Intute: from a distributed network to a unified database – lessons learned

Linda Kerr
INTUTE, United Kingdom

ABSTRACT: Intute (http://www.intute.ac.uk/) catalogues and describes the best Internet resources for education and research. It is funded by the Joint Information Systems Committee (JISC), and is primarily aimed at evaluating web resources suitable for undergraduate study. The service also offers Internet research skills tutorials, rss feeds of new resources added to the catalogue, a personalisation service (MyIntute), and a blog highlighting trends in Internet research skills and particularly good or topical subject-based resources. The current Intute catalogue of Internet resources is an aggregation of records from eight subject services previously funded by the JISC as the Resource Discovery Network (RDN). This paper describes the process and challenges of integrating these eight databases into one unified catalogue with one standard metadata schema, whilst continuing to satisfy the needs of different subject communities. The paper also outlines a current project to evaluate and compare the cost-effectiveness of manual and automatic metadata creation.

KEYWORDS: Web resources, subject gateways, cataloguing, indexing

1. Introduction

Using the Internet to search for information is now second (or first) nature to many people, in particular students in higher education. A recent Joint Information Systems Committee (JISC) and British Library-funded report on the "Google generation" concluded that the "Internet is one of the first places a student will look for information: 89% of college students use search engines to start an information search…. Students rely on the most basic search tools and do not possess the critical and analytical skills to assess the information that they find on the web" (CIBER, 2008).

Intute is a UK-service funded by JISC which provides access to the best Internet resources for study and research. Intute is a distributed network of partners with an executive based at Mimas, a nationally designated data centre at the University of Manchester. The main partners in Intute are: University of Manchester, Heriot-Watt University, University of Birmingham, University of Bristol, University of Nottingham, Manchester Metropolitan University and the University of Oxford. Each one of these partners has a responsibility for developing a particular subject area of Intute. These subject areas correspond roughly to the individual separate subject gateways, originally funded under project funding by JISC (and in some cases joint funded by Research Councils). Intute was launched as a unified service in July 2006, and combines the best elements of the subject gateways and aims to provide a coherent service searchable and browsable across different subjects. It also provides Internet Research Skills (the Virtual Training Suite) and a number of Web 2.0-type repurposing of the content – via twitter, blogs, rss feeds, and personalised lists of resources.

The current service offers a catalogue of over 120,000 Internet resources.
Technological advances and diminishing budgets have led to an examination of automatic cataloguing of resources, and the manual cataloguing of web resources by subject-experts is under pressure when set against the less costly community-based models of finding, organising and sharing information and information-about-information – for example Wikipedia and del.icio.us.


The oldest established gateway that would eventually become part of Intute, the Social Science Information Gateway (SOSIG), was funded as a pilot project by the Economic and Social Research Council (ESRC) in June 1994. Based in the University of Bristol’s Institute for Learning and Teaching, the gateway aimed to select, describe and classify quality Internet resources for the social sciences. This was the first of a range of subject gateways in the UK covering different subject areas, mostly funded by the JISC. It was a period of great interest in and funding for subject gateways, and exploring issues of the classifying and evaluation of Internet resources in the UK and internationally. The JISC-funded Electronic Libraries Programme (eLib) funded or part-funded a number of gateways: SOSIG (Social Sciences), Biz/ed (Business Education on the Internet), EEVL (Engineering), ADAM (Art, Design, Architecture and Media), History, RUDI (Urban Design) and OMNI (Medical Information). Also funded under the eLib Programme was the ROADS project: open software and tools that would support the development of the eLib gateways and contribute to the wider resource discovery environment.

BUBL (http://www.bubl.ac.uk), a gateway based at the University of Strathclyde, began organising resources along subject lines in 1991 (Dempsey, 2000), using first UDC and then the Dewey Decimal Classification System (Menzies & Nicolson, 2009). This hierarchical organising of information along subject lines was a natural process in response to the increasing number of resources discovered.

3. Cataloguing of Internet Resources

The eLib philosophy of “let a hundred flowers bloom”, encouraged the subject gateways to adopt different approaches to implementing classification and metadata schema. There was a shared perception that web resources required a lighter-touch cataloguing process than bibliographic resources. Chapman, Day and Hiom (1998) quote Vianne T. Sha’s three broad approaches to Internet resource discovery (1995):

The development of search services using robot-based search engines to index Web pages, an approach exemplified by services like Lycos and AltaVista.
The “manual subject guides” approach where human intelligence is utilised to identify and evaluate Internet resources. These are typically subject-based guides that take the form of HTML lists. Examples include the various WWW Virtual Library sites and the subject-guides listed by the Argus Clearinghouse.
The library cataloguing method, creating bibliographic records for Internet resources in library catalogues. The classic example of this approach is OCLC’s InterCat project.

