MARKETING WITH METADATA – HOW METADATA CAN INCREASE EXPOSURE AND VISIBILITY OF ONLINE CONTENT

M. Moffat

The concept of metadata is briefly introduced and the benefits of exposing metadata for publishers and content providers are discussed. Various approaches for standardised exposure of metadata are described including; harvesting via OAT-PMH, distributed searching via Z39.50 & SRU/SRW, and exposure of content for syndication via RSS. Case studies are used throughout to provide illustrative examples of these methods in action and a range of typical concerns relating to metadata sharing issues are addressed.

Introduction

Many producers of online content, such as journal publishers, professional societies and database owners, have descriptive data available, giving details of the actual content they produce. The descriptive data in question does not contain the content itself, but instead consists of records which describe the content. Often, this descriptive data is online at the content providers own website, in the form of a searchable database or ‘catalogue’. This can be a crucial tool which lets searchers, end users or customers discover, and ultimately decide whether to access, the actual content. The content itself may be freely available, available to subscribers or available via pay-per-view – it does not matter.

Now, imagine that same descriptive data (but not the content itself) also being accessible and searchable from a varied range of other websites, with appropriate links back to the original content provider’s web site. This would result in more eyeballs, more hits, more traffic, and ultimately increased exposure and visibility of the actual content to a wider audience.

‘Metadata’ is the accepted term used to describe information about content, and the essence of this document concerns ways to market with metadata. This document focuses on illustrating and explaining the benefits of achieving this by exposing metadata via standard and interoperable means. In a nutshell, this document introduces the means by which content providers can share, or embed, their descriptive data (metadata), with other websites, in standard and reusable ways.

It is primarily intended for a non-technical audience, who require an overview in order to allow them to make decisions regarding the best means of exposing metadata. However, the document does provide illustrative case studies and links to technical specifications, which will provide useful starting points for those tasked with actually implementing the exposure of standardised metadata.

If you wish to increase the visibility and exposure of your content in the online environment, then read on!
Benefits of Exposing Metadata

Before defining metadata, let us outline why exposing it is desirable and has tangible benefits for publishers and content providers. Exposing metadata in standardised ways can:

- Allow your content to be found from a large number of locations (e.g. portals, aggregators, search engines).
- Allow aggregators to expose and thereby help to promote your material in novel ways.
- Enhance the visibility and awareness of your available resources.
- Be a useful way to expose material to new markets.
- Allow potential users to determine the relevance of resources without having to first access them.
- Facilitate the production of interoperable services.
- Improve the visibility of your content in search engines, such as Google, Google Scholar and Yahoo.
- Drive traffic and business to websites.

**FIGURE 1**
Exposing Metadata can enhance the visibility of content
There are many advantages to exposing metadata. If you want people to be able to find your content, then exposing your metadata in standardised ways makes real sense. This is equally true for data providers with content they wish to give away freely, and also for those who wish to charge for their content or restrict access to registered users. Exposing your metadata increases the visibility and awareness of these resources, whether users are expected to pay to access the actual content or not. Increasingly, many types of data providers (small and large publishers, libraries, government agencies, professional bodies, and companies of all sorts) are providing some standard means to access their metadata. For example:

NASA provides metadata records from their technical reports via harvesting. As a result, the NASA Technical Reports Metadata can now be searched from a large number of websites including: Scirus, OAIster, Science Research, AERADE, SAIL-eprints.

Institute of Physics (IOP) provides access to journal article metadata records via distributed searching. As a result, the IOP Journal Archive can be searched from a wide variety of locations (e.g. Library Portals, Aggregators, Gateways), for example TechXtra, PSLgate, MetaLib at the University of East Anglia. (See Case Study – Implementation of Z39.50 at Institute of Physics Publishing.)

Inderscience is a commercial publisher which provides access to journal article metadata records via harvesting, in order to drive traffic to their subscription-based/pay-per-view content. Inderscience found that it was relatively easy to set up a harvesting interface, and that this has been a successful way to drive users to their content (see Case Study – Implementation of an OAI Repository at Inderscience).

COPAC provides free access to the merged metadata catalogues of 24 major university research libraries in the UK plus the British Library, the National Library of Scotland, and the National Library of Wales. COPAC provides access via distributed searching (Z39.50). This allows COPAC to be searched via other sites, such as the Library of Congress and Die Deutsche Bibliothek.

