A METHODOLOGY FOR PUBLISHING BUILDING REGULATORY DOCUMENTS BASED ON OPEN DATASETS

Lewis McGibbney[1], Bimal Kumar[2]
[1]lewis.mcgibbney@gcu.ac.uk
[2]bimal.kumar@gcu.ac.uk
School of Engineering and Built Environment
Glasgow Caledonian University
Glasgow UK

ABSTRACT

In recent years, the global issues surrounding sustainable development of our land and resources have become (amongst others) a huge area of economic as well as historic interest. The significant majority of construction work is subject to rigorous sustainable planning procedures and regulatory compliance measures; however in the last decade, one can witness a marginal and continual increase as we strive to ensure the longevity of design decisions to projects ranging from new buildings, extensions to external appearance of buildings, residential or commercial. Generally speaking, factors which characterise the drafting practice of primary and secondary legislation within the domain of Planning and Sustainable Development involve arguments including the extreme complexity and difficulties we encounter when attempting to model the domain, which is partly due to the inherited nature of historic development and national relevance (amongst many others). Additionally, it is observed that there are high degrees of inconsistency within the work-flows designed to enforce the aforementioned legislation. In this paper, we propose a new methodology for legislation/regulation drafting that addresses these issues with such documents. We show that the creation, amendment, authoring and publishing of a subsection of UK Secondary Legislation, in particular Scottish Technical Standards, can be undertaken in a fashion which embraces the fast moving area of Open Government. This paper describes an attempt at developing a standards focused, XML-based version of the Scottish Technical Standards which enables them to be processed outside of the source code repository. Processing has three phases, i.e. mapping legacy elements to Crown Legislation Mark-up Language, resolving identifiers and entity extraction based upon ifcXML dictionary definitions. We categorise this research within the initial stages of a semantic annotation process with the ultimate goal of opening up the information in these documents as semantically rich open datasets which will make their usage much more transparent and meaningful.
Keywords: Crown Legislation Mark-up Language, ifcXML, Open Government, Planning, Scottish Technical Standards

1. INTRODUCTION

Globally, in recent years social and political perceptions regarding how public data is obtained or produced, authored, amended and disseminated have changed beyond all formal recognition. The fundamental driver behind this global paradigm shift is our necessity to increase the openness and transparency of our democratic states, whilst in the process embracing our digital world by linking past, present and future segments of our society. Rob McKinnon of TheyWorkForYou.co.nz argues "It's much broader

than just government; government is actually just one participant in our society. What we're seeing happening is actually the cusp of a major social change, a global social change. Together we have the ability to transform the way that society works; globally we're bringing together through sites such as Wikipedia or collections of structured data, a global overview of how society works and how we organise ourselves. So we're just at the beginning of a major change in the way we operate. " (Working Group on Open Government Data 2012) As the argument gains momentum, we increasingly hear that open data is the digital fuel of the 21st century, however a huge part of this process involves making Government information more accessible to its users. With this ethos in mind, this research delves into the domain specific problems encountered when we consider opening up construction legislation within this paradigm shift.

As a subsection of the UK planning system, both the Scottish planning system and the legislation used within it are artefacts modelled to ensure we have control over changes to the built environment within Scottish towns and cities. In practice however the tasks of ensuring that compliance is met and that change is controlled are clouded by the extremely complex, highly subjective, performance based legislation produced as a result of dated drafting work-flows. This research proposes a novel methodology for authoring and publishing Scottish Statutory Instruments, in particular Scottish Technical Standards (STS) as Open Linked Data for widespread use at local and national level as well as by professionals and the public alike. This research describes an XProc (W3C 2012) managed authoring work-flow for developing an RDF/XML-based version of the STS, modelled on on-going work at legislation.gov.uk (the National Archives 2012); the most successfully recognised global implementation of open legislation. This paper is structured as follows: Section 2 provides a comprehensive account of related work, the authors have been extremely fortunate to model our current research on an existing production implementation being used to empower millions of individuals within the UK with Open Government Legislation, globally this production implementation is the largest and most mature within the open legislation paradigm. Section 3 delves into our background research; focused on extracting and combining intricate but extremely relevant literature within the fields of legal and construction informatics, the open data movement, and Scottish and UK legislation. Section 4 substantiates on the underlying target data model. Section 5 underpins our experiences with getting STS to work within the open data archetype. Finally we conclude with future work and acknowledgements.

