DESIGNING SOFTWARE FOR MANAGING INTERNET RESOURCE CATALOGUES: A CASE STUDY WITH IRCAT-M

by

Prasanan T S *
Shravan Kumar S*
Rajashekar T B*

ABSTRACT

The Internet is beginning to be seen as a serious, useful source of information, over the past few years. There is significant interest among users to find relevant Internet information, of relevance to their day-to-day tasks. Libraries are finding it necessary to take proactive action in identifying, evaluating and reporting relevant Internet resources to their users. As the number of relevant Internet resources increase, libraries need to develop adequate solutions to manage access to these resources. Libraries have followed different, often ad-hoc, approaches in providing such solutions. We propose a database-driven, web-based Internet Resource Catalogue Manager as a general-purpose solution, based on international standards. We describe the design factors one has to consider in developing such a system and a prototype package we have developed, called IRCAT-M. It is Web-based system, developed for Windows-based back-end server. It facilitates setting up and managing a database of Internet resources and provides web-based content management, search and browse interface. We discuss the design features in detail, implementation considerations, operational features and further work to be carried out.

Keywords: Internet Resources, Catalogues – Management, Software Design – Content Management

---

*** National Centre for Science Information, Indian Institute of Science, Bangalore**

0 Introduction

There is tremendous growth in the number and variety of online information resources available on the Internet today. There is also growing evidence that users increasingly rely on Internet-based information for their day-to-day tasks and they spend considerable time in finding such information. It is therefore becoming very important for libraries to develop appropriate means for organizing and providing quick access to relevant Internet information sources to their users. Given the large number of Internet sources that could be of potential relevance to organizations, it would appear that libraries will need some kind of software support for organizing and providing managed access to these resources. What features are to be supported by such software, what are the design and implementation issues? We address these aspects in this paper, in terms of a prototype software IRCAT-M (Internet Resource Catalogue Manager) that we recently developed at

NCSI. We discuss the following aspects in the paper: need for managing access to Internet resources; strategies currently adopted by libraries; key issues that need to be addressed in developing software support; and design, implementation and operational features of IRCAT-M. We conclude with further work that needs to be done.

1 Need for managing access to Internet resources

Parallel to the rapid growth in number of online resources on the Internet, several tools have been developed to facilitate resource discovery. These include: general and specialty search engines, general and specialty directories, and meta search tools. Though these have been of great help, users still need to invest considerable effort and time in making informed use of these tools and in judging the appropriateness and currency of the retrieved sources before deciding to visit them on the web. Such investment will be very expensive to organizations if every user independently explores the Internet to find more or less the same resources, which another colleague would have already found out. This is so since there is a broad commonality of interests among users within organizations for which one could identify a core set of Internet resources of direct relevance to most users. This means that organizations will do well to develop and maintain a local Internet resource catalogue (IRC) of highly relevant Internet resources and provide access to this on the intranet. Library users can access the IRC on the library website, select one or more sources and then visit these on the Internet/ intranet. Further, such a catalogue can be used to provide managed access to both subscribed and free Internet sources and also those hosted on the intranet.

2 Strategies adopted by libraries

Libraries have evolved different strategies for developing and providing access to IRCs via library websites. These include: static HTML pages with brief description and links to Internet sites, dynamic web pages generated from a back-end database with support for browse and search functionality, and extension of library catalogues (OPACs) to include links to Internet sites. The number and types of fields used for describing Internet resources also significantly varies across libraries. Many library automation packages, which use MARC cataloguing, have adopted the field 856 to describe electronic information sources. However, extensive work done in evolving standards (e.g. Dublin Core), best practices (e.g. DESIRE Gateway Handbook) and services (e.g. EEVL, SOSIG, InterCat and CORC services of OCLC) for describing and managing network resources, have demonstrated the limitations of traditional approaches. It is well understood that traditional library catalogues, which have been designed mainly to describe physical documents stored in a library, are not adequate to describe network-based virtual information sources. Catalogue standards like MARC, AACR and CCF focus mainly on bibliographic and descriptive elements, with poor or no support for administrative and rights management aspects of network-resources. Most library automation packages do not provide adequate support for creating and maintaining Internet resource catalogues with these features and also lack browse and search features appropriate for these resources. This explains the reason why most libraries, in spite of

using a library automation package, adopt independent strategies to provide high quality access to Internet resources. It is to be hoped that future generation library automation packages will provide adequate support for describing and managing network resources. IRC designs, such as the one presented in this paper, can facilitate such improvements.

3 Factors to be considered in developing IRCs

What factors need to be taken into consideration in developing IRCs? One could identify several key factors from a study of best practices, standards and services in this area. These include both content and system level considerations.

