Legal Information Management, 16 (2016), pp. 265–269

doi:10.1017/S1472669616000530

CONFERENCE REPORT

Law Via the Internet 2015 Conference

Abstract: Free access to public legal information for the general public and professionals promotes justice and the rule of law. Presenters at the 2015 Law via the Internet conference discussed projects using the power of technology combined with expert human input to make information accessible, and to extract new information from large document collections. Although the scale is different, there are similarities in the ways in which indexers and informaticians explore meaning, develop standards and consider user needs to make information widely accessible. Glenda Browne reports on the conference.*
Keywords: legal information; free legal information; informatics; indexing

INTRODUCTION

I attended the 'Law via the Internet' conference hosted by the Australasian Legal Information Institute (AustLII)[1] at the University of New South Wales from 10–11 November 2015. The conference proved to be interesting in many ways – for a view of the interaction between different professionals in the world of legal informatics[2]; for the innovative projects described; and for a sense of the global collaboration in this field. Topics of particular interest to indexers include the different approaches to information access (the balance between computer and human input), the use of standards (especially those relating to metadata), and the use of topic modelling to analyse content.

One interesting feature of the conference was the use of panels where one person spoke, and then other experts in the same field commented on the talk or asked questions of the presenter, followed by general questions from the floor.

In the opening session, Gabrielle Upton, the NSW Attorney General, saw technology as being a way of improving access to justice. Professor Lesley Hitchins, Dean of Law at UTS, spoke about the way that open access – both national and international – has shone a light on how the law operates.

In the first plenary session, Paul Chadwick, from Law Media, spoke about open access providing two foundations for justice:

- An accountability mechanism – publicity gives checks on legal decisions
- A protective mechanism – the public protects the independence of judges.

*The Editor of LIM is grateful to Maureen MacGlashan, the Editor of The Indexer, for giving permission for this article to be reprinted in this issue of LIM. The article was originally published in (2016) The Indexer 34(1), 12–16.

INFORMATICS VERSUS INDEXING

Most of the projects discussed at the conference dealt with large, open-ended data sets, and therefore had more in common with collection indexing, with its large teams, grouping of topics, and creation of bibliographic and subject metadata, than they did with book indexing.

Some of these projects existed to point users to known information, but others were aimed at analysing large collections of data to bring to light new information. Indexing does this too, to some extent. In book indexing, an author might say 'the index showed me themes I didn't know I'd focused on', and collection indexes can become explorable documents of their own. Xu Shu has suggested that the ISO 999 list of functions of an index should be expanded to cover 'serving as a reference tool in its own right without the need to access the text itself'[3].

Another difference from professional indexing is the tolerance of error, with the understanding that extraction or grouping can't be perfect with such large collections (of course, they're not always perfect in small indexes, either). One of the programs uses heuristics to identify the Act being referred to in a phrase such as 'we saw in s5 above'. I asked what they would do if errors were found. The answer was that there is always a trade-off between recall and precision (getting everything that is possibly relevant, and getting only content exactly on topic), and that 10 % error is acceptable for current awareness information. If something needs fixing, they would make the correction in the algorithm so that the error doesn't continue if the output is re-generated (the same principle applies to embedded indexing).

COMMUNITIES

There was a strong focus on community in the conference – both the importance of providing information

access for the community, and the community of legal informaticians working together towards common goals. There is a strong global focus in the LII world, with people sharing ideas and technology, and working together to reach their goals.

Online gravity

Keynote speaker Paul X McCarthy spoke about the concept of online gravity – that in the digital economy huge planet-like companies form and overwhelm their competition. Traditionally there have been leading brands (eg, Coca Cola), challengers (eg, Pepsi) and niche brands (eg, Dr Pepper). In the digital economy, there is far greater concentration at the leading brand (eg, Google) with only small markets for the outliers (eg, Bing). There are no dual planets.

Some points from his talk:

- If you're not a planet, you have to use the gravity of the big planets to your own advantage. He calls this gravity assistance.
- We need STEAM (Science, Technology, Engineering, **Arts**, Mathematics) not just STEM.
- Portfolio (multi-pronged) careers, eg, scientist/editor/indexer, are on the rise, personified in Generation Slashy.
- Talent is now based in the team, not only the individual.

The relevance for us as information workers, is that we should attach ourselves to the planets (eg, the International Digital Publishing Forum or Ingenta) to make use of gravity assistance, that we should value our strengths in both the arts and technology, that we should keep expanding our skills and the scope of work that we do, and that we should work in teams where the whole will be greater than the sum of the parts.

Community legal information access

Stephanie Booker spoke on two projects for access to Northern Territory law – the *Law Handbook Online* and the *Plain Language Law Portal for Northern Territorians*. These both act as pathways from secondary to primary sources.

