

Delivering OAI records as RSS  
An IMesh Toolkit module for facilitating resource sharing.

Monica Duke  
UKOLN  
University of Bath  
Bath, UK  
+44 1225 384930

M.Duke@ukoln.ac.uk

## **Background**

Subject Gateways act as a main point of access to high-quality resources on the Web. They are resource discovery guides that provide links to information resources that can be whole web sites, organisational home pages and other collections or services, themed around a specific subject, such as the physical sciences or humanities. At their core is a catalogue of rich metadata records that describe Internet resources - subject specialists identify and select the resources and create the descriptions. This high level of manual intervention, coupled with collection management policies and procedures, ensures a high level of quality control. The descriptions can be searched or browsed through a web interface.

Subject Gateways developed in the UK as part of the e-Lib programme [1], whilst similar initiatives are also found throughout Europe [2], in Australia [3] and in the United States of America [4]. In the United Kingdom, Subject Gateways operate under the umbrella of the Joint Information Systems Commission (JISC), and continue to develop within the JISC Information Environment [5].

The IMesh Toolkit project [6] was funded in 1999 under the JISC/NSF Digital Library Initiative [7] to develop tools for subject gateways, with partners in the UK (UKOLN, University of Bath and Institute of Learning and Research Technology, University of Bristol) and in the United States (Internet Scout Project, University of Wisconsin, Madison).

Over the last few years, collaborative efforts between gateways have led to Renardus [2], a pan-European service formed by collaborating gateway services across Europe. The Resource Discovery Network (RDN) [8] is a co-operative national network in the UK that brings together a number of 'hubs' with each hub specialising in a subject area and creating records within its domain. The RDN has taken the approach of using OAI as a mechanism to share records between gateways and provide a central search service [9]. Metadata records created within the RDN are based on the Dublin Core element set. Although some divergence across gateways for particular elements is allowed, there is consistent use of the title, description and identifier fields [10].

## **Motivation**

There has been an increasing interest in embedding subject gateway content into other web sites. The RDN provides a facility, called RDN-Include, which functions

similarly to the Google toolbar, allowing the RDN service to be searched without visiting the central RDN home page. When records are presented to users through frameworks external to the gateway, they become accessible in alternative ways, adding value through re-use. The IMesh Toolkit Project has provided a module that supports the delivery of subject gateway records as a newsfeed, using the RSS standard. By integrating the IMesh Module, subject gateway services will be able to support the export of their records in the RSS format, thus encouraging the sharing of gateway content. Having identified a selection of resources discovered through the subject gateway, the user will be able to generate a file containing an RSS document, or newsfeed, which can then be made available to other users through some presentation system, such as an institutional portal or a course web site.

## RSS

RSS is an XML-based format for sharing content on the web, best suited for list-oriented content such as news items and job listings. An RSS file (also known as an RSS feed or RSS channel) consists of a list of items, each of which contains a title, description and a link to a web page. The link in the RSS file leads to the full content. There is a convenient one to one mapping between the record title, description and identifier in RDN records, and the title, description and identifier of RSS list items.

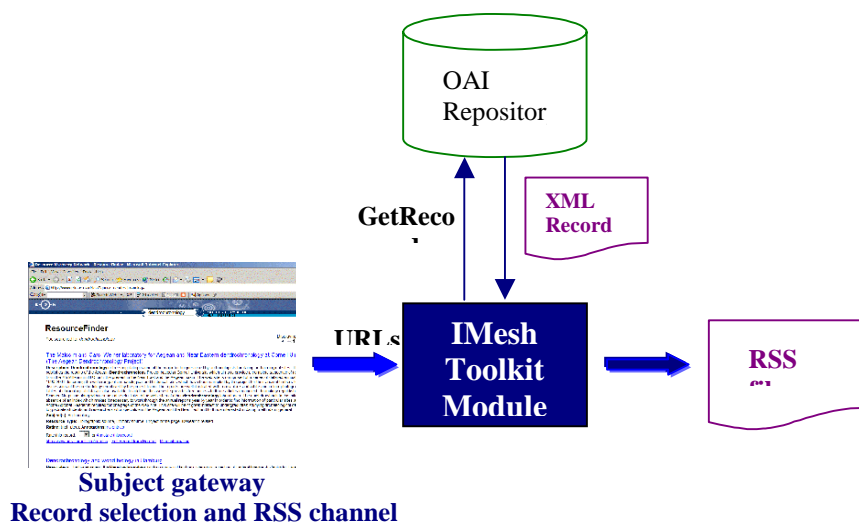
The RSS format can thus be a basis for presenting a reading list, consisting of a subset of gateway records. Each gateway record corresponds to an item in the RSS list, showing the title, description and URL to the full content. A channel title and description can also be added. For further information on RSS see [11]

