ARIADNE

Immaculate catalogues, indexes and monsters too…

David E. Bennett reports on the three day residential CILIP Cataloguing and Indexing Group Annual Conference, University of East Anglia, Norwich, UK, 13-15 Sepember 2006.

Introduction

Restful accommodation and pleasant food prepared the delegates for the carefully balanced mix of social networking sessions and challenging seminars. Everyone was extremely friendly and most proved to be erudite socialites, networking in some cases with great assertiveness and sense of purpose.

Cataloguing and classification was revealed as an area of library and information science that has survived years of neglect by most library schools to reveal itself as the much-needed solution to online resource accessibility. Cutting-edge advances in information technology were showcased, including the promising prototype AUTINDEX indexing software for indexing digital and digitised documents, the latest research into novel methods for automated image compression and indexing, searching and retrieval methods, and reviews of user-operability studies of image retrieval search interfaces.

Traditional print material cataloguing was not neglected. The development of the powerful yet simplified set of cataloguing rules known as the Research Development and Access protocol, set to supercede and replace the current Anglo-American Cataloguing Rules (AACR2 *rev.*) in 2008, was reviewed and the latest developments modelled. The technology behind the newly released union serials catalogue SUNCAT for attaching and tagging minimally described institutional holdings records to full bibliographic records using uniquely-tailored MARC bibliographic fields and bibliographic record retrieval was explained.

The questions and answers sessions following each seminar were almost as informative as the seminars themselves, probing and teasing out diverse threads of discussion from both within and beyond the scope of the original seminars.

Day 1: Overview and 5th UK Cataloguing and Indexing Standards Forum

Immaculate catalogues: taxonomy, metadata and resource-discovery in the 21st century

Alan Danskin, British Library

Challenges facing cataloguing

The exponential increase in the rate of print resource publication [1], the arrival of an ever-expanding collection of online material, rising costs and falling numbers of cataloguers were cited as reasons for selecting and filtering material for cataloguing on the basis of academic worth [2]. It was suggested that by necessity, non-academic resources now need to be catalogued and classified using simplified and derived metadata, including the automated and social indexing of web resources. Current collaboration efforts to engineer interoperable metadata standards for resource description for publishers and librarians and automation of metadata extraction were suggested as ways of transforming cataloguing from a cottage industry into a means of mass production. The benefits of cataloguing should also be marketed [3].

Library OPACs - RIP

It was asserted that online public access catalogues (OPACs) must be rapidly transformed from antiquated browsing interfaces with crude search tools useful only for locating specified materials into intuitive, aesthetically pleasing search tools that high quality search results and aggregate similar resources, in a similar way to commercial book retailers.

Moving from AACR2 to RDA

Ann Chapman, CILIP/BL Committee on AACR

The Research Development and Access protocol (RDA) was developed to satisfy the perceived need to simplify the cataloguing rules, increase their international acceptance and to permit cataloguing of new resource types together with technical improvements such as the separation of general and specific material designator terms. The committees [4] and processes [5] set up to develop the new cataloguing rules, and attempts to map them to MARC 21 machine readable cataloguing format were discussed. The current version of RDA was demonstrated by Alan Danskin on day 2 *(not described).*

MARC 21 Update

Corine Deliot, British Library

The process for submitting and evaluating proposed changes to the MARC 21 machine-readable cataloguing format was described. Methods for converting from Unicode to the archaic MARC-8 format where records included characters not recognised in MARC-8 due to its much smaller character set were apprasised. Deleting records or unrecognised characters was agreed to be unacceptable. Substituting a numeric character reference identifying the Unicode character in hexadecimal would have allowed the character to be mapped back into Unicode without information loss. This method may be adopted in the future. For the moment, the chosen method is to use a single placeholder character, the vertical bar (ASCII hex 7C), in place of all unrecognised Unicode characters. Recent changes to MARC 21 were described and discussion papers introduced.

When the rules change: cataloguing rare books

Dr Karen Attar, Senate House Library

The unique cataloguing requirements of rare and antiquarian materials was described, including the need for extensive physical descriptions, clear citation and aggregation of different publications and imprints of the same work on OPACs. Cataloguing standards and codes applicable to the description of antiquarian and rare books were compared. Controversial developments, such as the International Standardised Book Desciption guidelines for Antiquarian materials (ISBD(A)), Area 4, Option B to enter bibliographic information as printed in the book rather than as described under standards, were explored. Instances where the Bibliographic Standards Committee (BSC) objected to changes to standards were highlighted, such as changes to pagination descriptions under the 2006 re-draft of the ISBD (A) standard, which the BSC attacked as being less clear and less parsimonious.