The *Law Handbook Online* is replacing the print version. Online creation and access are seen to provide many benefits including ease of updating and ease of searching (indexing was not even mentioned).

The Plain Language Online Portal provides layered information – bite, snack, and meal (with the *Law Handbook Online* and government resources being the 'meal').

Phillip Chung spoke about the Australasian Legal Scholars Community collection – this comprises over 130 Australasian databases with 65,000 articles. Authors contribute by:

- Uploading metadata about their own publications and about co-authors
- Disambiguating entries from suggested inclusions (generated from automated data mining by LawCite)
- Uploading full-text articles.

STANDARDS

No matter the format of the content, standards are crucial for optimal sharing of, and access to, information. Use of international standards saves time in preprocessing of documents, enhances findability and enables data sharing. Even small steps to standardize drafting help, for example, the use of ISO format for judgment dates (eg, 2016-01-05 for the fifth of January), and numbering of paragraphs, not pages.

Print – resilience of authority in law

Francis Johns spoke on the resilience of authority in law. Robert Berring, a librarian and law educator, promoted the bibliographic view of authority, saying that authority was bestowed by publication, and warned about the risks of undisciplined research, saying 'Make sure they go to the indexes'.

Johns noted that law is not the only discipline concerned with authenticity, and said that while there had been a prediction that online access would break the hierarchy of information (showing importance), in reality, not much has changed. He found that Judith Lihosit[4], a librarian, had done similar work and had found that the predicted 'collapse of the legal universe' hadn't happened. Instead, law is an apprenticeship and authority is often based on social interaction, and not tied to texts.

HTML, XML and EPUB

The preferred standard for text output depends on the source and the audience. Courts write in Word, and can't be asked to draft judgments in XML. Word documents are usually converted to HTML, which is considered to be the easiest standard to work with. HTML 5 has more structure than earlier versions of HTML, but is not as good as XML.

EPUB is used primarily to access distribution channels. HTML is considered to be acceptable for a university textbook as students are a captive audience, and having the work on an ebook reader gives no browsing or search advantage. Calibre is one of the tools used for conversion to EPUB.

Akoma Ntoso

Akoma Ntoso (meaning 'linked hearts' in the Akan language of West Africa) is an international XML-based

https://doi.org/10.1017/S1472669616000530 Published online by Cambridge University Press

standard for legal documents. It was designed for various documents (laws/bills, debates, etc) and to accommodate the needs of many countries and organisations.

The Japanese government encourages adoption of international standards, but a group from Nagoya University spoke about the difficulties they have in converting documents compliant with the Japan Statutory Schema (JSS) into Akoma Ntoso format. It takes 84 conversion rules for 67 JSS elements to be converted to the Akoma Ntoso standard. Rules for 37 elements (for supplements and amendment acts) are still under consideration.

Akoma Ntoso is much more flexible than the JSS, leaving the potential for ambiguity. For example, Akoma Ntoso places no restrictions on the order of elements, so <subpara> can (and does in some cases) go before <para>.

European Case Law Identifier (ECLI)

Marc van Opijnen spoke about the European Case Law Identifier (ECLI). The ECLI ecosystem contains identifiers and metadata. The identifiers are human and computer readable, and don't necessarily replace national IDs. They have five elements in a fixed format: ECLI:country code: court identifier:year of decision:specific identifier. A High Court judgment from the Netherlands in 2012 would look like this: ECLI:NL:HR:2012:938.

The ECLI system also uses uniform metadata to improve search facilities for case law. This metadata follows the Dublin Core standard, and includes nine mandatory and eight optional elements. The system has to allow for different language versions, eg, an English summary of a Spanish case. Controlled vocabularies are used for the type of decision and the field of law. Version information, coverage and date are examples of mandatory elements; subject and abstract are among the optional elements.

One objective is to improve access to case law by creating linked open data, eg, to find all works on Supreme Courts mentioning 14–0871. Consistent depiction of cases would also be of benefit to indexers who are creating Tables of Cases, where variant forms of the same case name often have to be edited. Differences include the use or not of abbreviations (eg, the ghastly 'Anor' and 'Ors' for 'Another' and 'Others') and the inclusion or not of 'Pty Ltd' and other extras.

ECLI has been implemented by the European Union Court of Justice, the European Patent Office and several EU Member States.

European Legislation Identifier (ELI)

Jean-Michel Thivel and Manuel Siaud spoke on the use of the European Legislation Identifier (ELI). ELI is to legislation much the same as ECLI is to case law. Their aim is to further digitize public services to increase transparency and efficiency and to promote democracy. Or, as someone put it more casually, 'the job is to hand out data so people can do things with it'. ELI is user friendly, and provides a flexible framework for identifiers, metadata, datasets and ontologies.