**Fig.1** The correspondence between record metadata and an RSS item.

<pre> &lt;metadata&gt;   &lt;dc xmlns=<a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a>   xmlns:xsi="<a href="http://www.w3.org/2000/10/XMLSchema-instance">http://www.w3.org/2000/10/XMLSchema-instance</a>"   xsi:schemaLocation="<a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a>   <a href="http://www.openarchives.org/OAI/dc.xsd">http://www.openarchives.org/OAI/dc.xsd</a>"&gt;     &lt;title&gt;Canadian journal of     forest research&lt;/title&gt;     &lt;description&gt;The Canadian Journal of Forest     Research is published monthly by NRC Research     Press, which is part of the National Research Council     of Canada. .... The full text of articles can be     obtained by subscription or by payment per article. An     electronic alerting service is provided. Information     about the journal, the tables of contents, and the     abstracts are available in French.&lt;/description&gt;     &lt;subject&gt;forestry&lt;/subject&gt;     &lt;subject&gt;journals&lt;/subject&gt;     &lt;subject&gt;Canada&lt;/subject&gt;     &lt;language&gt;eng&lt;/language&gt;     &lt;type&gt;Journal / Contents and     abstracts&lt;/type&gt;     &lt;identifier&gt;<a href="http://pubs.nrc-cnrc.gc.ca/Fcgi-bin/Frp2/Frp2_desc_e%3Fcjfr">http%3A%2F%2Fpubs.nrc-cnrc.gc.ca%2Fcgi-bin%2Frp%2Frp2_desc_e%3Fcjfr</a>     &lt;/identifier&gt;   &lt;/dc&gt; &lt;/metadata&gt; </pre>	<pre> &lt;item rdf:about="<a href="http://pubs.nrc-cnrc.gc.ca/cgi-bin/rp/rp2_desc_e?cjfr">http://pubs.nrc-cnrc.gc.ca/cgi-bin/rp/rp2_desc_e?cjfr</a>"&gt;   &lt;title&gt;Canadian journal of forest   research&lt;/title&gt;   &lt;link&gt;<a href="http://pubs.nrc-cnrc.gc.ca/cgi-bin/rp/rp2_desc_e?cjfr">http://pubs.nrc-cnrc.gc.ca/cgi-bin/rp/rp2_desc_e?cjfr</a> &lt;/link&gt;   &lt;description&gt;The Canadian Journal of   Forest Research is published monthly by   NRC Research Press, which is part of the   National Research Council of Canada. ....   The full text of articles can be obtained by   subscription or by payment per article. An   electronic alerting service is provided.   Information about the journal, the tables of   contents, and the abstracts are available in   French. &lt;/description&gt; &lt;/item&gt; </pre>
--	--

## The IMesh Toolkit Module

The IMesh Toolkit module, written in Perl, generates an RSS file that contains a list of reading list materials, made up of a number of subject gateway records. Given a list of URLs each of which retrieves a record which complies to the OAI\_dc schema for GetRecord responses, the module will output an RSS-formatted list of items together with a channel title and description.



**Fig. 2.** Using the IMesh Toolkit Module

It is left up to the individual service to decide how best to design the interface for the selection of the individual resources and the underlying code to create a list of suitable URLs to be submitted to the IMesh Toolkit module. This ensures that to a certain extent the module is independent of the technologies used to deliver the interface of the gateway service.

The RSS file that is generated can be used in a number of ways. The file can be edited using an RSS editor, such as RSSxpress [12] to make changes, add comments or even links to other related resources. The RSS file can then be presented to users through a presentation system that understands RSS e.g. a desktop reader or newsfeed aggregator, such as Feedreader [13]. Alternatively, newsfeeds can very easily be displayed in webpages, by transforming the XML into HTML, for example by using RSSxpress-lite [14].

