The NewsAgent for Libraries electronic current awareness service for information professionals - extracting useful metadata from electronic records

Michael Keen, Jane Secker, and David Stoker
Department of Information & Library Studies,
University of Wales, Aberystwyth, UK

Contact: emk@aber.ac.uk

(This paper was presented at a Conference in May 1997, in Crimea. This edited version produced 16 July 1997)

BACKGROUND TO THE PROJECT

The NewsAgent for Libraries project, funded by the UK Electronic Libraries Programme, is seeking to provide an experimental, user configurable, electronic alerting service for library and information professionals. It will be a distributed system, based on three servers whereby relevant current information can be delivered electronically to users in a variety of formats, and at different levels of detail. Alerts may be received via e-mail, via individual web-sites accessed through a browser such as Netscape or else by a purpose built Windows 95 'NewsAgent Client'. The software is being developed by Fretwell Downing Informatics and is loosley based upon an existing client/server software package known as Dali (Document and Library Integration). Full details of the project are available at the project's website <http://www.sbu.ac.uk/litc/newsagent/>.

The project is the result of a merger of separate bids to FIGIT/eLib for current awareness services to LIS professionals. It involves consortia of the Library Information Technology Centre at South Bank University, the Department of Information and Library Studies at University of Wales Aberystwyth, the Centre for Research in Library and Information Management at University of Central Lancashire, the UK Office of Library Networking at University of Bath and Fretwell Downing Informatics. Three of these institutions are also the compilers and publishers of existing current awareness products or publications. However the project is also involving the collaboration of a number of publishers and information providers in this field, notably, Aslib, the Library Association, the Institute of Information Scientists, and the commercial publisher Bowker-Saur.

The two-year project began in April 1996 with a survey of Filtering Agents on the Web followed by a user needs analysis. This was followed by the design and determination of the editorial content of a demonstration system which it is hoped will be launched in late 1997. This will be subjected to an evaluation process which will be fed back into possible proposals for the establishment of a commercial system. The object of this paper is to report on the progress of the project to date and to consider ways of extracting useful metadata from a number of different categories of electronic publication, for use in the demonstration system.

USER NEEDS ANALYSIS

The User Needs Analysis made use of group discussion of six categories of information professionals, and an e-mail questionnaire of subscribers to the LIS-LINK discussion list. This was conducted in order to determine the factors which should influence the design and content of the service (Stoker and Secker, 1997). This showed the need for providing information on all kinds of subjects, in a variety of formats. However it particularly highlighted the growing importance e-mail, web pages and also electronic newsletters for current awareness purposes. However at the same time it identified the major problems of

information overload, and the changing nature of individuals' current awareness needs. It also stressed the significance of informal sources such as discussion lists, meetings, etc. as opposed to more formal current awareness services.

Yet at the same time there was an identifiable concern for the 'quality' of the information in terms of its authority, reliability, and currency. Thus the informality and timeliness of an un-moderated Usenet Newsgroup Posting was frequently regarded as being invalidated by its unreliability.

The ease of use of a system was more important than the visual attractiveness of its presentation. Desirable features include a logical structure, short bites of information with the ability to focus in as required. Information should be delivered daily to an individual's desk, where possible fitting in with current practices, such as the regular checking of e-mail. The ability to tailor the output of any system was a crucial element in the saving of users' time by ensuring only relevant material is delivered. The system would have to be flexible and offer users the ability to amend it as required. Many individuals were concerned that if they were required to define a narrow profile they might be depriving themselves of information. The ideal service might contain aspects of both these solutions, giving the ability to focus, but also access to a more general service where required.

There was a recognition that no one service would ever be likely to meet all the needs of different professionals, but taking as many of these factors as possible into account would be likely increase the chance of any such service being successful.

