MAGiC – SHINING A NEW LIGHT ON A GREY AREA

Paul Needham

The MAGiC project aims to provide the UK engineering community with a greater awareness of, and access to, key collections of technical reports in the UK. MAGiC has laid the foundations of a National Reports Catalogue and created a web-based demonstrator service (METReS) by employing the Open Archives Initiative Protocol for Metadata Harvesting. The system raises the visibility of technical reports and removes the need for libraries to catalogue reports themselves. OAI metadata harvesting can be employed in a wide range of contexts beyond grey literature, and represents a new approach to exposing resources and making them available at low cost and with minimal effort.

Paul Needham, Electronic Information Specialist, Kings Norton Library, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK
E-mail: p.a.s.needham@ntlworld.com

Introduction

In the last few years the engineering community has gained enhanced access to electronic information resources. However, while major collections of science and technology journals from mainstream publishers have received a great deal of attention, comparatively little attention has been given to grey literature such as technical reports.

Grey literature is information which is not conventionally published and is therefore not readily available through the usual channels of booksellers and subscription agents. Typical examples of grey literature include theses, technical reports, conference proceedings, trade literature and official documents. Within the engineering sector one of the largest and most important types of grey literature is the technical report.

Technical reports have been produced, in one form or another, since the early part of the twentieth century. They are an important source of raw data, with a level of detail often omitted from published articles, and may contain experimental procedure, production data, specifications, standards, operating plans, drawings and/or raw results [1]. Importantly, they will set out the conclusions and recommendations to come from the research, and include all aspects of the research, including those parts that failed. This means that the information is very specific, addressing a particular problem, and very rich in experimental detail.

For these reasons, it is often the report, rather than the refereed journal article, which is the information medium favoured by the engineering community.

In the USA, much recent report literature from agencies such as NASA and the Departments of Energy and Defense is readily available on the Internet. In the UK, apart from the substantial

holdings of the British Library, other major collections of technical reports tend to be scattered across academia, government and industry. These resources are difficult to identify, locate and access.

The MAGiC project [2], funded jointly by the British Library's Co-operation and Partnership Programme [3] and the Research Support Libraries Programme [4], has taken the first steps towards enhancing awareness, access and utilisation of key collections of technical reports for the benefit of the UK engineering community.

Most importantly, seeking to ensure that access to technical reports becomes part of the continuing development of a distributed national electronic resource, MAGiC has laid the foundations of a National Reports Catalogue and an associated full text archive, delivered through a web-based 'proof of concept' demonstrator service - METReS [5].

National Reports Catalogue

The last few years have seen the meteoric rise in use of the Internet, with portals and search engines being particularly popular resource discovery tools. Increasingly, engineers rely on the World Wide Web as their primary source of information. So the challenge faced by MAGiC has been to provide a web-based solution that raises the visibility of engineering technical reports in the UK and provides an easy-to-use single entry point for the discovery of those reports. At the same time it is necessary to provide an infrastructure that will improve and ease their management. Here, we consider the type of database that is required to underpin such a solution.

Our original intention was to create a distributed National Reports Catalogue (NRC) by implementing a system architecture based around the Z39.50 protocol to give access to specifically identified library catalogues, other external databases and web resources.

This was based on the assumption that bibliographic references to report literature are embedded in library catalogues. However, this assumption proved to be only partially true.

Significantly, most academic libraries treat reports series as serials (which they are not!) and have only a single catalogue entry at series level.

Where there is item level cataloguing, report numbers are entered inconsistently and few library catalogues allow collections to be searched by reports as a separate category.

This means that the role of library catalogues in the NRC is limited by existing report cataloguing practices. This has significant implications for the shape and structure of the service. If libraries routinely catalogued individual reports and linked those catalogue records to a 'parent' record detailing the report series, it would be possible to create a system to search different catalogues in a distributed environment based around Z39.50 parallel searching. However, with the current lack of a cataloguing resource base, this was not feasible.

Furthermore, parallel searching in itself contributes nothing to the management or rationalisation of scientific and technical literature, or its associated records.

Instead, we decided to adopt a system where reports metadata would be harvested from available sources into a relational database, linking together descriptions of corporate sources, the reports series they produce, the reports themselves and the collections where reports are held.