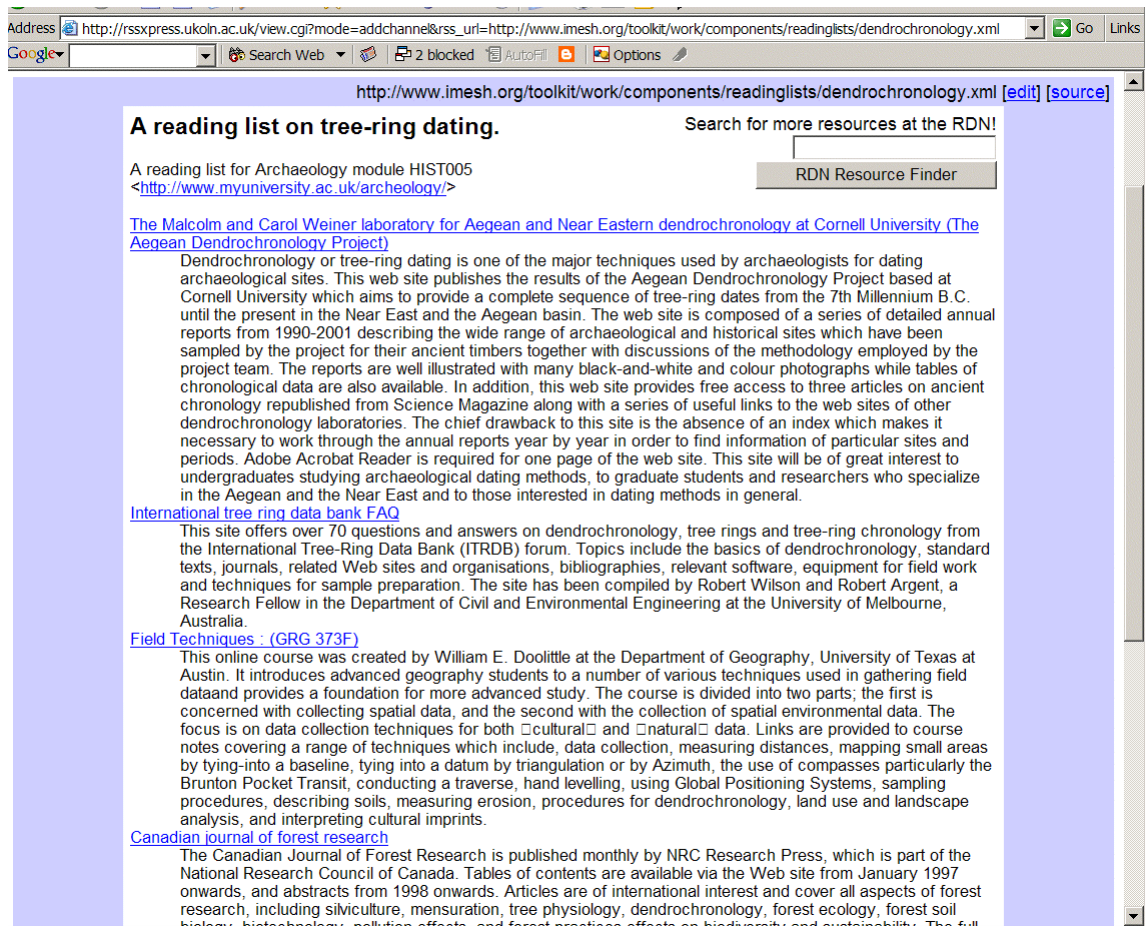


Fig 3. An example of a reading list generated from a search on the RDN, viewed using RSSxpress.

## Interoperability Issues

For the purposes of interoperability, to date the Open Archives Initiative Protocol for Metadata Harvesting requires repositories to disseminate Dublin Core, without any qualification. This provides a baseline of interoperability since the schema [15] mandated by the protocol accommodates the three fields of description that can be reused in the RSS item description (dc:title, dc:identifier, dc:description, equivalent to the RSS item title, item link and item description respectively). However the schema does not mandate any of the fields, therefore there are no guarantees that the fields will be present in all the records returned.

Where no descriptions of resources are available in the record, this makes for a rather limited RSS feed. The user may, of course, add a description after generating the RSS, by using an additional tool, for example an RSS Editor.