2. RELATED WORK

According to Rufus Pollock of the Open Knowledge Foundation "There are three ways that open government data can make the world a better place... number one;...it enables companies, individuals, not for profits to go and build interesting, useful, valuable applications and services. Number 2; I think it's about democracy; it's about participatory government it's about transparency and accountability it's an ability for us to see what our governments are doing. Number three; I guess my one is Why not? It's basically costless in most cases or close to costless to open up government data in most cases, why not open that government data up? It's already there, it's already been collected." (Working Group on Open Government Data 2012)

The notion of improving access to legal resources has been a long standing area of concern within public life; subsequently the problem has attracted interest from researchers within informatics and legislation research communities, with a

noticeable increase of interest within the last decade or so. Globally our (sometimes unbelievably) complex legislation drafting and publishing work-flows denote that with increasing frequency, a huge part of legislation compliance actually involves finding the actual resources we require. Inevitably this daemon of our own creation attracted lots of attention from commercial entities, namely LexisNexis® , Westlaw® and Justis to name a few. All of these services offer some level of search functionality, primarily aimed at simplifying the process of legislation compliance.

One particular example of work which played a fundamentally important role in improving both the direction and substance of this research was Hoekstra's *The MetaLex Document Server* (Hoekstra 2012), which explicitly details the process of publishing national regulations in the Netherlands. Hoekstra expands to detail the inconsistencies present within legacy XML-based regulations and justifies why it is necessary to transform them to open data syntaxes; namely CEN MetaLex[1], RDF[2] and other open standards, which can subsequently be used as open linked data across the semantic web.

Interestingly a familiar pattern emerges when we begin to understand why more legislation is not published openly as *5-star open data.* In the overwhelming majority of cases we witness web services serving legislation which is "fancy and costs a tonne, but is not flexible." In the context of this specific topic area, these certainly justify why we may wish to question the simplistic "Why not?" approach proposed earlier by Pollock. When scrutinised more detail, one sees that the aforesaid merely begins to describe the problems we face when attempting to publish legislation as open linked data, therefore it is excruciatingly important to emphasise the following points:

1) Legislatures/parliaments are full of very structured documents: bills, resolutions, journals, calendars, statutes, annotations...all have readily apparent structure.
2) XML is all about handling very structured documents.
3) Therefore, classic XML approaches fit legislatures/parliaments.

Unfortunately, according to (McGrath 2010), there are a variety of reasons why the above analysis is wrong[3], namely;

1) The centrality of line/page number citation in amendment cycles
2) The complex nature of amendatory actions
3) The critical nature of fidelity with historically produced renderings
4) The fluid nature of work-in-progress legal assets
5) The complexity of amendment cycle business rules that often pre-date computers and cannot be changed to make life easier for modern software
6) The subtle inter-play between legal content and legal content renderings
7) Content aggregation and derived document types

With this in mind, we now focus predominantly on the most influential work relating to our own research; The UK National Archive's legislation.gov.uk.

---

[1] http://www.metalex.eu/

[2] http://www.w3.org/RDF/

[3] It is also extremely important that Mr McGrath's comments are represented in the correct context, therefore we should add that "XML *has* an enormous role to play in legislatures/parliaments but it cannot be simply applied blindly per the standard XML value model without causing significant problems" (McGrath, 2010)

"We had two objectives with legislation.gov.uk: to deliver a high quality public service for people who need to consult, cite, and use legislation on the Web; and to expose the UK's Statute Book as data, for people to take, use, and re-use for whatever purpose or application they wish. In particular, our aim was to show how the statute book can contribute to the growing Web of data as well as to the Web of documents." (Sheridan 2010) In appraising the above, the outcome has resulted in the best globally recognised example of a web service aimed to satisfy the increasing requirements upon publicly funded institutions to publish their legislative data in formats users are able to consume. Hoekstra states that as of May 2011 the MetaLex Document Server maintained a store of some 119 million triples, which he compares to legislation.gov.uk's 1.9 billion. Further into his dissection of legislation.gov.uk's genetic make-up, Sheridan explores why the CEN Metalex standard utilised by Hoekstra is not descriptive enough in its entirety, stating that "general interchange formats for legislation …lack the expressive power we need for UK legislation".