Such a database would: remove the need for libraries to catalogue reports holdings in detail themselves; serve as a comprehensive reference point for the understanding of reports series; and offer a standard by which existing collections could be measured.

The system we adopted employs the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

The Open Archives Initiative

"The Open Archives Initiative (OAI) was conceived to develop and promote interoperability standards, with the aim of facilitating the efficient dissemination of content." [6]

OAI divides participants into two classes, 'data providers' and 'service providers':

- *Data Providers* administer systems that support the OAI Protocol for Metadata Harvesting (OAI-PMH) as a means of exposing metadata;

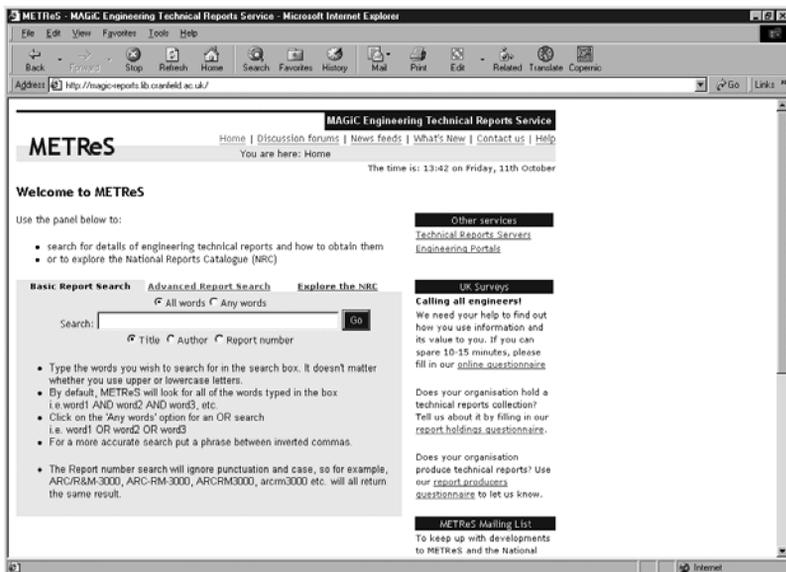- *Service Providers* use metadata harvested via the OAI-PMH as a basis for building value-added services. [7]

Figure 1. METReS homepage

By exposing bibliographic data in this manner, employing open standards, organisations that act as data providers minimise the effort needed to harvest their metadata, and greatly enhance the possibilities of interoperability and integration. Exchanging and sharing existing bibliographic records, in many cases, removes the need for organisations to create records themselves at all. And once a system is in place, the transfer of metadata is carried out automatically, on an ongoing basis, by computer scripts and requires no further human intervention.

Thus, the OAI provides a low-cost solution to the creation and transfer of metadata, while at the same time raising the visibility of available resources.

It is important to note that this is a system for the exchange of metadata, describing and pointing to the locations of documents and other resources. The documents themselves, whether electronic or hard copy, remain in their existing locations.

The protocol does not deal with issues of access restriction and management of intellectual property in exposed metadata, which remain the responsibility of the data

provider. Extensive documentation regarding the implementation of the OAI Protocol is available on the OAI website at http://www.openarchives.org/

METReS

The core service of METReS is the delivery of harvested bibliographic records from the National Reports Catalogue and documents from the full text archive, via a web interface (see Figure 1). METReS is, in effect, the window looking into the NRC. This demonstrates the strength of using harvested records which are held locally but describe resources in disparate locations.

