JISC-REPOSITORIES: "Subject Classification" Thread Summary

Compiled by Pete Cliff, Repositories Support Project, UKOLN

18th July 2008

http://www.jiscmail.ac.uk/cgi-bin/webadmin?A1=ind0806&L=jisc-repositories

Introduction

Between about 1pm Tuesday 17th June and 12pm Monday 30th June 2008 the JISC mailing list, JISC-REPOSITORIES[1], further discussed questions of subject classification, repositories and automation. The discussion totalled some 10,284 words (not including headers and quoted text) over 67 messages and the thread ("Subject Classification") spawned two others: "It's Keystrokes All the Way Down" and "Current Awareness". During the course of this discussion someone asked that a summary be created and this document represents an attempt to do just that. It does not attempt to attribute points to individuals just as it does not take any credit for the ideas expressed within.

Background

We begin with a question: "Do Institutional Repositories that make use of Library of Congress Subject Headings (LCSH) ask depositors to select the headings, or get cataloguers to do this work? Would it be better to simply use author chosen keywords (tags) or use a classification like ISI (to support REF)?"

Importantly, and implied by reference to REF (and the subsequent discussion), is that any requirement to subject classify should be made in the context of usage. (This is a general principle in the creation of any metadata). It would be right to ask: What is the purpose of subject classification of *Institutional (or other) Repository content*? (Note, that is *not* asking what is the purpose of subject classification *per se*). Other questions then arise: What is the cost of subject classification and how does compare with the benefits? Is human subject classification necessary, nice if you can get it or simply a waste of time? These are reoccurring questions within the repository community, suggesting that they have not as yet been formally explored.

What is the purpose of subject classification of repository content?

In theory at least there are interesting services that can be built using the subject classified content of repositories. These include systematic searching and browsing, filtering by subject area (discussed later), support for REF, and auditing of research grants (by attaching grant codes – which may carry subject information - to papers). The latter two are not demanded by end users of repository content, but by administrators and funders. The former are standard methods used to discover resources and it was felt that to remove support for these types of discovery without proper investigation would be premature and unfair to those people who rely on them.

The discussion seemed to veer towards full-text indexing, coupled with sophisticated search algorithms (such as those used by Google) and boolean queries, as sufficient mechanisms for discovery of repository content. There was a strong feeling that subject descriptors attached to metadata records of papers would not enhance/aid discovery and that if subject classification was required it would be difficult to see the value added by "human classification" (at deposit) over automatic classification (at deposit or any time after).

That said, some posters advised caution, suggesting that to entrust scholarly research to the power of the search engines was not something to be taken lightly and that to dismiss

subject classification, a standard discovery tool used by researchers and librarians, might carry some risks. Further, it was felt that there are limitations of full-text indexing and there was a question over whether or not a document's content (devoid of context) was sufficient to facilitate discovery (or automatic classification) of that document. Some felt this was a minor problem that would only occur with a specialised set of documents and that this set of documents would perhaps have no place in an Institutional Repository. Others felt this might be a very real issue for the content of IRs.

The discussion seemed largely based on opinion and impressions rather than studies assessing the usefulness of full-text indexing versus enriched metadata and the question was raised to ask if there were any studies looking into this.

Subject classification to information overload

Some felt that while subject classification did not aid discovery via search engines, it was still useful to distinguish content for subject based harvesters and to filter result sets, for example current awareness alerting services. IRs are, by default, as subject agnostic as the Institution itself. How then does a subject focussed harvester determine which full-texts to retrieve and index? Some services do not place any subject metadata into their records because it is be clear from the repository in question what the subject area is. However, machine to machine interfaces do not necessarily have the luxury of knowing the subjects each repository might cover.

A further issue was raised relating to current awareness and the limitations of alerting services built on top of full-text indexes. Often such alerts (via RSS feeds) would return false positives and it was suggested that a finer grained filtering (perhaps aided by subject classification) would be of use in solving these problems.

However, there was a strong feeling that machine classification would address these issues, adding subject classification after submission (or at retrieval), but as yet no one is very sure of how successful that would be now or how much better it might get in the future.