The resulting XML-based datasets we produce within the scope of this research are modelled on top of the existing data interchange standards adopted at legislation.gov.uk. Using the Crown Legislation Mark-up Language (CLML) as our target syntax we elaborate to state intricate differences involved in modelling specifically STS as open, linked UK Secondary Legislation and Scottish Statutory Instruments data.

3 BACKGROUND RESEARCH

3.1 A Brief History of the Building (Scotland) Act 2003 (The Act)
The Building (Scotland) Act 2003, as passed on 20th February 2003 states:

1. Scottish Ministers may, for any of the purposes of -
   a) securing the health, safety, welfare and convenience of persons in or about buildings and of others who may be affected by buildings or matters connected with buildings,
   b) furthering the conservation of fuel and power, and
   c) furthering the achievement of sustainable development,

   make regulations ("building regulations") with respect to the design, construction, demolition and conversion of buildings and the provision of services, fittings and equipment   in or in connection with buildings.

"The Act gives powers to Scottish Ministers to make building regulations, procedure regulations, fees regulations and other supporting legislation as necessary, to fulfil the purposes of the Act. The purposes include setting building standards and dealing with dangerous and defective buildings. The various regulations are made by Scottish Ministers, but must be approved by the Scottish Parliament before coming into force" (Statute Law Database 2003). From the inset it is essential for us to identify and disambiguate two elements of utmost relevance to this research

1. the legislative ecosystem (The Scottish Building Standards system (SBSS)) developed to manage the processes used to fulfil the purposes of the Act
2. the subsequent secondary legislation (the STS) produced to ensure construction oriented activities uphold firm compliance within the legislative ecosystem

3.2 The Scottish Building Standards System

At its heart the SBSS comprises of The Scottish Governments Building Standards Division (BSD) within the Directorate for the Built Environment, which on behalf of Scottish Ministers "prepares and updates building standards legislation and guidance documents, conducting any necessary research and consulting as the Act requires." In addition the BSD "approves verifiers and certifiers of design and construction... and checks how verifiers (including local authorities) and certification scheme providers are operating the system." (The Scottish Government 2010) Effectively the BSD operates at the top tier of the SBSS.

Working in parallel directly under the BSD we have the Scottish Association of Building Standards Managers (SABSM), comprising of principal officers from the 32 Scottish local authorities, the building standards divisions within local authorities themselves who manage the execution of tasks specific to the SBSS within their own geographical authority, and finally Verifiers and Certifiers; undertaking the tasks of independently checking applications for building warrants, and certifying that certain design or construction work complies with the building regulations respectively. The BSD has contact with all the aforementioned parties, in turn ensuring that the SBSS represents an ecosystem capable of working to the provisions of the Act. Central to the SBSS ecosystem exist the STS themselves, they define the accepted standard to which all domestic and non-domestic construction work must comply. In our early research however we have become aware of the noticeable discrepancies which exist within the workflows being used by verifiers to ensure that construction building warrant applications are granted on the basis of reasonable enquiry and regulatory compliance. We describe this as "the regulation loophole", where a *grey area* exists between how the BSD *envisage* or *intend* STS to be used, and how they actually *are* being used in practice. To summarise, the loophole exists due to ambiguities in the SBSS, which can be unintentionally used to circumvent or otherwise avoid the intent, implied or explicitly stated, of the system.

