wc:Abstract&gt;FAQ BOOK SOLVERS PLATFORMS VENDORS CALENDAR MORE WHAT'S
            NEW EXTENSIONS CHANGE LOG REPORTS NETLIB EXAMPLES CONTENTS HOME AMPL
            Modeling Language for Mathematical Programming Try&lt;/wc:Abstract&gt;
            &lt;wc:Keyword&gt;
                    &lt;rdf:Bag&gt;
                            &lt;rdf:li&gt;computer&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;programming&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;modeling&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;communication&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;mathematical&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;model&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;models&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;control&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;linear&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;nonlinear&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;discrete&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;interface&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;web&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;com&lt;/rdf:li&gt;
                            &lt;rdf:li&gt;language&lt;/rdf:li&gt;
                    &lt;/rdf:Bag&gt;
            &lt;/wc:Keyword&gt;
            &lt;wc:Classmark&gt;
                    &lt;rdf:Seq&gt;
```

```
                                <rdf:li>004.6   Interfacing and communications (Comp
                                <rdf:li>005.1   Programming (Computer programming)</
                        </rdf:Seq>
                </wc:Classmark>
                <wc:Word_count>193</wc:Word_count>
                <wc:Classification_date>11-Nov-98 15:22:31</wc:Classification_date>
                <wc:Last_modified>08-Nov-98 23:39:21</wc:Last_modified>
        </rdf:Description>
</rdf:RDF>
```

```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
 xmlns:wc="http://scit.wlv.ac.uk/~ex1253/wc/schema/">
        <rdf:Description about ="http://quimby.fla.com/Activities/Programming/APRIL/
                <wc:Accession_no>0</wc:Accession_no>
                <wc:Title>APRIL</wc:Title>
                <wc:Abstract>Home page of the Network Agent Research Group within
                Fujitsu Laboratories of America Inc</wc:Abstract>
                <wc:Keyword>
                        <rdf:Bag>
                                <rdf:li>computing</rdf:li>
                                <rdf:li>programming</rdf:li>
                                <rdf:li>system</rdf:li>
                                <rdf:li>communication</rdf:li>
                                <rdf:li>model</rdf:li>
                                <rdf:li>internet</rdf:li>
                                <rdf:li>language</rdf:li>
                                <rdf:li>rights</rdf:li>
                                <rdf:li>interaction</rdf:li>
                                <rdf:li>america</rdf:li>
                        </rdf:Bag>
                </wc:Keyword>
                <wc:Classmark>
                        <rdf:Seq>
                                <rdf:li>003.5   Theory of communication and control
                                <rdf:li>005.1   Programming (Computer programming)</
                        </rdf:Seq>
                </wc:Classmark>
                <wc:Word_count>46</wc:Word_count>
                <wc:Classification_date>11-Nov-98 15:24:31</wc:Classification_date>
                <wc:Last_modified>Not known</wc:Last_modified>
        </rdf:Description>
</rdf:RDF>
```


Yahoo - Home : Society and Culture : Religion and Spirituality

http://dir.yahoo.com/Society_and_Culture/Religion_and_Spirituality/