Through the use of metadata harvesting, in a very short time it has been possible to populate the NRC with over 13,000 records describing reports. Each record provides links to information about:

- obtaining the document, either a URL in the case of electronic full text reports, or a document supply option in the case of hard-copy reports;

- collections where the report is held;

- the organisation responsible for the production of the report;

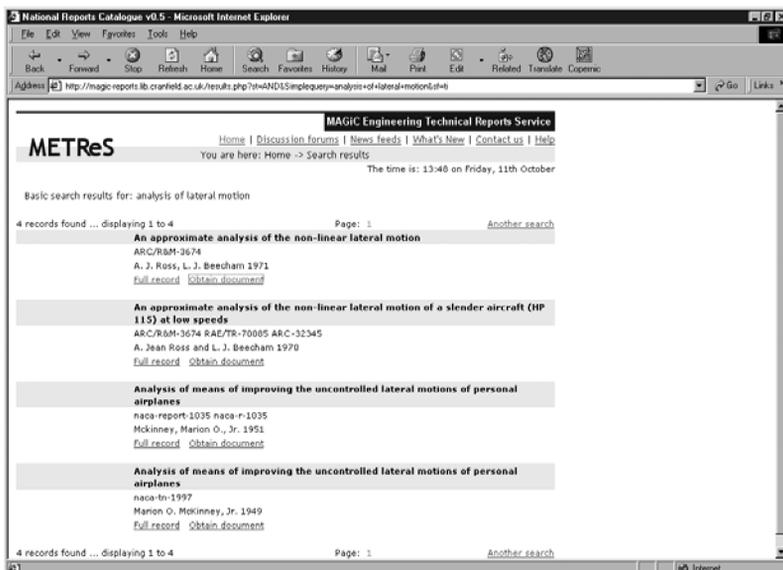- the report series to which that report belongs.
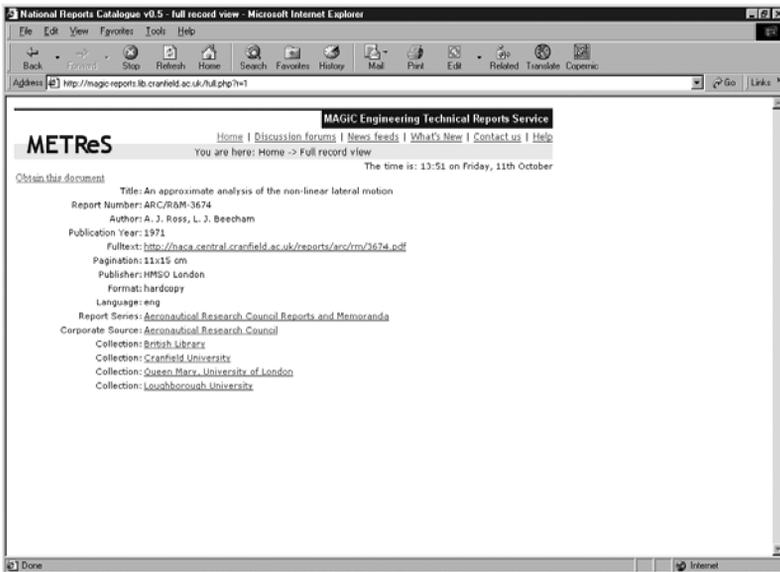

Figure 2. The results page

Figure 3. Full record view

catalogues of the other institutions would not have found the report. However, simply by supplying MAGiC with the report number ranges held in a given report series, those institutions have instantly increased the visibility of reports in their own collections.

However, holdings become visible not only to users, but also to collection managers. By using the advanced search page it is possible to restrict searches to selected corporate sources, reports series or collections. This allows reports production and holdings to be examined from a number of different angles and gives collection managers a standard against which to measure their own holdings, as well as offering the possibility for libraries to harvest records from the NRC to enhance their own local services. It is these facilities that also make METReS a very powerful tool for collection management and rationalisation.

METReS represents a powerful tool for resource discovery, document supply and collection management. As an example, a simple search from the home page for 'analysis of lateral motion' returns four records (see Figure 2).

From this page, it is possible to examine the full record (see Figure 3) and click through to the electronic full-text document, if available. By following the 'Obtain document' link, the user is taken through to a document supply page (see Figure 4). In this case the supplier is the British Library Document Supply Centre (BLDSC), and clicking on the 'Articles Direct' button takes us through to the Articles Direct Order Form, automatically passing across the publication details (see Figure 5).