The ELI is unique for each act and is structured by putting together the elements that describe the act, divided by slashes, eg, /eli/country/agent/year. It is possible to guess the identifier if you know the date of the act. In response to a question about searching, the speakers explained that all of the sub-components are optional except 'eli' and can be presented in any order. Search works as a filter, looking at each sub-component and narrowing the results set as search terms are added.

Metadata is used to describe 31 properties of the legislation. It is created by experts or machines.

Dublin Core

I was interested to see the use by legal informaticians of a standard that is also used in libraries and metadata indexing (Dublin Core) and another one used in libraries (FRBR).

The Dublin Core Metadata Element Set defines fifteen metadata elements for resource description in a cross-disciplinary information environment. ECLI follows the Dublin Core standard for its nine mandatory and eight optional elements. ELI encourages the use of relevant metadata elements to further describe legislation, and specifically recommends DCTERMS for country codes and language.

Functional Requirements for Bibliographic Records

Functional Requirements for Bibliographic Records (FRBR – pronounced FrrBrr) is a model developed by the International Federation of Library Associations and Institutions. It distinguishes between works, expressions, manifestations and items. This is important in libraries for showing the relationships between items – eg, the original concept, a translation, a new edition, and identical copies of a work.

ECLI and ELI both use the FRBR model as a work-level identifier. One conference attendee noted that it takes a lot of work to adapt FRBR for United States legislation. While the FRBR model works for legislation consolidation and amendment where the act is republished (and therefore becomes a new expression), it doesn't work so well for posthoc modification as done in the United States (where the work is continually modified).

LEGAL APPLICATIONS OF UNSUPERVISED TOPIC MODELLING

Tom Bruce and Sara Frug from Cornell LII spoke about the legal applications of unsupervised topic modelling.

Topic modelling is a tool for discovering 'hidden' topics (aboutness information) in large, often

https://doi.org/10.1017/S1472669616000530 Published online by Cambridge University Press

unstructured, collections of documents. The speakers describe it as sorting hay into smaller, thematic haystacks, but not finding the needle in the haystack.

In topic modelling, a document collection represents a discourse, a discourse contains topics, and topics are clusters of associated keywords often found together. The user guesses the number of topics, and machine-learning software generates topic clusters. Topics are labelled, and the results are refined by tuning stopwords and changing the number of topics to be generated.

Topic modelling is good for finding what topics are covered in a big collection, and for comparing discourse between collections, eg, seeing the differences in statutes from different states (eg, a landlocked one and a coastal one) and seeing how justice had changed over time (temporal comparisons).

Bruce and Frug used MALLET software (Machine Learning for Language Toolkit – http://mallet.cs.umass.edu/), and trained it with 25,000 US court decisions. All of the documents were given a topic number, label, and series of keywords. Stopwords are tuned to exclude jurisdiction so the results focus on subject matter (a single discourse).

The speaker said that this manual work was done 'by a hapless third year law student'. When I commented later that this was the bit I found most interesting, he said that with automated data mining he uses the 'last pickle in the jar' analogy. Someone asked a pickle packer how pickles fitted so well in the jar, and was told 'we always do the last one by hand'. Similarly, with automated data mining, some human attention is needed to ensure the quality of the final result.

Human input includes terminology extraction, so that the system can work with phrases not terms (eg, it separates 'collateral', 'collateral estoppel', 'estoppel' and 'collateral damage'). Stemming and lemmatization are also used, so a search for 'whale' also finds 'whales' (while making sure that a search for 'damage' doesn't retrieve 'damages').

Normalization of terms is also necessary, grouping the concepts 'Indigenous', 'Native American' and 'American Indian'. This work is labour intensive, and the results are corpus specific. The researchers tried some prominent Australian cases (including the Mabo native title case), and found many that worked badly. Terms such as 'Indian reservations' were not used in Australian documents, and even seemingly straightforward concepts such as natural resource rights, conveyances/deeds and authority didn't transfer well. For this system to work in Australia it would have to be trained on Australian materials.

In a large-scale, automated system such as this, you expect some junk topics, and they got less than 10%. For example, for a case where a student who injured their knee while kickboxing was suing a gym, the topics identified were 'premises liabilities, proximate cause, student employment, summary judgment damages'. The topic 'student employment' is a false drop.

In a previous conference presentation (http://conference.cali.org/2015/sessions/topic-modeling-swiss-army-knife-faculty-geeks-and-librarians), the speakers discussed potential uses for topic modelling, saying 'we're thinking about using them to:

- replicate hand-constructed indexes to large corpora like the Code of Federal Regulations or the Congressional Record
- discover the differences between the discourse surrounding crime and criminality in the 1980s and that in the period starting around the year 2000
- construct finding aids for large, confusingly titled bodies of guidance documents such as IRS written determinations and SEC no-action letters
- figure out what's in all those Congressional committee prints.

