

Markup and Metadata in International Criminal Law: A Case Study

Will Lowe¹, Olympia Bekou², and Emilie Hunter²

¹Methods and Data Institute, School of Politics and International Relations, University of Nottingham

²Human Rights Law Centre, School of Law, University of Nottingham

Email address of corresponding author: will.lowe@nottingham.ac.uk

Abstract. This paper sketches an ongoing project to construct a database of national implementing legislation of the Rome Statute establishing the International Criminal Court. The database applies a metadata scheme for international criminal law to legal texts from multiple countries, allowing comparative legal and socio-legal analysis at the paragraph level. In addition to marked up legal texts the database contains other materials for legal and international political context. The aim is to provide a more fully-connected dataset for comparative research in international law and politics, whilst also investigating metadata and markup issues specific to legal texts. We sketch the issues that arose during the project and the lessons we learned.

Project Background

The International Criminal Court (ICC) was established in 1998. Under its statute it has jurisdiction over genocide, crimes against humanity and war crimes. States that ratify the Rome Statute are under an obligation to incorporate the ICC cooperation regime into domestic law and are encouraged, since the ICC is complementary to national courts, to adopt legislation allowing them to prosecute nationally the crimes falling under the ICC's jurisdiction. Aside from these legal complexities there is also a political aspect to the database; complementarity creates a complex set of institutional obligations and incentives for states, even when those states do not ratify the Statute and have no formal role in the regime it creates. Of particular interest is the United States which has pursued bilateral immunity agreements ensuring that its personnel will not be affected by ICC actions. Variation between states is naturally substantial: national cultures, legal otherwise vary widely; legal systems are structured differently, e.g. between common law, civil law, and mixed systems; and individual states also differ in their understanding of the scope of their national legal system under the ICC regime. For the to provide an effective comparative view of this aspect of international law, these differences must be reflected in an encompassing search and metadata scheme.

The implementation database described in this paper (henceforwrd IDB) is the ongoing result of a collaboration between the Methods and Data Institute and the Human Rights Law Centre at the University of Nottingham. The IDB also forms part of ICC's own Legal Tools Project, an umbrella for academic and NGO projects around the the topics of the Rome Statute. Legal Tools projects include a unified war crimes prosecutors toolkit, the Case Matrix, developed at the University of Oslo, a database of Nuremberg war trials records at the University of

Marburg, and of the International Criminal Tribunal for the Former Yugoslavia, also at Oslo. The Court's Legal Tools project provides the international institutional context for the IDB.

Project data will ultimately (when HRLC's confidentiality agreement with the Court ends at the end of 2008) be available in two forms. It will be publicly available as part of the ICC's Legal Tool Project website in searchable and browsable form. Secondly it will be available in updated form linked from the Methods and Data Institute's pages. The Court's version of the data will focus on the practical legal aspects of the data whereas the Institute's version will continue to develop the political elements

The Problem

The aim of the IDB is to provide a searchable database of implementing legislation that is structured in four ways. First the documents must be available in original forms with a free text search function. Second, each document must be structured with metadata from the Court that is common across the legal tools. This metadata is intended as the lowest common denominator for describing content in all Legal Tools projects. Third, text spans of each implementation must be searchable using keywords corresponding to different aspects of international criminal law, viz. cooperation with the ICC, the crimes described, jurisdiction issues, procedural issues, and matters of substantive law. Fourth the same segments must be searchable according to the section of the Rome Statute that they implement. This latter segmentation allows a comparative view for legal or political scholars interested in comparing the details of specific country's implementation of e.g. provisions relating to torture, or the treatment of prisoners of war. The third (keyword) segmentation never makes coarser distinctions than the sections of the Rome Statute and is therefore derived data not requiring a separate markup task. Fifth, details of the geography, legal system, and treaty obligations of the state author of each piece of legislation must be available to joined with textual data for comparative purposes, or examined alone. An example of the latter analysis would be the relationship between the date a state signs or ratifies the Rome Statute, and the date it signs an immunity agreement with the United States (if it chooses to do so). This gives some measure of each state's resistance to competing diplomatic pressure from the U.S. and any regional pro-ICC grouping it may belong to such as the European Union.

A Technical Solution

The technical solution to this problem is simply described: We started with an Access database containing documents, document sections, document, section and country level metadata schemes and their relationships. We subsequently migrated to a web application framework¹ to provide a more consistently and accessible way to query the data. There are two points to make about this description. First: it is, from the point of view of computing, trivial; using mature open-source components to create a database-backed web application is a problem solved by programmers every day. Second, this description entirely hides the basic difficulties encountered in the project. We briefly consider the consequences of the technical straightforwardness before moving to practical difficulties.