Day 2: Collections, technology and users

Image, shape and multimedia resource discovery

Steven Rüger, Imperial College, University of London and Prof Frederic Fol Leymarie, Goldsmiths College, University of London

Novel multimedia search interfaces

Novel search interfaces for retrieval of image and multimedia resources were showcased. Medicine, personal collections, multimedia digital libraries, media archives, entertainment, tourism, e-learning and retail, especially multimedia catalogues were suggested as potential applications.

Bridging the semantic gap

Methods for bridging the 'semantic gap' between locations, structures, people and names and iconic meanings, and what computers can recognise by simplifying images and comparing them with a database of pre-labelled structures were discussed. It was suggested that, given approximately 30,000 indexed examples, computers could use statistical methods to identify and index similar images. Different objects and regions of an image can be labelled using simple terms. Aggregations of such objects within an image can then used to convey complex ideas, such as grass, sky, and people arranged around a dark object signifying "barbecue".

Novel image search interfaces

Prototype visual search methods were described, where the image to be searched for is presented centrally and similar images selected by lateral browsing are displayed in a circle around it, users dragging these images towards or away from the central image to increase or decrease their weighting in the search.

Existing document supply and OPAC catalogues were again attacked for their failure to match the intuitive and powerful platforms offered by commercial resource providers which also offer summary information, story boards and key frame summaries for video material, document clustering and cluster summaries.

Professor Leymarie described methods of simplifying two- and three-dimensional images to various simple components for compressed image storage. He confessed that accurate automated indexing and classification of such compressed elements was still difficult.

Three blind men and the elephant: current and future directions in image retrieval

Colin Venters, University of Manchester

Automated indexing of images

Computer methods used to analyse images were described as a Euclidian comparison of local physical properties, such as reflectance, brightness, hue, chroma and brilliance, in different regions of the image. It was suggested that following on from traditional cataloguing codes which attach concept-based metadata to images, semi-automated systems could use the physics of an image to extract, index, retrieve and cluster search results based on the physical similarity of images. Computers still fail to distinguish between visually similar images of different objects or with different iconic meanings. Other problems include noise and difficulties with inference, for example computers struggle to infer shapes composed of parallel lines or unconnected smaller shapes.

Evaluation of different user interface designs for image retrieval

Users dislike making repeated fine adjustments to search parameters. Querying by browsing is possible only if images are classified according to an intuitive taxonomy. A usability study conducted on small group of computer science students suggested classifying images by their predominant colour impedes searching. Most students sought images of specific objects, even when asked to find images to represent abstract concepts. Studies showed that searching for images by sketching components using drawing applications was time-consuming and difficult for untrained users. Classifying images by iconic meaning was criticised because it relies on the same interpretation being placed on an icon by both cataloguer and end-user.

From spectator to annotator: possibilities offered by user-generated metadata for digital cultural heritage collections

Seth van Hooland, Université Libre de Bruxelles

Overview of distributed indexation of images

Retrieval of high-level semantics within image databases traditionally relies entirely on human indexing. This is extremely time-consuming and therefore expensive, especially where digital images are created on a large scale. Social indexing comments can put diverse and scattered information into context and add information to images [6]. Social tagging also allows subjective accounts of personal experiences and memories to be added to images. For cultural heritage archives this could be considered important. The relevance, quality and even the accessibility of such metadata after large amounts of text has accrued is, however, questionable, although van Hooland was optomistic. Neither the form nor content of the indexation is usually controlled.

Fitness for purpose, the overriding criterion for judging the value of semantic tagging, is reduced by polysemy, synonymy and the low semantic value of social tags, together with the uncertainty that a researcher could trace a tag back to the person who made it in order to investigate their story. Emotional responses, especially superlatives, also reduce the accessibility of useful comments to users, although most can be removed by automated processes.

Analysis of comments and search purposes

Queries of the image database of the National Archives of the Netherlands and comments attached by social tagging to images held in the database were categorised using the Shatford-Panofsky grid. 82.5% of queries were for specific person(s), event(s), location(s) or date(s). No queries for abstract subjects were found. Comments attached to images similarly focussed on specifics. Only 2.86% of comments concentrated on emotion or abstraction. Few comments reflected on personal experiences regarding the image. Some users pose questions, turning the metadata into a dialogue, helping to create virtual communities around heritage institutions.