The subject gateways chose option 2. The metadata schema was based on the IAFA (Internet Anonymous ftp Archive) templates (ROADS; Kirriemuir et al., 1998).
Subject terms were assigned to aid collection development, and allow subject browsing according to the perceived needs and habits of their subject communities. EEVL, for example, chose to use a modified version of EI Thesaurus main subject headings. SOSIG classified resources according to UDC (later modified and augmented), and OMNI according to the National Library of Medicine Thesaurus Medical Subject Headings (MeSH).

As an example of the classification work-processes in a gateway, the engineering subject gateway, EEVL, created broad subject headings and assigned one or more subject headings to a resource. These were supplemented by uncontrolled keywords. If a subject section became too populous (200 resources in a section was a cut-off point) then that section was split, and reclassified – a time-consuming process, but one that shows a desire to not overload the user with complex browsing and an understandable lack of fore-knowledge of how much information and what kind of information would be eventually available on the Internet (MacLeod et al., 1998). Gateways’ Collection Development Policies also influenced their subject classification – cataloguing at a finer level of granularity (for example, an individual paper or tutorial) required a more detailed classification scheme. SOSIG implemented HASSET thesauri terms to help refine searches (Chapman, op. cit.).

Further gateways were funded in 1999 - 2001 with the creation of the RDN. By this stage, some top-level gateways were made up of sub-gateways with different classification schemes – BIOME (Health and Life Sciences) was a grouping of the pre-existing OMNI with four new gateways: AgriFor (covering agriculture, forestry and food sciences), VetGate (covers all aspects of animal health) BioResearch (covering biological and biomedical research) and Natural Selection (covering the natural world). In some cases, this structure was dictated by the nature of the funding. For example, a bid to fund a gateway for aeronautical engineering based on an existing institutionally funded gateway (AERADE) was twinned with EEVL’s bid for continuing funding. The aeronautical engineering section in EEVL therefore had a different classification scheme to the main gateway.

When the gateways were separate, the differences in structure between them did not matter to most users. Again, the perception was that users came from distinct subject communities, although multi-disciplinary users such as librarians had to “learn” to use different interfaces. The idea of meta-gateways/portals – that brought together/aggregated content from a number of different sources – developed.

4. Resource Discovery Network

A final tranche of gateways were funded in 2001, and by this stage were grouped by JISC under an umbrella organisation: the Resource Discovery Network (RDN). The emphasis at this stage was to develop a service that would allow users to search across subject gateways but preserve the individual gateways. To this end, an aggregated search known as the RDN ResourceFinder was created (Cliff, 2004). This used the OAI metadata harvesting protocol to gather metadata from the gateways on a regular basis and present this data via a single search interface. The gateways were therefore required to map their fields to the Dublin Core simple metadata element set.

The ambition to provide a cross-browse service by subject was not realised within the timescale of the service.
However, during this period the gateways moved towards a standardised Resource Type listing and the RDN Interoperability and Standards Framework was published in 2002 (Powell).


The subject gateways were further integrated in 2006 when they were funded by JISC to combine to form one multi-subject catalogue and service, which was rebranded very successfully as Intute.

The objective was to create a multi-disciplinary, seamless search and browse for the user over a range of subject areas, and allow the development of supporting services.

A new database was created to store the combined Internet resources from the eight subject gateways (EEVL, SOSIG, PsiGate, GeSource, BIOME, ALTIS, HUMBUL, Artifact). Four different subject group cataloguing interfaces to the database were created: Social Sciences, Arts and Humanities, Science, Engineering and Technology, and Health and Life Sciences. These corresponded to the subject groupings on the public website. A new database was created; a requirement gathering process was undertaken by consulting each of the subject gateways, and looking at existing practices, metadata fields, classification and cataloguing procedures.

It was considered desirable not to lose any metadata held in the subject gateway records and to standardise the metadata schema where possible. Although one new metadata schema was created, cataloguers only saw the fields relevant to their own discipline. So, for example, historical period was viewable from Arts & Humanities, as only their records included this information.