CONCLUSION

Important work is being done with legal data to obtain new information, and to improve access for all users. Technology is essential for coping with large data sets, but expert human input is still crucial to planning and quality control in these projects. Applying the concept of gravity assistance, perhaps indexers should be making more connections in the field of informatics.

Footnotes

[1] See an earlier article on indexing in AustLII: Building a global legal index: a work in progress, by Madeleine Davis, *The Indexer*, 2001, vol. 22 no.3, pp. 123–127, http://www.theindexer.org/files/22-3/22-3_123.pdf

[2] Wiktionary defines informatics as 'A branch of information science and of computer science that focuses on the study of information processing, particularly with respect to systems integration and human interactions with machine and data' and informatician as 'someone who practices informatics'. I have previously written about health informatics and its intersection with librarianship (Browne, Glenda. 'Forward-looking seminars: AGLIN and HLA in Canberra', *Online Currents*, v.25, pp. 305–312, http://www.webindexing.com.au/forward-looking-seminars-aglin-and-hla-in-canberra/).

[3] Xu Shu. 'The Shen Bao Index: its academic significance and effect on the development of Chinese indexing', *The Indexer*, Vol. 33, No. 4 December 2015.

[4] Lihosit, Judith Research in the Wild: CALR and the Role of Informal Apprenticeship in Attorney Training, *Law Library Journal*, Vol. 101:2 [2009–10] 157–176, http://www.aallnet.org/mm/Publications/llj/LLJ-Archives/Vol-101/pub_llj_v101n02/2009-10.pdf

https://doi.org/10.1017/S1472669616000530 Published online by Cambridge University Press

Biography

Glenda Browne has been a freelance indexer of books, journals and websites in a wide variety of subject areas since 1988. She is also a medical librarian one day per week. She is co-author of Website indexing and The indexing companion, and author of The indexing companion workbook: book indexing. Glenda was awarded Highly Recommended in the ANZSI Medal for her index to The Indexing companion. Glenda teaches indexing at Macleay College and for ANZSI and other professional groups, and has been the ANZSI representative on the IDPF EPUB Indexes Working Group. She is ANZSI Education Officer, and the inaugural convenor of the NSW Indexers and Education Groups. More information at: http://www.webindexing.com.au/

Legal Information Management, 16 (2016), pp. 269–272

doi:10.1017/S1472669616000542

Current Awareness

Compiled by Katherine Read and Laura Griffiths at the Institute of Advanced Legal Studies

This *Current Awareness* column, and previous *Current Awareness* columns, are fully searchable in the *caLIM* database (Current Awareness for Legal Information Managers). The *caLIM* database is available on the Institute of Advanced Legal Studies website at: http://ials.sas.ac.uk/library/caware/caware.htm

The 'Cardiff Index to Legal Abbreviations' is available at http://www.legalabbrevs.cardiff.ac.uk/

EUROPEAN UNION

Nick Barber, Tom Hickman and Jeff King 'The Article 50 'Trigger'' (2016) August Counsel: the Journal of the Bar of England and Wales 18

Mateja Durovic, *European Law on Unfair Commercial Practices and Contract Law* (Hart Publishing 2016)

Evanna Fruithof, Alexandria Carr and Gordon Nardell 'A New Relationship' (2016) August Counsel: the Journal of the Bar of England and Wales 20

Patrick Overy 'European Union. Guide to Tracing Working Documents. Update' (GlobaLex July/August 2016) http://www.nyulawglobal.org/globalex/European_Union_Travaux_Preparatoires1.html

Anne Peters, *The Freedom of Peaceful Assembly in Europe* (Hart Publishing 2016)

Peter Thompson 'Roll up that Map' (2016) 166 (7710) NLJ 8

Stephen Weatherill, *Law and Values in the European Union* (Oxford University Press 2016)

Michael Zander 'Act in Haste' (2016) 166 (7707) NLJ 6

INFORMATION POLICY

Eric Barendt, *Anonymous Speech: Literature, Law and Politics* (Hart Publishing 2016)

Maria Biasiotti and Sebastiano Faro 'The Italian Perspective of the Right to Oblivion' (2016) 30 International Review of Law, Computers and Technology 5

Antoon De Baets 'A Historian's View on the Right to be Forgotten' (2016) 30 International Review of Law, Computers and Technology 57

Claudia Kodde 'Germany's 'Right to be Forgotten' – Between the Freedom of Expression and the Right to Informational Self-Determination' (2016) 30 International Review of Law, Computers and Technology 17

Ronald J. Krotoszynski, *Privacy Revisited: A Global Perspective on the Right to be Left Alone* (Oxford University Press 2016)

Ian Long, *Data Protection: The New Rules* (Jordan Publishing 2016)