Increasingly, universities and institutions are establishing institutional repositories which permit harvesting of metadata records. At the time of writing, the Registry of Open Access Repositories (ROAR) contains over 600 such repositories, and the Directory of Open Access Repositories (OpenDoar) lists over 300. The ARROW project, which harvests metadata from Australian institutional repositories, is a good example of a service which aggregates institutional metadata to good effect.

Sometimes people don’t quite understand that exposing metadata actually makes them more visible and drives traffic to their site. In today’s world, it is sharing metadata, not creating a monopoly of data, that makes you more valuable. (Eric Lease Morgan ? Head, Digital Access and Information Architecture Department University Libraries of Notre Dame)

Discoverability is very important to the Institute of Physics, as we want to encourage as many users as possible to visit our web site and read the articles.
published in our journals. By exposing our metadata through a variety of means – Z39.50, OAI-PMH and RSS, as well as search engine indexing – we can reach new and existing readers, attract new authors, and increase the profile of the journals we publish. (Judith Barnsby – Senior Product Manager, Institute of Physics Publishing)

What is Metadata?

Metadata is literally ‘data about data’; information that describes an object, but is not the object itself. The NSDL Metadata Primer\(^1\) defines metadata as ‘structured, standardised descriptions of resources, whether digital or physical, that aid in the discovery, retrieval and use of those resources’.

A library catalogue record is an example of a metadata record which, in its simplest form, might contain details about the title, author and date of publication of an item. Metadata can be produced for all sorts of objects (e.g. books, journals, images, learning materials, etc) and a number of metadata standards/schemas have emerged which attempt to standardise the metadata for various types of materials (e.g. MARC for materials in library catalogues, MPEG for images, LOM for learning materials, etc). Metadata, therefore, allows a precise and standardised way of describing content in discrete packages called metadata records.

The Dublin Core is a metadata standard for describing a range of digital objects, and contains a set of 15 metadata elements (e.g. Title, Creator, Subject, Description, Publisher, Contributor, Date, etc.). Dublin Core is important as it is often mandated as a minimum metadata requirement.

A simplified example of a Dublin Core (dc) metadata record describing this article is included below.

```xml
<record>
  <metadata>
    <dc:title>‘Marketing’ with Metadata</dc:title>
    <dc:creator>Moffat, M.</dc:creator>
    <dc:publisher>Heriot Watt University</dc:publisher>
    <dc:date>2006</dc:date>
    <dc:format>HTML</dc:format>
    <dc:language>eng</dc:language>
    <dc:description>this document introduces the means by which content providers can share their metadata in standard and reusable ways.</dc:description>
    <dc:identifier>http://www.icbl.hw.ac.uk/perx/advocacy/exposingmetadata.htm</dc:identifier>
  </metadata>
</record>
```
Often, content providers already produce metadata for their materials, although this may be in a non-standard proprietary format. In many instances, this proprietary metadata can simply be mapped to a standard scheme, such as Dublin Core, rather than having to be recreated from scratch.

Further Information

Metadata demystified: a guide for publishers: http://www.niso.org/standards/resources/Metadata_Demystified.pdf

Understanding Metadata

A general introduction to metadata, from NISO – the US National Information Standards Organisation: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

Dublin Core http://dublincore.org/

Brief description of the 15 Dublin Core Elements: http://dublincore.org/documents/dces/

Why Adopt a Standardised Approach?

It may seem reasonable to ask ‘If it is all about sharing content why cannot we just provide links to our content?’ Providing links to your content does not allow information about it to be shared and re-purposed easily and in a standard way. The beauty of exposing metadata in a standard way is that little effort is required for third parties to reuse your metadata, and make it available to their visitors. Standard metadata is, therefore, an investment in current and future interoperability.

There are a few large organisations, such as Google and Amazon, which have the muscle to impose non-standard interfaces for exposing metadata (e.g. the Google Search API). Few others do, hence the need for standards. Without adopting a standardised approach to exposing metadata, the result is hundreds, if not thousands, of different non-standard interfaces, each of which requires an individually tailored means of implementation.

In the UK, content providers are encouraged to expose their metadata via standard means in order to make their systems interoperable within the Joint Information Systems Committee (JISC) Information Environment² (Powell 2002). In the USA and Australia, developments, such as ‘search engine bridges’ (Suber 2004), are improving search engine indexing of certain types of exposed metadata. This may ultimately enhance the ranking of these materials in services, such as Google, Google Scholar and Yahoo.

How can Metadata be Exposed?