?? **Factors related to content:**

- How many resources are to be covered?
- What is the scope of resources to be covered? (Subject, content level in terms of user needs, language, geographic boundaries, time, free/ subscribed, etc.)
- What resource types are to be covered (e.g. databases, data sets, e-journals, patents)
- What criteria we adopt for selection, evaluation and rating of resources?
- What strategy (procedure) we adopt for identifying relevant resources
- What metadata elements constitute an IRC record, in terms of descriptive (e.g. title, publisher, subject, URL), administrative (e.g. creation/ modification date, staff) and rights management (e.g. access rights, free/paid) related attributes? What is the syntax (constituent elements) of these elements?
- How do we render the content of these data elements (cataloguing rules for rendering field contents)?
- What hierarchical classification scheme we use for subject categorization of resources, to facilitate browsing by subject?
- Any special content formats supported by the resources to be identified and categorized?

?? System level factors:

- How do we add and update content into the IRC? What content management interface is required?
- How do we facilitate users to browse and search the content in IRC? What search/ browse/ display interface is required?
- How do we track the usage of IRC? What resources are accessed more, by whom? How do we report the usage? (Usage tracking and reporting system)
- How do we maintain the currency of content in IRC (e.g. validation of URLs)
- What back-end database is required to support content management, search and retrieval? What is its design?
- What additional application level features we want to support? (e.g. user feedback gathering, new sites recommendations, etc.)

4 IRCAT-M: Prototype software for IRC management

We have developed a prototype general-purpose software package, called IRCAT-M for IRC management. This has been prepared to understand better the design issues involved in developing such a package. The prototype version we have developed supports many of the factors mentioned in the previous section. IRCAT-M is a tool for creating, developing and management of Internet resource catalogues with features for content management and with a well-designed search and browse interface for the users of the IRC.

4.1 Development environment

IRCAT-M has been developed for the Windows platform, with Personal Web Server (PWS) as the web server, MS Access as the back-end database and PHP as the server side scripting language. This has been done to keep the development process simple. Windows is the most popular operating system used in libraries. MS Access is a very popular RDBMS package, part of the MS Office suite and is reasonably good for small and medium size databases. PWS is a freely available web server software made available by Microsoft. PHP is an open source server side scripting language for providing dynamic access to web site content, including that residing in databases.

As no software can be developed in vacuum, we have used real-life content in the area of 'Pharmaceutical Science and Technology' (PS&T) for development purposes. Further, to simplify content gathering process, we have restricted ourselves to free sources, in English language.

4.2 Design

We have grouped the various design factors discussed in Section 4 above, into four components in IRCAT-M: Content design, database design, content management interface design and user interface design. The four components are shown in Figure 1. We briefly discuss these in subsequent sections.

Fig. 1: Design components of IRCAT-M

4.3 Content design

Various content definitions we have provided in IRCAT-M are discussed in subsequent sections below.

4.3.1 Metadata elements:

We have adopted the Dublin Core (DC) Metadata element set for describing Internet resources covered in IRCAT-M. The Dublin Core is a 15-element metadata set intended to facilitate discovery of electronic resources. It is ideally suited for describing web-based sources. Table below shows the original DC elements, those DC elements not used in IRCAT-M, modified elements and additional elements we have defined.

Original DC Elements	Title, Author/ Creator, Subject/ Keywords, Description, Publisher, Other Contributor, Date, Resource Type, Format, Resource Identifier, Source, Language, Relation, Coverage, Rights Management
DC Elements NOT used in IRCAT-M	Other contributor, Relation
Modified DC elements	Rights Management – Access Type
Additional elements defined for IRCAT-M	Record Status, Keywords, Created by, Created date, Modified by, Modified date.

4.3.2 Resource types

Generally Resource types are the nature or genre of the content of the resource (type of content). Resource type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the working draft list of Dublin Core Types). 'Resource type' is a very useful attribute to describe an Internet resource, as many users are known to use this as an approach element for finding sources (e.g. patents, conferences). For IRCAT-M we have defined the following resource types: Conference, Guides, Discussion groups, Electronic Journals, Patents, Theses and Dissertations, Abstracting and Indexing Databases, Digital Collections, Product Catalogues, Library catalogues, Museum and Archives, Virtual Libraries, Reference sources, Employment, Libraries, Organizations, Companies, Directories and Miscellaneous.

4.3.3 Resource formats

Resource formats are the physical or digital manifestation of the resource. Typically, 'Format' may include the media-type or dimensions of the resource. Dublin Core defines resource formats as a list of worldwide-accepted MIME types. Identified resource formats are Text, Application, Images, Video and Audio.