¹ Django (python) and MySQL both hosted remotely.

Between Two Stools

This is the third university-internal ‘documents, metadata, and search’ project undertaken by the Methods and Data Institute, and the second also involved legal data on insolvency cases and the other is a data set of militias and violent pro-government groups, jointly directed with researchers at the University of Aberdeen.² In each case none of our social science or law researcher audiences were capable themselves of implementing appropriate technical solutions. Moreover, our experience suggests that for every data structuring exercise we hear of at least five more are pursued without technical assistance. Such data sets are typically managed with file naming conventions, in Excel spreadsheets, or inside statistical packages such as Stata or SPSS. The minimal metadata and weak or absent relational structure made available by these packages seldom makes best use of the data.

On the other hand, few such projects require or would benefit from any currently available explicitly grid computing or text mining resources we are aware of at Nottingham or elsewhere. These not only perform much more complex functions, but often fail to provide the requisite metadata or relational structure necessary, and are simply too sophisticated to be usable in the small to medium scale projects pursued by our audience. This appears to be because e-Social Science research has historically been concerned with the provision of compute power to relatively generic data analysis problems, whether expressible as mathematical optimization problems such as fitting large scale microeconomic models (Gros et al. 2006), massively parallelizable algorithms such as those necessary to run large scale agent simulations (Birkin et al. 2006), or localized low level but domain-specific extraction processes, such as term or entity extraction in computational linguistic problems (e.g. Ananiadou 2007; Forsyth et al. 2006). The increasing focus on qualitative data has brought a return to traditional representational concerns. The problem for this domains is effective representation and conventional search rather than efficiency of analysis. The corresponding challenges are flexibility of metadata scheme, comparability across document collections, and targeted search. These are, in short, all of conventional problems of digital asset management familiar to librarians and archivists everywhere but previously addressed by proprietary software infrastructure implemented at the institutional level.

Legal Data

Legal materials are a particularly interesting case for e-Social Science applications for at least four reasons. First, whatever one’s view of legal positivism (the theory that laws are normative in themselves and do not retain their normative status by virtue of any underlying ethical or power structure) it is clear that laws themselves *are* structured normatively, so the usual social scientific assumptions about measurement and statistical dependencies are not obviously applicable. Second, the legal materials in the IDB are intended to be used for instrumental, historical, *and* social scientific purposes: section level decompositions are used instrumentally by participants, NGOs or private citizens for understanding the application of international criminal law in war crimes and genocide trials; historians may use the document level information to construct accurate chronologies and track diplomatic relations, and social scientists may use the metadata itself to study diplomatic and institutional constraints on state action. For example, ratification dates are data for political scientists interested in diplomatic

² Accountability and Government Militias. ESRC grant RES-062-23-0363

pressure and international organizations, but metadata for lawyers interested in comparing the structure of implementations across Gulf states.

Existing Legal Data and Metadata

Lawyers needing access to legislation, cases, and commentary for national law have a wide variety of systematically organized, albeit proprietary, digital library resources such as Westlaw³, LexisNexis⁴, and Heinonline⁵ for older materials. In contrast, international law materials, particularly in the areas of international criminal law are surprisingly weakly covered in open sources. Users must typically search each Court or Tribunal website individually, websites are seldom similarly structured, and often provide little more than a collection of relevant documents comprehensible primarily to other international lawyers. This applies also to the International Criminal Court itself and particularly to the International Criminal Tribunal for Rwanda and for the Former Yugoslavia. One of the less emphasized aspects of the original e-Social Science agenda was widening access to digital materials. The IDB, as part of the Legal Tools project is intended as a small part of this process.

Turning from existing legal collections metadata schemes we found that many cover only one tradition of law, e.g. common law or focus on one country only e.g. the US Legislative Indexing Vocabulary⁶, or the UK's Office of Public Sector Information.⁷ More legal metadata schemes exist in commercial products, but these are often prohibitively expensive and often also semantically opaque. Correspondingly much academic work has also been done on legal vocabularies and legal informatics, e.g. the Metalex, and Ontologies for Legal Information Sharing projects. Unfortunately these proved considerably too complex to apply for the IDB project. We were left then to construct our own metadata scheme. What did we learn from the exercise?

Lessons Learned

How Small Projects Happen

The ICB project started as an extended consulting session between members of the Human Rights Law Centre (including the third author) and the Methods and Data Institute (the first author). With intermittent small bursts of funding from various governments, and volunteer coders from the Law Department, the project evolved from being a web page with documents to a small relational database and some simple keyword metadata. The first six months of the project were primarily devoted to the HRLC putting project management structure in place, and the MDI motivating, explaining, revising and re-revising metadata schemes.