Swings and roundabouts: a look at the role of Cat and Class in the LIS curriculum today

Kathleen Whalen Moss

A comparison of cataloguing and classification training in the 15 UK universities offering library and information science education library school syllabi from across the country were appraised by means of semi-structured interviews, a web survey, telephone interviews, and a literature review. In 2005/06, only five undergraduate courses taught cataloguing and classification. Only ten postgraduate courses offered six or more weeks practical cataloguing and classification training. With different members of the information and library science community speaking out for and against cataloguing and classification as a necessary skill, different library schools have chosen to teach it to varying degrees. Some schools ignore it altogether in favour of specialised modules, such as those in business information, which standing alone have been denounced as "satisfactory for no-one"; others have thoroughly endorsed it as a core skill that underpins a clear understanding of metadata and the description of online resources and as a pre-requisite for being able to interpret catalogues for end-users [7].

Owing to the lack of cataloguing and classification training provided to an entire generation of librarians, a skills gap has opened up which is putting library schools willing to support cataloguing and classification modules under pressure to find suitable lecturers with practical workplace experience. Increasing interest in metadata may encourage a more thorough treatment of cataloguing and indexing. Experts in the field have stated that their sophisticated concepts of cataloguing and classification are ideally suited to describing web resources [8].

Despite RAE pressures, a reluctance to spend money on working materials, shortages of skilled teaching staff and pressures on teaching time brought about by the modularisation of syllabi cataloguing and classification instruction is increasing in library schools' curricula, supported by CILIP and assisted by the increasing status of metadata research. Post-professional university tuition is thriving, and co-operative schemes may help to fill gaps in tuition.

SUNCAT: the creation, maintenance and challenges of a national union catalogue of serials in the UK

Natasha Aburrow-Jones, SUNCAT

Launched in August 2006, SUNCAT aims to provide researchers with a union catalogue for all academic serials that details holdings and associated access rights and provides librarians with a central repository of high quality bibliographic records [9], which may be downloaded using the z39.50 file transfer protocol [10], in exchange for serial holdings records from the downloading institutions. It acts as a tool for locating resources for document supply and attempts to raise awareness of the need for quality serials information among librarians and researchers. Serial holdings are updated regularly.

Different serial titles tagged using "ONIX for serials" formats, given a unique SUNCAT identification numbers (SC-IDs) and automatically matched using an algorithm to the most complete bibliographic record available for that title, *i.e.* with a matching SC-ID. Inadequately or incorrectly catalogued records uploaded into SUNCAT do not automatically merge and are instead highlighted and merged manually by the SUNCAT team [11]. Librarians are alerted to non-merged records and encouraged to inform SUNCAT of mismatches and other errors.

Day 3: Subject access tools for the 21st century – ontologies, taxonomies and thesauri

Terminology mapping for subject cross-browsing in distributed information environments

Libo Si, Loughborough University

Approaches to mapping semantics between the different metadata standards used in different databases in order to provide platforms capable of simultaneously searching several databases with different individual metadata standards and controlled vocabularies were reviewed. Derivation of metadata standards, recombining metadata elements from different metadata schemes into one application profile, "crosswalk", *i.e.* metadata mapping specifications, metadata registries [12], aggregation, *i.e.* conversion of heterogeneous metadata standards into a consistent form and, for web resources, the use of the Resource Development Framework (RDF) as a platform for integrating different metadata schemes. Switch languages and co-occurrence mapping were discussed.

Metathesauri merging concepts from different controlled vocabularies were posed as one method of resolving differences between controlled vocabularies. It was recommended that resource providers publish metadata their schemes in semantic web enabled format, *e.g.* RDF, XML, to facilitate their re-use. The development of a common metadata scheme that can accommodate elements from other metadata schemes and upper level ontologies, onto which metadata schemes, concepts, intra- and inter-relationships in different knowledge organisation systems could be mapped was cited as a promising solution.

AUTINDEX: automatic indexing and classification of texts

Catherine Pease and Paul Schmidt, Institute of Applied Information Science (IAI), Saarbrücken

The shift in user focus from library-centred to internet-based research was blamed in part on the poor quality of search results resulting from inconsistent and overgeneralised human indexation, which together with the need for a full text match by most library search tools and inflexibility in semantic relations.