Broadly speaking there are three different ways in which standardised exposure of metadata can be achieved.
1. Exposing Metadata via harvesting.
2. Exposing Metadata via distributed searching.
3. Exposing content for syndication.

1. Exposing Metadata via Harvesting

When metadata is exposed for harvesting, it can be collected by third parties and subsequently used by them to provide services (e.g. searching or browsing the metadata records). The Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) is the most established means of exposing metadata for harvesting (see Section 6). The terms ‘data provider’ and ‘service provider’ have specific meanings within the OAI community. A data provider maintains one or more repositories which support OAI-PMH. A service provider harvests metadata using OAI-PMH and uses the metadata as a basis for building value added services (Figure 2). Typically, service providers harvest multiple repositories and use them to provide services targeted at specific communities or audiences.

2. Exposing Metadata via Distributed Searching

When exposing metadata via distributed searching, a standardised search interface is produced by content providers (Figure 3). Remote service providers, such as portals or aggregators, can then query this interface and receive back search results in a standard reusable format. Z39.50 and SRU/SRW are protocols commonly used to implement distributed searching (see Section 7). Typically, service providers use distributed searching to query a number of data providers at one time – this is known as ‘federated searching’.

Generally speaking, harvesting is a simpler and more straightforward option than distributed searching. Once metadata is harvested, service providers do not require constant interaction with data providers (although periodic
reharvesting of metadata is necessary to keep it up to date). The harvested metadata is stored locally and used repeatedly by service providers. Distributed searching, however, requires a more complex and ongoing interaction. Here, every search request generates a new query to the data provider database and the transfer of search results. The metadata transferred in a distributed search is not usually stored, and is used only transiently for the duration of a single search.

3. Exposing Content for Syndication

Metadata about recently produced ‘new’ content can be exposed for use by third parties via a file format known as Rich Site Summary (RSS). Put simply, RSS is a standard format (XML) for sharing topical and timely content, such as news items, job adverts, or event announcements. RSS can be a useful complementary mechanism to harvesting or distributed searching, in that it presents an additional means by which interested parties can gain access to only the latest newest content (see Section 8).

Exposing Metadata via Harvesting

OAI repositories

OAI-PMH is a simple protocol that allows data providers to expose their metadata for harvesting (Figure 2). It supports the regular gathering of metadata from one service to another. OAI-PMH is based on common underlying Web standards – HTTP, XML and XML schemas – which makes it fairly easy to implement for services already running a web server.

OAI-PMH is widely used for eprints archives and has its roots in the ePrints community. However, the concepts of OAI-PMH – exposing multiple forms of metadata through a harvesting protocol – can be applied to a wide range of digital materials, for example: images, learning materials, assessment materials, technical reports or catalogue records. For the sake of enabling the maximum, broadest level of interoperability, OAI-PMH mandates that metadata should be exposed as Dublin Core. However, it is important to appreciate that the protocol

FIGURE 3
Exposing Metadata via Distributed Searching

OAI repositories
enables multiple forms of metadata to be exposed. These alternative forms of metadata can be as rich as necessary to suitably describe content (e.g. IEEE LOM for learning materials, MARC for library records, etc.). Thus, any form of metadata can be exchanged using OAI-PMH as long as it can be encoded as XML, and a suitable XML schema is created.

Case Study – Implementation of an OAI Repository at Inderscience

‘Inderscience’s motivation was commercial in nature – to make their metadata available to aggregators and drive more users to their full text subscription based materials’

Inderscience is a medium-sized commercial journal publisher, covering engineering, technology, and management and business administration. As part of a JISC/PALs funded project, in 2004 the company was involved in an exercise to create an OAI-PMH compliant repository. Their motivation was commercial in nature, in that they wished to ‘make their metadata available to aggregators and drive more users to their full text subscription-based materials’.

Inderscience viewed the opportunity to get as much information as possible about their content into the public domain, essential to their commercial success. Dissemination of metadata via OAI-PMH offered the possibility of users being able to discover Inderscience content from a wide variety of locations, allowing the company to play to its strengths and compete on a more level playing field with some larger competitors.

‘Experience at Inderscience revealed that it was relatively easy to set up an OAI repository’

Experience at Inderscience revealed that it was relatively easy to set up an OAI repository, the process involving approximately 30 hours of technical time. The publisher’s current Content Management System (CMS) offered a well-structured database, which required only minor modification in order to support OAI harvesting. The process involved installation of OAI data provider’s software on the publisher’s web site and integration with the publishers CMS in order to populate the repository with the necessary metadata elements (Dublin Core). The implementation required knowledge of XSLT transformations, PHP, Java servlets and MySQL. A full report on the Implementation of the Inderscience OAI Repository is available.