A more problematic aspect is the use of dc:identifier. For the purposes of reading lists, the linkage between the record and the identifier of the associated resource is rather important, since the aim of the end-user of the reading list is to access that associated resource. It is clearly a problem if this link cannot be established. With respect to the resource identifier, the OAI-PMH states "the nature of a resource identifier is outside the scope of the OAI-PMH. To facilitate access to the resource

associated with harvested metadata, repositories should use an element in metadata records to establish a linkage between the record (and the identifier of its item) and the identifier (URL, URN, DOI, etc.) of the associated resource. The mandatory Dublin Core format provides the identifier element that should be used for this purpose." We have encountered two examples of difficulties in defining the URL of the resource that the record describes. These are:

- multiple IDs are given in the record, i.e. the dc:identifier field is repeated. RSS items only allow one URL link per item. Where two or more IDs are present in the OAI record, if a URL which contains http in the address is found, this is chosen as the ID to be used in the RSS feed. In the case of two identifiers with http URLs, the choice is currently arbitrary (the first one listed gets picked).
- the dc:identifier is not a URL but a link to the OAI repository identifier

### **Other Issues: Branding and Copyright**

Given the intellectual effort invested by gateways in the creation of metadata records, there may be a reluctance to give away their content, and in particular, to allow that content to appear in a context where the branding associated with the service is lost. On the other hand, it has been argued that RSS can help to drive traffic and increase brand awareness by making content available in a more convenient manner for the user [11]. For example a search box can be included in the news feed (see the RDN example above), which may entice users to carry out further searches and lead them to the central service.

Moreover, gateways may be funded specifically for an intended audience, for which re-use of the records may be restricted. The gateways own the copyright to the content of the records, which carry a statement to this effect, describing acceptable use.

### **Conclusion**

The IMesh Toolkit project has provided a simple module which enables the sharing of subject gateway content (and indeed any other repository which supports the OAI-PMH) by means of another recent and popular XML-based format (RSS). Some requirements for optimal use of the module are that the metadata records contain the dc:title and dc:description elements, and that the dc:identifier field contains a URI that leads to the full resource content.

### **Further Information**

<http://www.imesh.org/toolkit/work/components/readinglists/>

## References

[1] Dempsey, L. (2000), "The subject gateway: experiences and issues based on the emergence of the Resource Discovery Network", *Online Information Review*, Vol. 24 No. 1, pp. 8-23. Also available: <http://www.rdn.ac.uk/publications/ior-2000-02-dempsey/>

[2] [www.renardus.org](http://www.renardus.org)

[3] <http://www.nla.gov.au/initiatives/sg/index.html>

[4] <http://scout.cs.wisc.edu/>

[5] JISC Information Environment Development  
[http://www.jisc.ac.uk/index.cfm?name=ie\\_home](http://www.jisc.ac.uk/index.cfm?name=ie_home)

[6] The IMesh Toolkit Project  
<http://www.imesh.org/toolkit/>

[7] International Digital Libraries Programme  
[http://www.jisc.ac.uk/index.cfm?name=programme\\_nsf](http://www.jisc.ac.uk/index.cfm?name=programme_nsf)

[8] The Resource Discovery network  
<http://www.rdn.ac.uk/>

[9] Andy Powell "An OAI Approach to Sharing Subject Gateway Content"  
Poster paper, 10th WWW conference, Hong Kong. May 2001.  
<http://www.rdn.ac.uk/publications/www10/oaiposter/>

[10] RDN Cataloguing Guidelines  
<http://www.rdn.ac.uk/publications/cat-guide/>

[11] RSS - A Primer for Publishers & Content Providers  
[http://www.eevl.ac.uk/rss\\_primer/](http://www.eevl.ac.uk/rss_primer/)

[12] RSSxpress  
<http://rssxpress.ukoln.ac.uk/>

[13] Feedreader  
<http://www.feedreader.com/>

[14] RSSxpress Lite  
<http://rssxpress.ukoln.ac.uk/lite/include/>

[15] [http://www.openarchives.org/OAI/2.0/oai\\_dc.xsd](http://www.openarchives.org/OAI/2.0/oai_dc.xsd)

**Acknowledgments:** Rachel Heery and Andy Powell of UKOLN for ideas and feedback  
The IMesh Toolkit Project was funded under the JISC/NSF Digital Libraries Initiative II