4.3.4 Criteria for selection of resources

Usefulness of an IRC is directly related to the quality of Internet resources it covers. We have to define a set of criteria for the IRCAT-M cataloguer to evaluate and select an Internet resource for inclusion. Based on extensive work that has already been done in this area, we have prepared a toolbox for selecting the resources as below:

- ?? Scope- Breadth, depth, time and format
- ?? Content – Accuracy, authority, currency, uniqueness, quality of graphics and writing, purpose and audience, reviews, user friendliness, search and browse features.
- ?? Cost - Costs can be divided into: (1) costs of connecting to the resource, and (2) costs associated with the use of the intellectual property contained in the resource.

4.3.5 Strategy for identifying resources:

How do we identify web-based sources that may possibly be considered for inclusion in an IRC? Web is a very large resource base. One needs to evolve a systematic strategy for this purpose. We followed the following strategy for the domain 'pharmaceutical science and technology' used for IRCAT-M development. First we reviewed domain-specific virtual libraries and portals. Then we browsed general-purpose directories. We adopted this strategy since sites covered in these sources are generally selected by experts and hence are of high value. We followed this with web searching, first using meta search engines like Copernic, followed by general search engines. While formulating search strategies for search engines, we also included the resource types, as search parameters.

4.3.6 Criteria for resource rating

An Internet Resource Catalogue will be of great value if the sites included in it are ranked (rated). We have developed a rating criteria based on careful study of rating criteria followed by some of the scholarly sites. We also considered the rating system used in systems such as Argus Clearinghouse Ratings system and Science and Engineering Network News. Criteria for rating the resources are:

- ?? Content - Validity, Authority, Substantiveness, Accuracy, Comprehensiveness, Uniqueness, Composition and organization
- ?? Form – Ease of navigation, Provision of user support, Usage of standards, appropriate use of technology
- ?? Process – Information integrity, Site integrity, System integrity.

We have used a '*' rating system, with '*****' being the highest rate.

4.3.7 Classification scheme

It is highly desirable if the resources covered in an IRC are categorized using a hierarchical classification scheme, as classification schemes provide several advantages for resource discovery, in terms of browsing, searching and filtering. None of the general-purpose classification schemes (DDC, UDC) supported a good hierarchy for PS&T, the subject domain used for IRCAT-M development. We adopted the PS&T portion from Library of Congress Subject Headings and developed a three-level subject scheme. Though IRCAT-M is currently limited to this scheme, we hope to make the production version of the software to support incorporation of any scheme (of up to three levels).

4.3.8 Guidelines for preparing resource summary

A key metadata element in an IRC record is the summary (abstract) of the Internet resource covered by the IRC record. A user's decision to select a resource from the IRC is largely dependent on this summary. It is therefore crucial to prepare this summary carefully. Can we define a template in guiding the IRC cataloguer to focus on these aspects? We have defined the following template for IRCAT-M: objective of the site, target audience, scope of the subject areas covered, and source of content and volume of content.

4.4 Database design

IRCAT-M uses MS-Access as the back-end database. It has 7 tables – 4 master tables for holding the main contents of the resource and 3 link tables for linking multiple occurrences of subject, resource type and resource format, with the 'Resource' master using the resource identifier as the linking field. List of tables that are required for the database are:

?? **Master tables**
- Resource (main table containing details of each resource like resource id, title, publisher, URL, etc.)
- Subject (classification scheme, including the notation and the subject heading)
- Resource type (resource type code and resource type)
- Resource format (resource format code and resource format)

?? **Link tables**
- Resource – Subject Master (resource id, subject notation)
- Resource – Resource type (resource id, resource type code)
- Resource – Resource format (resource id, resource format code)

4.5 Content management interface design

A key component of IRCAT-M design is the requirement for developing completely web-based content management tool. This has to be designed to facilitate a content manager (cataloguer, librarian) to add a new Internet resource or modify an existing resource or delete an existing resource. The interface has to support features like content validation (e.g. non-empty condition for mandatory fields, formats), duplicate check (using URL), and updating of database tables. Further, it was required that 'deletion' of a record be a logical deletion only and not a physical deletion. This was to facilitate later restoration of the resource record, if necessary.

4.6 User interface design for search and browse

The most important component of IRCAT-M is the web-based user interface for search, browse and display of content from the database. This component was designed to meet several key features. For example, the browse interface should support browsing of resources by subjects, resource types, resource formats and by titles. Search interface should be able to search a word or combination of words or a part of a word. It should also be able to search in a particular field or combine search parameters across fields. Limiting the search for a particular subject or resource type or resource format option is also to be made available. Results display will be for displaying the resources with major fields like Title, URL, Author, Description and Site rating. Provision should be made for obtaining full details of a resource with all the fields. Also provision should be made for opening a separate browser window and visit the website of a particular resource from the results page.

5 Implementation

We have discussed earlier the development platform we have used for developing IRCAT-M. The prototype version of IRCAT-M consists of five major software segments: Content Management Segment, Simple Search Segment, Browse Segment, Advanced Search Segment and Display segment. We have also carried out testing of all the software segments. These include the following tests:

?? Content Management segment has been tested for insertion, modification and deletion of resources, subjects, resource types and resource formats.