Since being involved with the project, Inderscience has further developed their OAI repository which is now publicly available. Their experience of implementing an OAI repository has been positive and instrumental in establishing partnerships with a number of content aggregators. In some instances, OAI-PMH has been used to exchange metadata, and in others, Inderscience have elected to utilise a non-standard approach (e.g. their metadata exchange with Google Scholar). Service providers which aggregate data from the Inderscience OAI repository include: TechXtra, Collection of Computer Science Bibliographies and Research Papers in Economics.

It is perhaps salient to point out that large commercial aggregators (such as Google Scholar, Yahoo, CSA etc.) may prefer to utilise their own proprietary
means for collecting metadata, and this has certainly been the experience at Inderscience. The establishment of one-to-one metadata exchanges between aggregators and content providers benefits aggregators as it may enable them to establish a competitive advantage in the marketplace. However, for content providers, this can ultimately result in a proliferation of bespoke agreements between a large number of different aggregator services, some large and some small. While it may be necessary to implement bespoke solutions for the benefit of the large commercial aggregators, OAI-PMH remains an effective mechanism to expose multiple forms of metadata, and is a good candidate to enable interoperability with the broadest possible audience.

Further Information
Open Archives Initiative: http://www.openarchives.org/
OAI For Beginners – Tutorial: http://www.oaforum.org/tutorial/

Static OAI Repositories

A static OAI repository is a low tech, low barrier alternative to the provision of a fully fledged OAI repository. A static repository provides a simple approach for exposing relatively small and fairly static collections of metadata records via OAI-PMH. In a nutshell, a static OAI repository is a single file which contains all of the metadata records in OAI Dublin Core format. The file is registered at a remote static repository gateway which takes care of dealing with the OAI-PMH requests serving of the metadata. It is possible to create static OAI repositories with little effort and low cost, as the only requirements are: (a) the creation of a single XML file which is accessible on the Web, and (b) registration of this file at a suitable static repository gateway.

The static OAI repository approach is suitable for relatively small and unchanging collections (e.g. under 5,000 metadata records in which the metadata does not require frequent amendment, addition or deletion). Examples of collections might be those which are created by one-off digitisation programmes, or those which are infrequently updated, perhaps on a monthly basis.

Case Study – Implementation of a Static OAI Repository for the Springburn Virtual Museum

‘existing metadata has been exposed for harvesting at low additional cost’

As part of a one-off digitisation project, the Glasgow Digital Library digitised a selection of photographs from the Springburn Community Museum, resulting in the Springburn Virtual Museum. In order to widen access to these digitised materials, a static OAI repository was produced. The static repository is simply an XML file containing the metadata records from the digitised photos.

A simple script was created to generate the static repository file from an existing Microsoft Access database. The file was made available on the Web and registered with a Static Repository Gateway hosted at the OAI Scotland Information Service. The gateway enables access to the static repository as if it
was a standard OAI repository (see OAI-Identify Request), and existing metadata has thus been exposed for harvesting at low additional cost.

Further Information
Specification for an OAI Static Repository and an OAI Static Repository Gateway: http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm

Exposing Metadata via Distributed Searching

Z39.50

Z39.50 is a client/server protocol for searching and retrieving information from remote computer databases. It specifies procedures and formats for a client to search a database provided by a server, retrieve database records, and perform a number of related information retrieval functions, such as sort and browse. Z39.50 is commonly used in the library environment, where it has its developmental roots, but is also widely used in other sectors. Z39.50 is an ANSI/NISO standard.

Initial work on Z39.50 began in the 1970s, and there have been a number of versions of the protocol, the current one being Z39.50-2003, which incorporates several amendments and clarifications on the previous version (Z39.50-1995). Z39.50 has been criticised for being overly complex, difficult to implement and a ‘pre-web’ technology, as well as the fact that much of the functionality that it offers is limited by differences in the way it has been implemented by developers and commercial vendors. The Bath Profile was produced in an effort to improve the situation. The Bath Profile describes and specifies a subset of Z39.50. It clearly specifies the search syntax which should be used for common bibliographic searches, and details the expected behaviour of Bath-compliant servers in response to these searches.