<table>
<thead>
<tr>
<th>Field</th>
<th>Standard/Convention</th>
</tr>
</thead>
<tbody>
<tr>
<td>*Handle</td>
<td>Automatically generated</td>
</tr>
<tr>
<td>*Subject Group</td>
<td>Automatically generated from cataloguer login</td>
</tr>
<tr>
<td>*Subgateway</td>
<td>Selected from Intute listing</td>
</tr>
<tr>
<td>*Title</td>
<td>Sentence case, capitalising proper nouns (based on AACR2 (11))</td>
</tr>
<tr>
<td>Alternative Title</td>
<td></td>
</tr>
<tr>
<td>*URL</td>
<td>Repeatable field</td>
</tr>
<tr>
<td>Language</td>
<td>ISO 639-2 Code</td>
</tr>
<tr>
<td>Description</td>
<td>Free text</td>
</tr>
<tr>
<td>Keywords (controlled)</td>
<td>Choose from look-up tables or enter keywords from suggested thesauri</td>
</tr>
<tr>
<td>Keywords (uncontrolled)</td>
<td>This field is used for free text keywords (terms not included in relevant thesauri), or for the non-preferred terms, which come from thesauri not available as look-up tables.</td>
</tr>
<tr>
<td>Resource Types</td>
<td>Intute Resource Type List (in-house)</td>
</tr>
<tr>
<td>Classifications</td>
<td>Use Individual classification schemes</td>
</tr>
<tr>
<td>Country of Origin</td>
<td>Based on listing created originally by EEVL</td>
</tr>
<tr>
<td>Resource Creator</td>
<td>Enter personal names, where possible, in the order suggested by AACR2 [11] chapter 22 for headings of persons or Library of Congress tables</td>
</tr>
<tr>
<td>Resource Publisher</td>
<td></td>
</tr>
</tbody>
</table>
**5.1 Major differences and issues**

During this catalogue integration project, it was necessary to balance the differences and similarities between the subject gateways’ approach to cataloguing and resource discovery to achieve a workable service. The main differences were

- How resources were classified by subject and type - use of controlled vocabularies and subject headings.
- The granularity of resources catalogued.
- The subject scope (e.g. commercial sites are relevant to engineering students but are less likely to be useful in the medical sector).
- The geographical and chronological coverage of resources.
- The importance of publisher/creator.
- The target audience (Higher Education or Further Education).
5.2 Classifications by Subject Group
When creating a new catalogue, the thorny issue of introducing a single classification scheme for subject headings was raised. It was decided not to introduce one unified classification scheme across the subjects, in order to maintain the specialist nature of the service and also that such a mammoth task would not have been possible in the time available to launch the new service. Four "subject group" views onto the database (public and cataloguing interface) were created.

As can be seen below, a range of subject specific classification schemes and thesauri were and are used. Most are subject-specific.

The catalogue supports the use of uncontrolled and controlled vocabularies, and subject headings. The subject headings can correspond to an existing classification scheme (for example with engineering, which is based on the Engineering Information Subject Headings) or have developed in natural language from the community. Records can have subject headings, controlled keywords and/or uncontrolled keywords. In addition, the medical and nursing gateways do not allocate broad subject headings but thesaurus terms only.

Arts and Humanities
Subject classification (in-house)
- E.g. Modern languages and area studies > Asian > Japanese > Poetry

Controlled Keywords:
2. Place: TGN (Getty Thesaurus of Geographic Names)
3. Keywords AAT: AAT (Art & Architecture Thesaurus)
4. People Index: TGN or ULAN (Union List of Artists Names)

Uncontrolled Keywords
Period/Coverage
Geographic Name (TGN)

Social Sciences
Subject classification – modified UDC
- E.g. Social sciences > Hospitality and catering > Catering - general > Events catering

Controlled Keywords – choose one or more
1. CABI Thesaurus
2. HASSET (Humanities and Social Science Electronic Thesaurus)
3. IBSS (International Bibliography of the Social Sciences thesaurus)
4. LIR legal thesaurus
5. SCIE (Social Care Institute for Excellence thesaurus)
6. SRM (Thesaurus of Social Research Methodology)

Uncontrolled Keywords
Science, Engineering and Technology
Subject classification /heading
1. Ei Thesaurus (Engineering Information)
2. ACM (Association for Computing Machinery)
3. DDC modified

Uncontrolled Keywords

Health and Life Sciences
Controlled Keywords – choose one or more
1. CABI Thesaurus
2. MeSH (National Library of Medicine Thesaurus Medical Subject Headings)
3. RCN (Royal College of Nursing)
4. LCSH (Library of Congress Subject Headings)
5. Species
6. Histmed-name
7. JACS (Joint Academic Coding System)
8. LEARNDIRECT
9. Altis

In general, the philosophy in classifying and cataloguing Internet resources was community based and pragmatic – a bottom-up approach. Perhaps the best analogy is to specialist libraries, created to serve the needs of their users, who may be familiar with a particular thesaurus. This has consequences, discussed below, and it is interesting to speculate if this approach would have been taken if a cross-subject service had been developed from the start. For some disciplines specialist thesauri are considered essential, and a great deal of credibility would have been sacrificed if they had been removed.

In July 2009, a top level of subject headings based on JACS was introduced, and the four subject-groups removed.