The full record view (Figure 3) also provides links to information about the corporate source and the report series, and gives a list of known collections that hold the document. The record lists several collections where the report is held: the British Library, Cranfield University, Queen Mary College and Loughborough University. Of these institutions, only Cranfield University has all of its reports individually catalogued. A search in the

OAI in action

The use of the OAI fits in with current mainstream thinking. Metadata harvesting is gaining global acceptance and a growing number
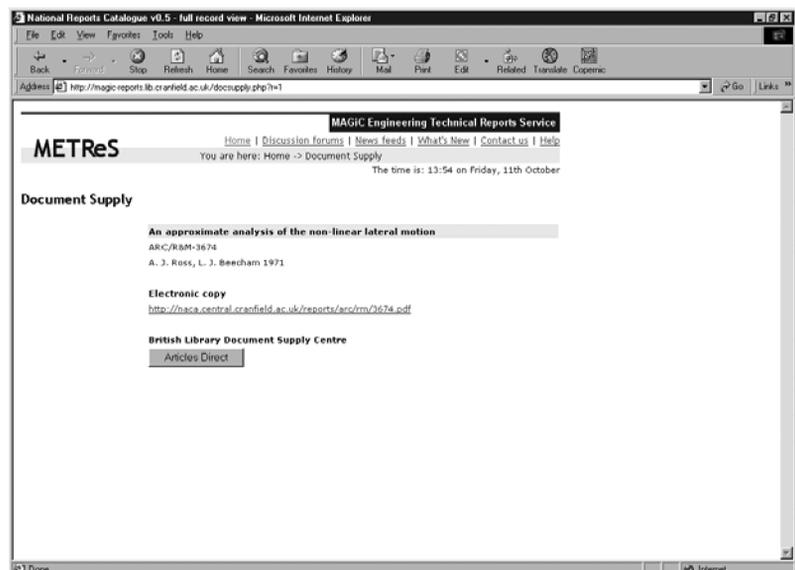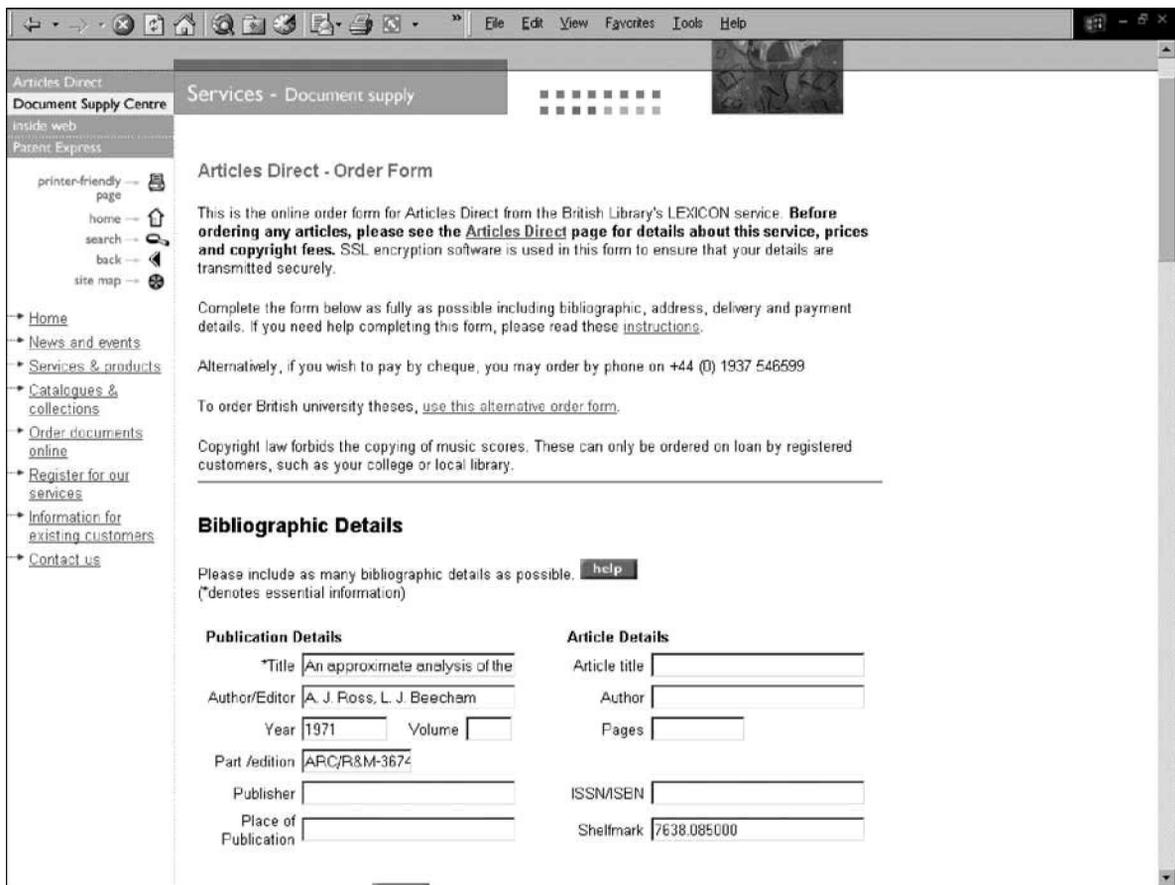