A number of initiatives, known collectively as ZING (Z39.50 International: Next Generation), are attempting to take Z39.50 forward into the current web HTTP environment. The most important of these are probably the SRU/SRW protocols, which drop the Z39.50 communications protocol in favour of HTTP and XML. SRU/SRW, is considerably simpler than the original Z39.50 protocol, and is, therefore, more readily amenable to a wider number of developers.

Case Study – Implementation of Z39.50 at IOP

The IOP is an international professional body and learned society, established to promote the advancement and dissemination of physics. IOP Publishing is an integral part of the Institute, with an active publications programme covering over 40 established journals, and is at the forefront of electronic publishing developments.

IOP chose to adopt Z39.50 for distributed searching, as this standard is used extensively in the UK and in Europe, and is recommended by the JISC. Implementation began in September 2002, and took around three weeks (not
A third party Perl module, called Z3950::simpleserver, was used, and the deployment itself was relatively straightforward with few problems. An initial dilemma of deciding which format to use for the returned results was resolved by selecting functional area C of the Bath profile upon the recommendation of UKOLN. The initial IOP Z39.50 implementation was tested from mid to end October by several universities, and subsequently released as a live Z39.50 gateway service.

Developing a Z39.50 gateway has allowed the IOP to be searched from within library portals, such as MetaLib, and in services, such as EEVL Xtra, providing an alternative route to IOP content. In addition, IOP have found that librarians appreciate early adoption of standards and protocols, which helps cement their relationships with these key customers.

IOP are now considering the implementation of SRU as the next step in their discoverability strategy.

Further Information

Z39.50 International Maintenance Agency: http://www.loc.gov/z3950/agency/
Z39.50-2003 Application Service Definition and Protocol Specification (236 Pages): http://www.loc.gov/z3950/agency/Z39-50-2003.pdf
The Bath Profile: http://www.collectionscanada.ca/bath/bp-current.htm
ZING (Z39.50 International: Next Generation) : http://www.loc.gov/z3950/agency/zing/
Z39.50: Part 1 – An Overview. Part 2 – Technical Details: http://www.biblio-tech.com/html/z39_50.html; http://www.biblio-tech.com/html/z39_50_part_2.html

SRU/SRW
(Modified Version of ‘SRW and SRU in Five Hundred Words or Less’ by Morgan (2004). Reproduced with permission.)

SRU (Search/Retrieve via URL) and SRW (Search/Retrieve Web Service) are search protocols for querying databases over the internet and returning search results. Both protocols are maintained by the Library of Congress. The development of SRU and SRW has been informed by over 20 years of experience with the Z39.50 protocol, and SRU/SRW build upon the important aspects of this predecessor, but are less complex, easier to understand and simpler to implement.

Basic ‘Operations’

Both SRU and SRW protocols define only three basic operations: explain, scan, and searchRetrieve.

Explain – Explain operations are requests sent by clients as a way of learning about the server’s database. At a minimum, responses to explain operations return the location of the database, a description of what the database contains, and what features of the protocol the server supports.
Scan – Scan operations are processes for enumerating the terms found in the remote database’s index. Clients send scan requests and servers return lists of terms. The process is akin to browsing a back-of-the-book index where a person looks up a term in a book index and ‘scans’ the entries surrounding the term.

searchRetrieve – SearchRetrieve operations are the heart of the matter. They provide the means to query the remote database and return search results. Queries must be articulated using the Common Query Language (CQL). The results of searchRetrieve operations can be returned in any number of formats, as specified via explain operations. Examples might include structured but plain text or data marked up in XML vocabularies, such as Dublin Core.

Differences in Operation

SRW and SRU are essentially ‘brother and sister’ protocols which achieve the same ends. The differences between them lie in the way operations are encapsulated and transmitted between client and server, as well as in how results are returned.

SRW – client requests and server responses are returned as SOAP messages (HTTP is not used).

SRU – client requests are encoded as name/value pairs in the query string of a URL and are transmitted via HTTP. The results of SRU requests are simple XML (i.e. the same content as a SRW result without the SOAP ‘wrapping’).

Case Study – Implementation of SRU at JSTOR

‘SRU provides a way for the archive to be available to users that wish to rely on these search alternatives’

JSTOR (Journal STORage) is a digital archive of over 150 core scholarly journals, which start with the very first issues, i.e. complete back runs. The collection covers material from the 1800s up to a ‘moving wall’ of between one and eight years before current publication. It covers a number of subjects, particularly in the Humanities and Social Sciences. JSTOR started life as a pilot project in the USA, funded by the Andrew W. Mellon Foundation. The aim was to provide a solution to the increasing costs to libraries of storing back runs of journals. The project proved so successful that a full service was launched in 1997. JSTOR is now an independent, non-profit making organisation.