What is the cost (to Institutional Repositories) of subject classification?

The discussion suggested that deposit into repositories is disappointing and the poor rates of deposit can be directly and solely attributed to the effort (in terms of “keystrokes”) required to submit a paper. There was a strong feeling that reducing the metadata overhead (by, for example, not asking authors for subject headings) at submission would significantly increase the chances of authors depositing their work. That is to say the cost to IRs of subject classification is high: it prevents content deposit. (There was also the question of the author's qualifications for cataloguing a work in accordance with a subject scheme).

As aside to this discussion, the question was put to the list whether or not it really was the case that “keystrokes” were the main cause of the disappointing deposit rates. Some on the list felt that there were other, equal, if not more significant factors – such as copyright clearance/fears. If “keystrokes” were not the main factor, it could be argued that the cost of subject classification to deposit was less than envisaged, but there was only anecdotal evidence to support this.

That subject classification implies “keystrokes” that the authors are unwilling to make begged the question does all metadata requested/required form a barrier to deposit? If it does, should IRs be asking for any metadata *at all* other than that which can be gained

automatically? What if all barriers were removed and the submission interface for an IR were simply a Web site to which files could be uploaded/copied? How would such a Web site differ from an IR? (A few ways were mentioned: for example that an IR allows the institution to manage the scholarly output and that OAI-PMH was a better dissemination technology than screen scraping). However, the question remains: Are IRs themselves barriers to deposit? Barriers to Open Access?

Metadata Standards

There were implications for metadata efforts within the community and application profile work was mentioned in this context. The problem is that if IRs will remain empty if there is an insistence on high levels of complex metadata, what role is there for things like SWAP? Should (could?) SWAP stipulate a subject classification scheme? How will it be possible to get authors to construct the relationships SWAP requires if they will not/are not capable of selecting a subject heading? There was a feeling that software tools currently do not support the easy creation of complex metadata coupled with a concern that they never will. “Developer bewilderment” was cited as the reason; that is to say that the software developers themselves do not understand or accept that structured metadata is a requirement for discovery and because of this will not invest the time and effort developing the tools to create it.

Where now?

A number of questions were raised on the list as part of this discussion. Among these significant ones appear to be:

What are the requirements of IRs/services that subject classification supports?
Is subject classification an aid to resource discovery – from full-text indexing to alerting? Do we know either way or is it just a feeling?
Is the disappointing deposit rate still attributable to just “keystrokes”?

Just where we go from here is left to the reader.

During the course of the discussion it was suggested that the thread itself might be interesting to automatically classify. The following is the output from OpenCalais:

```
URL:
    http://www.driver-community.eu
    http://www.iriss.ac.uk/openlx
    http://tinyurl.com/62bmvk
    http://metalogger.wordpress.com
    http://search1.driver.research-infrastructures.eu
    www.digitalpreservationeurope.eu
    http://www.hull.ac.uk/golddust
    http://eprints.ecs.soton.ac.uk/11006
    http://search.arrow.edu.au
    http://cadair.aber.ac.uk
    http://www.eduserv.org.uk/foundation
    http://www.digitalpreservationeurope.eu
    http://eprints.ecs.soton.ac.uk/12094
    www.iriss.ac.uk/openlx
    http://eprints.utas.edu.au/view/authors/Sale
    http://elpub.scix.net/cgi-bin/works/Show?
    http://efoundations.typepad.com
    http://metalogger.wordpress.com/&gt;
    http://zoomii.com/&gt;
    http://www.libworm.com
    http://www.ukoln.ac.uk
    http://openaccess.eprints.org/index.php?
    http://www.franklin-consulting.co.uk
    http://nzresearch.org.nz/index.php/browse/browseSubject
    http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
    http://www.amazon.com/review/product/0691020728?filterBy=addFourStar
    http://eprints.ecs.soton.ac.uk/11125
    http://arxiv.org/abs/cs/0312018
    http://www.eprints.org/openaccess/policysignup
    http://www.intrallect.com
    http://tomfranklin.blogspot.com
    http://www.dcc.ac.uk
    http://edina.ac.uk
    http://road.aber.ac.uk
    http://www.icbl.hw.ac.uk/~philb

PhoneNumber:
    01970 628724
    02890 974824
    07989 948 221
    +44 (0)23 8059
    0161 434 3454
    0131 451 3278
    +44 870 234 3933
    +44 (0)23 8059
    +44 (0)131 651
    +44(0)141 330

MedicalCondition:
    bewilderment
    Paralysis

ProvinceOrState:
    Tasmania

IndustryTerm:
    brilliant open web interface
    http://search.arrow.edu.au/
    data-mining lack
    heavy-duty tools
    semantic-web site:latest
    subject search tool
    online repository
    search gateways
    systematic search
    web interface
    gogle search
    repository services
    subject search tool
    aggregated feeds available to other services
    Internet Resources Newsletter
    online repository
```