Figure 4. Document supply page

Figure 5. Articles Direct Order Form

of organisations are using the OAI to reveal and transfer metadata. Among those:

NASA has provided OAI layers for several of its reports servers, including the NACA Reports Server [8] and the Langley Technical Reports Server (LTRS), [9] both of which provide links to online versions of reports.

AERADE [10] is the UK Aerospace and Defence gateway, hosted and managed by Cranfield University, which provides access to a collection of key aerospace and defence resources on the Internet. Over a period of two years, information and subject specialists added 2,000 resources to the database. Using OAI metadata harvesting, in February 2002, it took less than two days to add 1,700 new records to the database, taken from the NASA TRS. The use of a scheduled script, run from the AERADE server, allows the Langley server to be interrogated on a regular basis. If new reports have been mounted at Langley since the last time the script ran, it automatically incorporates the new metadata into the AERADE database.

Involvement in the OAI has led to an

agreement between MAGiC and NASA to exchange electronic versions of NACA and ARC reports and this has resulted in the launch of the NACA UK mirror site [11]. In this particular case, the OAI protocol is used as the basis for replication of both metadata and documents.

While MAGiC has used the OAI as the basis for populating the NRC and providing master metadata sets that enhance the use, access and visibility of engineering technical reports in the UK, metadata harvesting has general applicability beyond that.

The eprints.org [12] site, which is dedicated to opening access to the refereed research literature online through author/institution self-archiving, uses OAI as the basis of interoperability between eprints archives.

Significantly, journal publishers are showing a great degree of interest in the potential of metadata harvesting for increasing the visibility of their publications. The Institute of Physics has recently implemented the Open Archives Initiative technical framework for their Electronic

Journals service and metadata records for article abstracts are now available for harvest from their server on request [13].

Summary

OAI metadata harvesting has made it possible for MAGiC to achieve its goals of making technical reports more visible and accessible in the UK. It offers a high level of integration and interoperability between heterogeneous systems and organisations and has allowed the creation of a system that addresses the problems of collection management caused by the complexity of grey literature and the paucity of cataloguing.

Going beyond grey literature, OAI metadata harvesting can be employed in a wide number of contexts, and represents a new approach to exposing resources and making them available at low cost and with minimal effort.

Perhaps most importantly, metadata harvesting, by its very nature, fosters an atmosphere of co-operation, collaboration and active participation. This can be only be of benefit to producers and consumers of information alike.

References

1. Mildren, K. and Hicks, P. (eds) *Information sources in engineering* (3rd edn) London: Bowker-Saur, 1996

2. http://www.magic.ac.uk/

3. http://www.bl.uk/concord

4. http://www.rslp.ac.uk/

5. http://magic-reports.lib.cranfield.ac.uk/

6. http://www.openarchives.org/documents/FAQ.html

7. http://www.openarchives.org/OAI/2.0/ openarchivesprotocol.htm#DefinitionsConcepts

8. http://naca.larc.nasa.gov/

9. http://techreports.larc.nasa.gov/ltrs/ltrs.html

10. http://aerade.cranfield.ac.uk/

11. http://naca.central.cranfield.ac.uk/

12. http://www.eprints.org/

13. http://www.iop.org/Physics/News/0467j