To increase the convenience of access to the archive, the ‘JSTOR XML Gateway’ was developed in 2005. The gateway uses SRU, and search results are returned as XML in Dublin Core format. The JSTOR XML Gateway was designed using SRW/U Open Source Software, developed by Online Computer Library Center (OCLC) Research. (see http://www.oclc.org/research/projects/webservices/ for further information).

This new approach makes it possible for search requests to be received, translated, and their result sets delivered within an agreed-upon framework, while eliminating the main shortcomings associated with earlier methods. For JSTOR, it provides a way for the archive to be available to users that wish to rely on these search alternatives and gives a greater level of assurance that the results returned for JSTOR by the metasearch engine will be accurate.3
The use of SRU has allowed JSTOR to expand its relationships with publishers, libraries, and other aggregators. In 2005 alone, 14% of all article views in JSTOR were a result of links from other websites. They reported ‘We expect this figure will grow over time as our efforts to enable pathways for users continue’.³

Further Information
SRU Homepage at the Library of Congress: http://www.loc.gov/standards/sru/
SRW Homepage at the Library of Congress: http://www.loc.gov/standards/sru/srw/

Other Emerging Standards

‘Over the last few years we have seen a general trend towards a simplification of search interfaces from complex standards such as Z39.50, through simpler revisions of it in the form of SRW and SRU leading ultimately to very simple proposed standards such as the Amazon A9 OpenSearch specification’ (Powell 2005). The A9 OpenSearch specification builds on a simple set of HTTP CGI parameters and the RSS format and is, arguably, simple enough to implement for the vast number of content providers. This approach is currently being considered by the NISO Metasearch Initiative as an even simpler means of implementing distributed searching than SRU/SRW.

In short, standards evolve and other, possibly even simpler, alternatives may appear in the future.

Further Information
A9 OpenSearch: http://opensearch.a9.com/
NISO Metasearch Initiative: http://www.niso.org/committees/MS_initiative.html

Expositing Content for Syndication

Dictionary definitions of ‘syndication’ are along the following lines:
‘Distributing a news article or picture through a syndicate for publication in a number of newspapers or periodicals simultaneously’
‘Syndication’ is often used in the context of RSS, because this file format is all about exposing content for reuse.

RSS

Put simply, RSS is a format for sharing content easily on the Web. What type of content? Commonly things such as news items, job adverts, or details of latest publications are ideal candidates for RSS, although almost any list-orientated information can be suitable. RSS is adequately described elsewhere (Moffat 2003; Nottingham 2003) but the key points include:
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. Typically, this is metadata, with the full content itself being available separately and accessed by the link in the RSS file.

Utilising an RSS feed is straightforward. Once an RSS file is made available on a web site, interested parties can simply gather the file from the site and reuse the content in a variety of ways.

There are a number of different versions of RSS. Although this has led to some confusion and debate, it is not a major issue, as almost all RSS applications can handle any RSS version.

RSS can stand for ‘Rich Site Summary’, ‘RDF Site Summary’ or ‘Really Simple Syndication’ depending on who you ask, and which version they are speaking about. RSS is an XML-based format.

To sum up, RSS allows potential users to see some of a content provider’s information without actually having to visit their site directly. As Nottingham (2003) explains:

Imagine that your company announces a new product or feature every month or two. Without an RSS feed, your viewers have to remember to come to your site and see if they find anything new – if they have time. If you provide a feed for them, they can point their aggregator or other software at it, and it will give them a link and a description of developments at your site almost as soon as they happen. News is similar; because there are so many sources of news on the Internet, most of your viewers won’t come to your site every day. By providing an RSS feed, you are in front of them constantly, improving the chances that they’ll click through to an article that catches their eye.

RSS is a very flexible format, and is now widely used to expose metadata about just about any type of online content, e.g. news headlines, job adverts, press releases, conference or events announcements, new book listings, journal tables of contents, marketing communications, product announcements, service announcements, tender opportunities, web logs (blogs), audio or video (podcasting).