    repository technologies
    sensible human boolean search
    subject search
    smart text-processing software
    taxonomy search
    well-managed general web site
    software development
    boolean full-text search
    learning tools
    Internet users
    browses repository search
    browses repository search interfaces
    in-house tool
    online research repository
    boolean full-text search
    repository software development
    mass-market newspaper
    software developers
    web services suite
    search engines
    magic solution
    search tools
    semantic web
    boolean search
    tomnfranklin web
    search engine
    friendly portal
    cloud computing
    wildcat Web site

City:
    Glasgow
    Zurich
    Hand
    Southampton

Technology:
    ASCII
    repository technologies
    http
    AJAX
    html
    ascii
    My algorithms
    search engine

Country:
    New Zealand
    Australia
    Scotland
    United States
    United States
    Scotland
    United Kingdom

FaxNumber:
    02890 976586
    +44(0)141 330
    +44 (0)23 8059
    0131 451 3327
    +44 (0)23 8059

Person:
    Stevan Harnad
    Philip J Hunter
    Julian Cheal
    Scott Welsh
    Stevan Harnad
    Andy Powell
    Antony Corfield
    Neil Godfrey
    Gwasanaethau Gwybodaeth
    John Smith
    Sarah Currier
    Peter Cliff
    Ricky Rankin
    Neil Godfrey
    Carr On

Tom Franklin
Peter Crowther
Mason Ingrid Mason Digital
TÃ®m Cynorthwywyr Pwnc
Ian Stuart
Stevan Harnad On
Joy Davidson
Steven Harnard
Arthur Sale
Ingrid Mason Ingrid Mason
Phil Barker
Steve Hitchcock
Pete Cliff
Philip Hunter
Hugh Owen
Ingrid Mason
Ingrid Ingrid Mason Digital
Alma Swan
Philip Hunter Storelink
Rosemary Russell
Neil Just
Simeon Warner

Facility:
Library of Congress
Digital Library Section Edinburgh University Library George Square
Bureau of Statistics
Kelburn Campus
Aberystwyth University Llyfrgell Hugh Owen Library
Mountbatten Building
Computer Sciences Mountbatten Building
Library of Congress

Organization:
University of Southampton
Arthur Sale University of Tasmania From
Heriot-Watt University
University of Edinburgh
Eduserv Foundation
University of Bath
School of Electronics
Australian Government
University of Tasmania
School of Oriental and African Studies
Queen's University
Heriot-Watt University
School of Electronics and Computer Science University of Southampton
Harvard
University of Southampton
Victoria University of Wellington
Institute of Maths & Physics
Training Coordinator Humanities Advanced Technology and Information Institute
University of Zurich
Congress
School of Electronics and Computer Science
University of Glasgow
Australian Bureau
Arthur Sale University of Tasmania From
Australian Bureau of Statistics
Information Institute
School of Mathematical
University of Edinburgh
Bureau of Statistics

Company:
IRs Export
NARCIS
Tom Franklin Franklin Consulting
Computer Sciences
Google
Yahoo
Intrallect Ltd.
Google