```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
```

```
 xmlns:wc="http://scit.wlv.ac.uk/~ex1253/wc/schema/">
	<rdf:Description about ="http://www.theatlantic.com/election/connection/reli
		<wc:Accession_no>0</wc:Accession_no>
		<wc:Title>Political Issues Religion</wc:Title>
		<wc:Abstract>RELIGION Articles from The Atlantic Monthly 's archive
		and related links Welcome to the Next Church by Charles Trueheart 19
		Seamless multimedia worship round the</wc:Abstract>
		<wc:Keyword>
			<rdf:Bag>
				<rdf:li>intellectual</rdf:li>
				<rdf:li>multimedia</rdf:li>
				<rdf:li>vision</rdf:li>
				<rdf:li>archive</rdf:li>
				<rdf:li>cultural</rdf:li>
				<rdf:li>school</rdf:li>
				<rdf:li>copyright</rdf:li>
				<rdf:li>service</rdf:li>
				<rdf:li>religious</rdf:li>
				<rdf:li>university</rdf:li>
				<rdf:li>public</rdf:li>
				<rdf:li>family</rdf:li>
				<rdf:li>church</rdf:li>
				<rdf:li>worship</rdf:li>
				<rdf:li>god</rdf:li>
				<rdf:li>christian</rdf:li>
				<rdf:li>christianity</rdf:li>
				<rdf:li>spiritual</rdf:li>
				<rdf:li>religion</rdf:li>
				<rdf:li>America</rdf:li>
				<rdf:li>politics</rdf:li>
				<rdf:li>political</rdf:li>
				<rdf:li>rights</rdf:li>
				<rdf:li>communities</rdf:li>
			</rdf:Bag>
		</wc:Keyword>
		<wc:Classmark>
			<rdf:Seq>
				<rdf:li>027	General libraries, archives, informa
				<rdf:li>210	Philosophy and Theory of Religion</r
			</rdf:Seq>
		</wc:Classmark>
		<wc:Word_count>243</wc:Word_count>
		<wc:Classification_date>23-Nov-98 17:34:38</wc:Classification_date>
		<wc:Last_modified>Not known</wc:Last_modified>
	</rdf:Description>
</rdf:RDF>
```

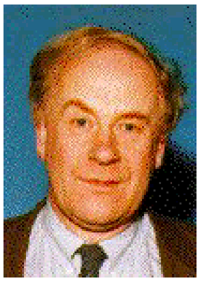
```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"
 xmlns:wc="http://scit.wlv.ac.uk/~ex1253/wc/schema/">
	<rdf:Description about ="http://www.inlink.com/~rife/religion.html">
		<wc:Accession_no>0</wc:Accession_no>
		<wc:Title>Dave's Controversial Religion Page</wc:Title>
		<wc:Abstract>Dave's Controversial Religion Page The Monroe
		Institute Spirit WWW site The Myth of the Historical Jesus Tibetan
```

```
                Book of the Dead Faqir Chand The Unknowing</wc:Abstract>
                <wc:Keyword>
                        <rdf:Bag>
                                <rdf:li>jesus</rdf:li>
                                <rdf:li>spirit</rdf:li>
                                <rdf:li>religion</rdf:li>
                        </rdf:Bag>
                </wc:Keyword>
                <wc:Classmark>
                        <rdf:Seq>
                                <rdf:li>210     Philosophy and Theory of Religion</r
                                <rdf:li>290     Comparative Religion and Other Relig
                        </rdf:Seq>
                </wc:Classmark>
                <wc:Word_count>76</wc:Word_count>
                <wc:Classification_date>23-Nov-98 17:37:05</wc:Classification_date>
                <wc:Last_modified>Not known</wc:Last_modified>
        </rdf:Description>
</rdf:RDF>
```

Vitae

Charlotte Jenkins is a Research Student at the University of Wolverhampton, UK. Her research is concerned with tools for information resource discovery on the Web and in particular automatic classification. Charlotte graduated from Oxford Brookes University in 1995 with a B.Sc. Joint Honours in Computing and English studies

Mike Jackson is Professor of Data Engineering at the University of Wolverhampton, UK. He is a Fellow of the British Computer Society. He obtained his first degree in Computer Science at Sheffield City Polytechnic and his Masters at Manchester University. Mike has served on the organising and programme committees of numerous database conferences including BNCOD, IDEAS, EDBT, ICDE, ER and VLDB.



Peter Burden graduated with a BA in Mathematics from the University of Cambridge in 1964. He is currently employed in the School of Computing and Information Technology at the University of Wolverhampton where he teaches systems and network programming and is responsible for the School's Unix based systems. His research interests include Internet resource discovery and cataloguing.

Jon Wallis works as an IT consultant in the pharmaceutical industry, specialising in laboratory automation and robotics. Prior to this, Jon was a senior lecturer at the University of Wolverhampton. His teaching was mainly in the area of computer networks and communication systems, with research interests in web search engines and the information management issues of corporate web sites.