Example – RSS in Action, OneStep Jobs and OneStep News

There are a number of ways in which RSS can be utilised, e.g. personal desktop readers, online aggregators or as part of other applications, such as web browsers and email clients. OneStep Jobs and OneStep News are online aggregator services which clearly illustrate the concept of RSS aggregation. These services collect RSS feeds from a number of content providers in the subject areas of engineering, mathematics and computing, and provide access via a searchable and browsable interface. The OneStep Services have proved to be popular within their particular niche, and the contributing content providers gain by increased exposure of their resources targeted to a specific community.

Further Information

RSS – A Primer for Publishers & Content Providers: http://www.eevl.ac.uk/rss_primer/
Common Questions Answered

Content providers may have legitimate concerns about sharing metadata about their prized content. Some of the most commonly expressed questions are addressed below.

If it's all about sharing content why can’t we just provide you with a link to our content?

Simply providing a link to your content does not allow it to be shared and re-purposed easily and in a standard way. The beauty of exposing metadata in a standard way is that little effort is required for third parties to reuse your metadata and make it available to their visitors.

I don’t like the thought of giving away our content for others to use.

Exposed metadata usually only contains a brief description of the actual content – just enough to generate interest in potential users. These users will be directed back to your site by links in the metadata in order to access the full content in the normal way (i.e. freely available, subscription based, pay-per-view, etc).

If I make metadata available, am I losing control over my look-and-feel?

Yes and no! Yes, you are giving away a little of your metadata and letting other services present it within their own look-and-feel. However, the overall result should be that more people subsequently visit your pages and see the complete content in the way you intended.

If I make my metadata available, will there be a loss of traffic to my site?

No, the overall effect of exposing metadata is that it will actually drive traffic to your site. By exposing your metadata records for use by other services, you are allowing people to find your content at other, independent web sites. In theory, this could result in fewer hits to your home page, but a far larger number of hits to your actual content, as a result of new users going directly to your resources from independent web sites.

Will exposing my metadata mean that it is indexed by search engines, such as Google or Google Scholar?

This depends on how your metadata is exposed and the indexing approaches taken by individual search engines. Exposing metadata via OAI certainly can improve ranking in search engines. ‘A normal Google or Google Scholar search favours OAI-repository material and normally ranks it higher than an individual’s own website’. Recent developments, such as ‘search engine-OAI bridges’ are improving search engines indexing of OAI compliant repositories (Suber 2004). Many OAI repositories are now indexed by a number of search engines, e.g. Cogprints, a repository for cognitive sciences, is indexed by Google, Google Scholar, Yahoo, Scirus and Citebase.
Why can’t I simply make my content available to Google and let people find my stuff that way?

You can, and in many cases this will be a perfectly appropriate thing to do. This is particularly true for freely available full text resources. However, in some cases, for example where most of your resources are not text-based, exposing them to Google may not help much. In other cases, you may not want to make the full content freely available. In these situations, exposing metadata may be more appropriate. By making your metadata freely available, you can allow people to discover your resources more readily.

If I want to expose metadata, I can do it via harvesting or distributed searching. How do I decide which of these options to go for?

A good question . . . and not an easy one to answer! There are some practical considerations. If you don’t want to give away your metadata, then make it available for searching using Z39.50 or SRU/SRW. If you are happy to make your data available for harvesting, then implementing OAI may well be simpler – it is certainly a less complex protocol. With OAI it is also easier to share multiple metadata formats. Some content providers choose to expose content via a number of different mechanisms, for example the Institute of Physics archive can be searched via Z39.50 or harvested using OAI-PMH.

Can I tell who is accessing my metadata?

When exposing metadata via distributed searching it is possible to monitor and analyse the usage of your metadata. The situation when exposing content via harvesting is more complex. It is possible to monitor who is harvesting your metadata, however, this is far from the whole picture – a single service provider can harvest your metadata and make it viewable to thousands of end users. More sophisticated techniques can be employed to help track harvested metadata usage, for example adding parameters to the URLs within metadata can allow a webmaster to monitor the hits to specific pages.

If I use the OAI-PMH, does that mean I have to make all my metadata freely available to all service providers?

No, not necessarily. The ‘open’ in OAI does not mean freely available. Data providers can choose to restrict who can gather metadata records from them (e.g. via IP Authentication).

Exposing metadata via OAI-PMH does not seem to be sufficient for our needs. What can we do?