SYSTEM DESIGN AND DELIVERY

The three basic issues to be addressed in the design of the trial system would be:

- what to cover (editorial content).
- what to deliver (full text, abstract, or metadata only)
- how to gather the data

As originally envisaged, NewsAgent would parse keywords and phrases from electronic documents using existing Web-harvesting technology, and filter them to users through a series of seven channels (genre topics) and sixteen subject topics. The former group would include:

- Publications
- Research
- Products
- People
- Statistics
- Events
- Discussions

The latter group would include:

- Information Management
- Information profession
- Libraries
- Information Users
- Management
- Buildings
- Library Technology
- Technical Services
- Information Work

- Information Retrieval Processes and Schemes
- Information Retrieval Systems
- Information Technology
- Reading
- Publishing
- Records Management

However during the course of the project the Dublin Core Metadata Standards have emerged and for the purposes of the demonstration system, NewsAgent will seek to use and where necessary, enhance this standard, as the primary means of identifying and classifying the data which will be delivered to subscribers. The ability to parse keywords and phrases will however be retained at a later stage in order to provide the degree of subject specificity demanded by certain categories of users.

The Dublin Core (DC) proposals <http://purl.org/metadata/dublin_core_elements> identify fifteen fairly straightforward metadata elements which may be used to describe "document like objects". These include elements such as Title, Creator, Subject, Date, Format, which have been used to catalogue printed documents for centuries. In addition there are others which are more specifically related to electronic environments - these include Resource Identifier (such as a URL), Relation (which describes relationships within a hierarchy such as files within a directory), and Rights Management. Metadata may be 'embedded' within documents, particularly Web pages <http://www.ub2.lu.se/tk/metadata/DC10cats.html>, in the same way as Cataloguing in Publication Data can be provided in the preliminaries of printed books. All of these metadata elements will not necessarily be present in any one document: the presence or absence of many of them will depend upon the type, format, and status of the document Where necessary metadata elements may be repeated, or else qualified to provide a more specific and sophisticated level of description.

One of the main problems of relying upon DC metadata elements embedded within a source is that so few currently include it. However, UKOLN, one of the collaborating partners, has developed a DC Generator known as DC-dot which will retrieve a Web page and automatically generate the HTML <META> tags suitable for embedding within the <HEAD> section of the Web-page, which will not be visible on the screen but may be identifiable by web-harvesters and search engines <http://www.ukoln.ac.uk/metadata/dcdot/>. A NewsAgent version of DC-dot is at <http://www.ukoln.ac.uk/metadata/NewsAgent/dc/>. Likewise it was felt that software could be developed to parse other categories of electronic record to obtain at least some of the necessary metadata elements without human intervention.

Of the fifteen DC metadata elements one (Title) is designated as Mandatory, and one (Subject) is Highly Desirable. Two others, Publisher and Rights Management, are also Mandatory and will be provided automatically for each document by a link to a publisher record. The rest are regarded as either Desirable or Optional (see Appendix 1). There is discussion as to whether two will be usable in their existing form by NewsAgent - these are Relation (described above), and Coverage (which is a free text description of the temporal coverage of a resource such as "Italy during the Reformation"), or whether they will have to be replaced by NewsAgent's own versions. Wherever possible the pure Dublin Core elements will be used, both to save labour and for the sake of adhering to common standards.

In addition, the NewsAgent system would require metadata elements not currently found in the Dublin core set. These are:

- genre topic (highly desirable),
- subject topic (highly desirable),

- contact - such as e-mail address, home page URL (optional)
- date of validity - defining the time period in which information is useful (optional)

The issues of 'information quality' will be addressed both by the reputations of the collaborating publishers and also by reference to the Dublin Core list of thirty-six standard resource types, eleven of which include some degree of indication of quality - such as the differentiation between refereed and un-refereed journals, magazines, moderated and un-moderated mailing lists and newsgroups etc (annotated list available at <http://www.roads.lut.ac.uk/Metadata/DC-ObjectTypes.html>).

EDITORIAL CONTENT

The main determinants of the editorial content of the demonstration system were that there should be a range of useful and relevant data, in different electronic formats, which was freely available without copyright or rights complications. Sources which were either published by those institutions involved with the eLib programme, or the NewsAgent project or else were edited by members of the team were therefore obvious candidates for inclusion. Five sources, all dealing with electronic libraries were chosen from many possible candidates. These are:

1. *Electronic Calendar* (University of Wales, Aberystwyth) - a calendar of forthcoming events compiled and published by the Thomas Parry Library on a Web Site <http://www.aber.ac.uk/~tplwww/eleccal/>. Compiled using Microsoft Access, it contains approximately 800 records per year.
2. *Ariadne - The Web Version* (UK Office of Library Networking and University of Abertay, Dundee) - a magazine dealing with Electronic Libraries, published in hard-copy every two months and in a Web version more frequently with more material <http://www.ukoln.ac.uk/ariadne/>
3. *Library Technology* (Library Association and Library Information Technology Centre) - the IT supplement to the *Library Association Record*, compiled by LITC published 6 times each year.
4. *Program - electronic library and information systems* (Aslib) a quarterly refereed journal, edited by Lucy Tedd of University of Wales published in hard copy and also available via the Web. Non-subscribers have access to the bibliographic details and abstracts in html, whereas subscribers may also view the full text as pdf files <http://www.aslib.co.uk/program/>
5. The *LIS-Elib* - The 'E-lib projects' un-refereed email discussion list.

Other possible candidates for inclusion at a later date would be various Bowker-Saur publications, including the *Journal of Librarianship and Information Science (JOLIS).* This publication will be available to subscribers in an enhanced web-version in the autumn of 1997, and will incorporate a feedback discussion group moderated by the journal editor David Stoker (also a member of the NewsAgent team).

Extracting metadata

If NewsAgent is ever to become a viable commercial service, it is essential that as much of the data required by the system as possible should be identified and retrieved automatically without human intervention. None of these sources currently contain embedded DC metadata, although it is hoped that both *Library Technology* and *Ariadne* will shortly do so on an experimental basis. The NewsAgent harvester will then have no difficulty in locating the web pages, translating them into a usable form and filtering them through subject channel. Exactly how it will cope with the seven NewsAgent genre topics (Publications, Research, Products,

Events, Discussions, People, and Statistics) remains a difficult and as yet unanswered question. Certainly some clues may come from the publication source, or the nature of the data, but at present there seems to be no watertight means of allocating using computer algorithms.

The highly structured record of *Electronic Calendar* which can be exported in a comma delimited form, means that it will be a relatively simple task to parse in order to extract metadata. It is this source which will most require the additional metadata element of date of validity.

Similarly, all e-mails sent to a discussion list will contain certain basic fields within the headers, which may be automatically translated into the metadata elements without complication. From example: From: equals DC Creator, Reply To: equals DC Publisher, Subject: Equals DC Title, etc. Also relationships between messages can perhaps be established for the DC Relation element. There will inevitably be complications whereby titles may not always represent true subjects, and which of the various dates found on e-mails is to be used, but these ought not to be insuperable.

The issue of how to deal with full-text journals will also have to be dealt with. As originally envisaged, these sources would be supplied as pdf or sgml files and separately mounted on one of the NewsAgent servers, but their availability on the Web has tended to overtake the earlier plans.

Further details of NewsAgent's approach to metadata, and to the use of channels, topics and keywords in matching users' profiles, will appear in a forthcoming paper (Keen, 1997).

References

1. Keen, Michael, 1997, Using Channels, Topics, Keywords and other Metadata in an Electronic Alerting Service. (Paper submitted to Online Information 97)

2. Stoker, David, and Secker, Jane, 1997, The design and content of an electronic current awareness service for information professionals, Electronic Library and Visual Information Research - Elvira 4, Aslib, 57-64.

Appendix 1 - Metadata Elements from Dublin Core (Number and DC.) and NewsAgent.

(1) DC.title	MANDATORY
(2) DC.creator (or author)	OPTIONAL
(2) DC.creator.email	OPTIONAL
(3) DC.subject	HIGHLY DESIRABLE
(N) NewsAgent.topic	HIGHLY DESIRABLE
(4) DC.description	OPTIONAL
(N) NewsAgent.contact	DESIRABLE
(N) NewsAgent.contact.email	DESIRABLE
(N) NewsAgent.contact.homepage	DESIRABLE
(5) DC.publisher	MANDATORY (BY LINK)
(6) DC.contributor (other contributors)	DESIRABLE
(7) DC.date	OPTIONAL
(N) NewsAgent.date.validto	OPTIONAL
(8) DC.type (resource type)	OPTIONAL
(9) DC.format	OPTIONAL
(10) DC.identifier.(resource identifier)	OPTIONAL
(11) DC.source	OPTIONAL
(12) DC.language	OPTIONAL
(13) DC.relation	OPTIONAL
(14) DC.coverage	OPTIONAL
(15) DC.rights (rights management)	MANDATORY (BY LINK)