?? Testing of Simple Search segment has been carried out in the following way:
   - o Getting the user input for query in simple search, browse and advanced search
   - o Communicating with the database
   - o Processing the query based on the search
   - o Processing the retrieved documents
   - o Displaying the final results in the required format, at10 results per page

?? Testing has also been carried out for each of the interfaces like simple search, browsing the catalogue by subject, resource type, resource formats and titles.

?? Software has been tested for advanced search with the limiting option, null query

6 Operational features

Operation of IRCAT-M is coordinated through its home page (Figure 2). The home page links to content management interface (for use by the IRC cataloguer or content manager); user interface (for use by the end user) in terms of simple and advanced search interfaces and also provided detailed information about the system itself (‘about IRCAT-M). While any user can open the user interface, content management interface is password protected restricting its use to only authorized staff to add and edit contents in the database. Content management interface facilitates the cataloguer to add/ edit content to the different tables – ‘Resource Master’, ‘Subject Master’, ‘Resource Type Master’ and ‘Resource Format Master’ (Figure 3). Simple search interface provides a keyword – based search on resource title and description fields, and browsing by subject, resource type, format and title (Figure 4). Advanced search interface facilitates searching on different fields, Boolean combination across search parameters and limiting the search to specific subjects(s), resource type(s) and format(s) (Figure 5). In search results, ten resources are shown per page, with provision for selecting other pages (Figure 6). For each resource, its title, author/publisher, description, site rating and URL are shown. A ‘More’ link is provided enabling the user to display all the metadata associated with a resource.

7 Further Work

In the context of design factors mentioned in Section 4, several enhancements can be made to IRCAT-M to make it more useful. These include the following:

- o Link currency checking
- o Usage tracking and reporting system
- o Enhancing the search interface to support phrase searching, word stem search, etc.

- Incorporation of relevant metadata element qualifiers recently developed to extend DC
- Porting the software to work on open platforms like Linux and MYSQL

8 Conclusion

The World Wide Web (WWW) consists of all the web servers that provide access to variety of hypertext documents, web pages, software, etc. Several organizations have developed their own Internet Resource Catalogues in their own way and using them in their intranets or websites. They have generally followed ad-hoc approach in their development. We propose a consistent development approach for IRCs. This can be facilitated by a general-purpose software. IRCAT-M is a demonstration in this direction. A fine-tuned, robust version of IRCAT-M can be of significant use to many libraries and information centers. The design we have presented here could also be incorporated into library automation packages to enable them to support high quality IRCs and make Internet information an integral part of OPACs.

9 References

1. MINJ (FILBERT) and RAJASHEKAR (T B) (ed.). Web servers- features, installation and configuration. Training Programme on Webmaster/Content Manager. Center for Continuing Education, IISc, Bangalore, 2001.
2. Ratings System. The Argus Clearinghouse Ratings System. 27 Mar. 1997. http://www.clearinghouse.net/ratings.html
3. SOSIG selection criteria. Evaluating Internet Resources for Social Science Information Gateway (SOSIG). http://www.sosig.ac.uk/desire/ecrit.html#cc
4. EEVL - The Edinburgh Engineering Virtual Library http://www.eevl.ac.uk/
5. CORC - The OCLC Cooperative Online Resource Catalog http://www.oclc.org/corc/
6. InterCat - Internet Cataloging Project http://www.oclc.org/oclc/man/catproj/catcall.htm
7. Evaluating Information on the Internet. Roger Williams University Libraries. 31 Aug 1999. http://www.rwuonline.cc/library/evaluat.html
8. PRASANNA (T S) and SHRAVAN KUMAR. Internet Resource Catalogue Manager (IRCAT-M): A case study with Pharmaceutical Science and Technology. Major Project Report. NCSI Training Course, October 2001.
9. SMITH (ALISTAIR G). Testing the Surf: Criteria for Evaluating Internet Information Resources. The Public-Access Computer Systems Review; 8 July 1997. http://info.lib.uh.edu/pr/v8/n3/smit8n3.html
10. DESIRE Information Gateways Handbook: Your guide to creating high quality portals on the Internet. http://www.desire.org/handbook/contents.html

11. Dublin Core Metadata Initiative (DCMI): Dublin Core Metadata http://dublincore.org/documents/dces/
12. Dublin Core Metadata Initiative (DCMI): Dublin Core Qualifiers http://dublincore.org/documents/dcmes-qualifiers/
13. http://www.ecs.fullerton.edu/pub/irl/docs/final.txt
14. http://bubl.ac.uk/link/
15. http://www.rdn.ac.uk/publications/terminology/