Remember OAI-PMH can be used to expose multiple forms of metadata. OAI Dublin Core is mandated as a simple format providing baseline interoperability which must be provided. But there are a number of reasons why it may not be sufficient to expose only Dublin Core. For example, the Dublin Core elements may not include enough of the elements you need, the Dublin Core elements may not be sufficiently precise or you may simply require a much richer metadata format. Using OAI-PMH, multiple alternative forms of metadata can be provided, and these can be as rich as is necessary to suitably describe content (e.g. IEEE LOM for learning materials, MARC for Library records, etc.). Any form of metadata
can be exchanged using OAI-PMH as long as it can be encoded as XML and a suitable XML schema is created.

Is it possible to licence the usage of exposed metadata?

Using OAI-PMH it is possible to explicitly associate a licence with each of the metadata records which are exposed. The OAI-PMH Rights Expression Specification (Lagoze et al. 2005) provides a general framework for stating rights about metadata records. As illustrated in the simplified code snippet below, this is achieved using optional \(<about>\) and \(<rights>\) tags.

Service providers should look for, and abide by, any rights expressions in the records which they harvest, although the protocol does not actually provide a mechanism for enforcing licence restrictions.

It sounds quite complicated – will it be a lot of work for our techies?

It depends! This document introduces a number of means of exposing standardised metadata (e.g. via OAI-PMH, Z39.50, SRU/SRW), some of which are relatively simple and others which are more complex.

```xml
<record>
  <metadata>
    <dc:title>‘Marketing’ with Metadata </dc:title>
    <dc:creator>Moffat, M. </dc:creator> [...Etc.]
  </metadata>
  <about>
    <publisher>Heriot Watt University </publisher>
    <rights>Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License </rights>
  </about>
</record>
```

Exposing Metadata is truly a win – win for all involved. If you’ve a concern or question that we have not covered here, we would like to hear it! email m.moffat@hw.ac.uk

**Acknowledgements**

This document has been produced as part of the JISC funded PerX project. Thanks go to Inderscience, Terry Hulbert and Judith Barnsby (Institute of Physics), Mike Chrimes (Institution of Civil Engineers), Phil Barker (CETIS), Roddy MacLeod and Santy Chumbe (Heriot Watt University) for their input and valuable comments. Thanks also to Eric Lease Morgan (University Libraries of Notre Dame) and Andy Powell (Eduserv) for permission to reproduce their materials within this document.
Notes

1. NSDL Metadata Primer:  &lt;http://metamanagement.comm.nsdlib.org/outline.html&gt;
2. JISC Strategic Activities – Developing an Information Environment:  &lt;http://www.jisc.ac.uk/index.cfm?name=about_info_env&gt;
3. JSTOR Interface Enhancement. JSTORNEWS 9(3), October 2005:  &lt;http://www.jstor.org/news/2005.10/interface.html&gt;
4. SHERPA (2006). Fifteen Common Concerns – and Clarifications:  &lt;http://www.sherpa.ac.uk/documents/15concerns.html&gt;

References

LAGOZE, C. et al. 2005. Conveying rights expressions about metadata in the OAI-PMH framework:  &lt;http://www.openarchives.org/OAI/2.0/guidelines-rights.htm&gt;

MOFFAT, M. 2003. RSS – A Primer for Publishers and Content Providers:  &lt;http://www.eevl.ac.uk/rss_primer/&gt;

MORGAN, E.L. 2004. SRW and SRU in Five Hundred Words or Less. D-Lib Magazine 10(5):  &lt;http://www.dlib.org/dlib/may04/05sinbrief.html#MORGAN&gt;

NOTTINGHAM, M. 2003. RSS Tutorial for Content Publishers and Webmasters:  &lt;http://www.mnot.net/rss/tutorial/&gt;

POWELL, A. 2002. 5 step guide to becoming a content provider in the JISC Information Environment. Ariadne 33:  &lt;http://www.ariadne.ac.uk/issue33/info-environment/intro.html&gt;

POWELL, A. 2005. The JISC Resource Discovery Landscape – A personal reflection on the JISC Information Environment and related activities. UKOLN, University of Bath:  &lt;http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/resource-discovery-review/ &gt; ;  &lt;http://www.jisc-resource-discovery-landscape.pdf&gt;

SUBER, P. 2004. The case for OAI in the age of Google. SPARC Open Access Newsletter (73):  &lt;http://www.earlham.edu/~peters/fos/newsletter/05-03-04.htm#oai-google&gt;

M. Moffat, PerX Project Officer, ICBL, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, EH14 4AS. E-mail: m.moffat@hw.ac.uk