AUTINDEX is a prototype indexation and classification application [12]. Digitised texts are subjected to morpho-syntactic analysis, isolating the lemma of each word and then tagging it with relevant information including which part of speech in which it occurs. Shallow parsing then resolves grammatical ambiguities and identifies noun phrases that may be used for indexation. Multiword terms and their syntactic variants are located and keywords, ignoring inflectional differences, identified based on their frequency and nouns are weighted according to semantic type. 140 symantic types are included in AUTINDEX's morpheme dictionaries. If the client organisation provides a thesaurus of controlled terms, AUTINDEX maps identified keywords to thesaurus terms. If a classification system is provided, AUTINDEX annotates text descriptors with classification codes and uses the frequency of thesaurus descriptor terms, hyperonym and synonym relations to calculate topic classification and assign a suitable class mark to the document.

Weaknesses identified in the application were that it can only index digitised documents and, although extensible, it currently only indexes English and German texts. Print materials therefore need to be scanned in and converted to text documents using OCR technology to facilitate automated indexation. No comparison was made between the time taken to scan in and convert printed materials to text documents using OCR technology and manual indexation.

The epidemiology of IPSV (Integrated Public Sector Vocabulary)

Stella Dextre Clarke, Independent information consultant

Integrated Public Sector Vocabulary (IPSV) was a product of the 1999 "Modernising government" white paper, which specified that all dealings with government should be deliverable electronically by 2008, enabled by an e-Government Interoperability Framework. The evolution of the IPSV from 1999 until 2006 was described and factors in its success were identified.

The original idea of indexing all e-government documentation for keyword searching using a single pan-governmental thesaurus was abandoned in 2002 in favour of the Government Category List (GCL) of just 360 preferred terms and 1000 synonyms used to broadly classify documents for efficient browsing. Local authorities, developed the Local Government Category List (LGCL) of 1400 more specialised preferred terms with support for indexing local government subjects. The Seamless Consortium, led by Essex County Council, independently developed a portal of 2600 preferred terms. Some local authorities were then obliged to index resources with GCL, LGCL and Seamless taxonomy. In April 2005, IPSV was launched to rationalise e-government. Its use was made compulsory. Whilst IPSV has 3000 preferred terms and detailed indexation support for local government and community information, most government departments are able to use an abridged version of only 500 preferred terms for broad classification.

A longitudinal study to evaluate the quality of e-government indexing is underway. It is speculated that the success of the initiative can only be judged after it has become established, a process that will take at least ten years.

Conclusion

Cutting edge information and computer science research promises to simplify information searching and provide high quality semantic, concept orientated descriptions for all materials, together with more detailed and consistent indexation. Library catalogue platforms need to be developed to compete with the intuitive interfaces, powerful search engines and the ability to aggregate similar resources offered by commercial services. Cataloguing and classification education is gradually recovering in universities. Changes to working practices, social indexing, and automated cataloguing, indexing and classification may ease the financial constraints threatening traditional cataloguing practice.

References

1. British Library's content strategy http://www.bl.uk/about/strategic/pdf/contentstrategy.pdf
2. Leysen, J.M. and Boydston, J.M.K. Supply and demand for cataloguers present and future *LRTS* **49(4),** 250-265.
3. Measuring our value: results of an independent economic impact study http://www.bl.uk/about/valeconf/pdf/value.pdf
4. CILIP-BL Committee on AACR http://www.slainte.org.uk/aacr/
5. Joint Steering Committee for the Revision of the Anglo-American Cataloguing Rules (JSC) http://www.collectionscanada.ca/jsc/
6. Example of a socially tagged digital archive http://na.memorix.nl/
7. Broughton, V. (2004). Classification – come back all is forgiven. *Library and Information Gazette* 17 December, 1.
8. Rupp, N. and Burke, D. (2004). From cataloguers to ontologists: changing roles and opportunities for technical services librarians. *The Serials Librarian* **46 (3/4),** 221-226.
9. SUNCAT: a brief history http://www.suncat.ac.uk/description.shtml
10. SUNCAT: processing files from contributing libraries http://www.suncat.ac.uk/librarians/data_processing_initial_load.html
11. AIMSS: automatic ingest of metadata on serial subscriptions http://www.jisc.ac.uk.cfm?name=project=aimss/
12. AUTINDEX website (in German) http://www.iai.uni-sb.de/iaien/de/autindex_dfg.htm

Author Details

David E. Bennett
Graduate Trainee
University of Oxford

Email: Pharm2_uk@yahoo.co.uk

Keywords:

1. *Cataloguing*
2. *Indexing*
3. *mapping*
4. *image*
5. *IPSV*
6. *SUNCAT*
7. *RDA*
8. *cross-searching*