³ <http://www.westlaw.com> accessed 25/05/2008

⁴ <http://www.lexisnexis.com> accessed 25/05/2008

⁵ <http://www.heinonline.org> accessed 25/05/2008

⁶ <http://thomas.loc.gov/liv/livtoc.html> accessed 25/05/2008

⁷ <http://www.opsi.gov.uk/> accessed 20/03/2008

Metadata Construction is Pedagogy

The latter process made it very clear that the sociological process of constructing metadata schemes and the pedagogical process of explaining the structure of international law were substantially the same. Specifically, the task was, first to map legal distinctions onto to the set of currently well-understood metadata structures, and second to find the subset of these that could be implemented by one person quickly enough not to slow coding and markup procedures from our volunteers.

While our discussions took a winding path through keyword lists, hierarchically structured keyword lists, thesaurus structures, taxonomies, and topic maps, a fundamental misunderstanding did emerge whose resolution provided sufficient clarity to move forward, viz. the difference between using metadata and relational structure to *represent*: structuring data and metadata to reflect the logical structure of the law, and using it to *organize*: structuring data and metadata to facilitate specific searches and audiences (see Rorty, 1979 and van Frassen, 1980 for the contrast). The representational task turns out to be quite familiar for lawyers, particularly the second author, who has considerable experience drafting as well as working with existing legislation. It was therefore not simply practical constraints that lead to a metadata scheme that does *not* reflect, in any deep sense, the structure of legal text, but nevertheless facilitates expert search and browsing.

In fact we remain agnostic about the superiority of our chosen approach to metadata structure and have also applied for EU project funding with a consortium of computer scientists to investigate the utility of a full-scale representational approach to our data based on extensions to OpenCyc and text mining methods.

Don't Expect Requirements

Ideally requirements and scope are relatively clear at the beginning of a project. This was not the case for the IDB, and we do not expect this ever to be possible in quite the way expected by software engineers. In a project that crosses two disciplines each side must learn at a minimum what is trivial, difficult, or impossible in the other's frame of reference. HRLC expanded their view of the what is possible with automatic search, but contracted it with respect to what is straightforward in constructing an exhaustive metadata code book. MDI were forced to rethink their previously clear-seeming distinction between normative and descriptive data. It is inconceivable that we could have come to this understanding without having engaged in the project first.

Build Several to Throw Away

The IDB project is small and cheaply pursued compared to most e-Social Science projects. A significant advantage of this is that the technology could constructed could be quickly broken down and repeated rebuilt to reflect substantively motivated changes in data and metadata structure; we 'threw away' at least four complete metadata entry systems (Brooks, 1975). If we had technology available that was more tailored to our initial view of the problem we expect our results would have been worse.

Documents, Metadata, and Web Problems are Ubiquitous

The IDB, like a large number of social science data problems reduce, technically speaking, to document and metadata management with web access and search. Since these appear to be solved problems in computer science, it is easy to forget that they still lie between two stools for most social scientists: the technology necessary to solve them is more complex than using a spreadsheet but often not complex enough to motivate computer science collaborations, and certainly beyond the reach of most social scientists and lawyers. Nevertheless this machinery is an excellent candidate for centralized provision in the manner of grid computing.

Conclusions

We have briefly sketched the Implementation Database project, some of the process of construction, and described what lessons we learned from it. IDB is in many ways an atypical e-Social Science project - it is much smaller than a grid computing application, makes no use of statistical or parallel processing machinery, and leverages no more than relational representation structures. Nevertheless it is, we claim, very representative of a class of problems facing social scientists for which there is no large scale centrally provided technical solutions available. We think there should be.

References

- Ananiadou, S. (2007) The National Centre for Text Mining: a Vision for the Future, *Ariadne* (53)
- Birkin, M., A. Turner, B. Wu (2006) A Synthetic Demographic Model of the UK Population: Methods, Progress and Problems. In the Proceedings of the 2nd International Conference on e-Social Science, Manchester UK
- Brooks Jr., F. P. (1975) *The Mythical Man Month*. Addison Wesley
- Forsyth, R., S. Ainsworth, D. Clarke, P. Brundell and C. O'Malley. (2006) Linguistic-computing methods for analysing digital records of learning. In Proceedings of the 2nd International Conference on e-Social Science, Manchester UK.
- van Frassen, B. (1980) *The Scientific Image*, Oxford University Press.
- Grose, D., R. Crouchley, T. van Ark, R. Allan, J. Kewley, A. Braimah, M. Hayes (2006) *sabreR*: Grid-enabling the analysis of multi-process random effect response data in R. In the Proceedings of the 3rd International Conference on e-Social Science, Manchester UK
- Rorty, R. (1979) *Philosophy and the Mirror of Nature*, Princeton University Press.