5.3 Resource Types
Each resource catalogued by the subject gateways is allocated one (or more) resource types, selected from a controlled list. For example, software, journal, e-book. The resource types used by the gateways were standardised to some extent by the RDN Interoperability Working Group in 1992 (Cliff). So although there were strong similarities in gateway’s individual resource type listing, there were differences in the number, nomenclature and definition of resource types in each gateway, and as part of the creation of a single catalogue it was decided to create a single resource type list. No subject group would be required to use all resource types, but naming was standardised and clarified – for example “field studies” and “field guides” had slightly different meanings in Geography and in Humanities. Some resource types were dropped and new scope notes written for others. “Learning resources” was a difficult resource area to tackle as some gateways categorised at the level of “tutorial” or “lecture note” and others at a broader “learning object” level; in the end a “catch-all” of “other educational resources” RT was created to allow subject groups to spend time reclassifying resources given originally a more general classification into the new specific resource types. Where possible, global edits were carried out, sometimes using a keyword, or title to identify resources. It required in-depth knowledge and familiarity with the catalogue by individual cataloguers, some of whom have been with their subject gateways since they were created. Even whether a Resource Type name should be
singular or plural had to be decided! In the end 77 resource types were agreed. This was further rationalised in 2009 to 51 RTs, grouped by related type.

5.4 Issues and solutions
This partial integration of subjects had the following consequences:

For Cataloguers

- Unable to classify across different subjects (outside the four subject group interfaces) and therefore unable to de-duplicate records – for example a record in a cross-disciplinary subject area. It should be noted that there was some debate whether these records were in fact true duplicates, as the descriptions and classifications were tailored for the subject communities. Perhaps it would be better to say these are records describing the same website. However, their continuing presence does represent a duplication in time and effort to maintain.

For Users

- Overlap of topics between classification schemes and subject headings.
- Cross-disciplinary browsing difficult.
- Different ways of finding records – consistency of approach.

The following solutions were adopted, although full integration was postponed due to funding issues. This full integration would have integrated the four subject group interfaces and allowed cross-classification between subject areas.

- Some sections renamed to distinguish between different classifications schemes.
- Introduction of manually assigned related terms.
- Reduction where possible of duplicate records by better delineation of the scope of subject areas.
- Provided a course-orientated view onto the subject hierarchies – 19 subject headings based on JACS. Linked to the underlying classification schemes which are still available for browsing and filtering at subject level. This allowed inter-disciplinary browsing – e.g. Geography became a top level heading, linking to resources in both Social Sciences and Physical Sciences.
- Four subject groups removed on the public interface.
- Same terms from different subject thesauri are linked.
- Rationalised metadata fields in response to user feedback: removed ISBN; ISSN; Geographical coverage.

5.5 Automatic Metadata Classification
As the service developed, the need for efficiencies and the desire to provide better resource discovery for users led Intute to explore automated cataloguing. The Value for money in automatic metadata generation (ViM) project (http://www.intute.ac.uk/projects.html - end Sep.2010) is comparing the cost-benefits of automatically generated metadata, manually generated metadata and user-contributed metadata. Autonomy IDOL software is being used to automatically generate data to populate a number of metadata fields in the standard Intute
record template. The metadata is created from the metadata/data on website. In the control method, Intute cataloguers create standard catalogue records.

The del.icio.us social bookmarking API to being used to retrieve suggested keywords/tags for user-suggested websites using “DC-dot” type metadata editor to populate metadata records.

Preliminary impressions suggest good metadata is dependant on quality of metadata on source website or if the website has a simple structure with lots of relevant text that software can capture and analyse. However the Autonomy system is being trained to improve effectiveness. The trial period has just begun, comparing the two automated and one manual method, with cataloguers keeping timesheets to compare “productivity” between the different systems. A report will be written on effectiveness at the end of the project. ViM’s sister project, EASTER (http://www.ukoln.ac.uk/projects/easter/), is evaluating a number of existing tools for automated subject metadata generation.

6. Summary

The viability of manual classification of Internet resources has been weighed up and alternatives considered since the very beginning. Most studies have concluded that the human selection process, in particular the informed election of resources and “hand-written” record descriptions are valued by users. However, this approach is labour-intensive and the number of resources that can be identified limited, and in these economic times, perhaps even the application of automated indexing and classifying tools may not be enough to ensure the continuation of even this semi-automated approach to resource discovery. Light-touch social networking tools, such as del.icio.us, can supply lists of resources. These are not quality-checked, but perhaps the future will lie in developing students into competent evaluators of quality by developing their Internet research skills, rather than by the manual selection of resources.

References

All URL valid as of 11.03.2010