3.3 Discussion on a Future Model for Reasonable Enquiry

In 2011 the BSD commissioned two reports resulting from a research project undertaken to "inform the decision to be made... concerning the appointment of verifiers under the Building Scotland Act 2003." The first of the two reports highlights that "...a number of issues relating to "reasonable enquiry" were identified. The main issue identified was the lack of consistency in the approaches of verifiers to reasonable inquiry through site inspection when accepting or rejecting a completion certificate submission." (Optimal Economics Ltd 2011) The same report continues to provide comprehensive details of respondents from local authorities, trade bodies, developers and consultants regarding topics surrounding compliance consultation and guidance for setting measurable performance standards for verifiers regarding suitable levels of site inspection. "Almost all respondents (96%) to the Consultation indicated that there was a need to establish greater certainty on compliance with building standards during the construction stages. Moreover, 82% of respondents considered that the balance between plan checking and inspection was not correctly weighted." In particular respondents thought it was important to establish compliance during construction, particularly in relation to works which would be covered up, namely;

1. Use of modern complex materials and solutions used to meet standards
2. Alternative performance designs used in lieu of guidance solutions
3. Elements of design that occur during the construction works
4. Environmental impacts of non-compliance

5. Possible health risks due to non-compliance
6. Cost and inconvenience of non-compliance

To build on the aforesaid, we hope that with the above in mind, it is clear that STS should not be opened up simply to satisfy some global paradigm shift towards open government, or in fact to promote transparency, neither should they be opened up merely to release social or commercial value, or to improve participatory governance from successive Government organisations (although any one of these points alone, never mind combined, merit clear mention and justify a requirement for work to be carried out), they should however most importantly, be opened up because there is a firm requirement for us to mitigate against the current failures of individuals to consistently infer information from the SBSS which in turn is having a detrimental impact throughout every aspect of how we manage, develop and maintain our built environment. The argument has been made that on one hand we either acknowledge that there is work to be done, and that it must start by making the STS, and a means of complying with them more user friendly and achievable, or we do nothing and continue de facto rather than de jure which is what we should be striving to achieve.

4 DATA MODEL

4.1 Some Musings on the Anatomy of Scottish Technical Standards
Before we can begin modelling our target RDF datasets, a hugely significant prerequisite involves undertaking extensive and fine grained content (or textual) analysis on the source STS. In abstract terms this involves a fully explorative study into the following, (i) the logical structure which the source and target texts must conform to; specifically undertaking extensive study and understanding of the typically hierarchical nature we find within legislative resources as well as the specific structure the STS conform to, (ii) the semantics of domain specific terminology; including additional content representation features scattered throughout the documents e.g. graphs detailing calculations for space heating, tables providing U-value measurements and equations focussing on energy performance compliance to name a few. Although we argue that both (i) and (ii) are inextricably linked, it is essential that in the process of developing the target data model, we attempt to separate presentation from content; unfortunately however this is a far from trivial task.
"The inter-weaving of pure content and semantics are too deep. Douglass Hofstadter's article in Scientific American about Knuth's Meta-Font system is a great examination of how deep this problem really is... Hofstadter coined the term *Ambigram* to show how even simple typographic constructs can lead to interesting semantic ambiguities.

The field of mathematics has long struggled with this in its search for an executable representation of mathematical constructs. Some notations are so visual that it becomes difficult to see how there could ever be a useful separation made between content and interlinked presentation. For example consider Penrose's Tensor Diagram Notation (Penrose 2007) which represents extreme cases, where the presentation *is* the content. No wonder that TeX remains the weapon of choice for mathematicians.

And yet, the legal world manages to survive the ambiguities and contradictions in its corpus. How? Via what is known in semiotics as a dynamical interpretant known as the Judiciary. It is a beautifully simple idea. If there is a doubt as to the meaning of a text, the Judiciary tells you what it means. The explications provided are then

themselves captured in textual form known as case law and the case law becomes legally powerful thanks to stare decisis.

An analogy from software development is unit testing. The code is the code is the code but the true meaning of the code? The unit tests tell you that. The code "means" what the unit tests tell you it means. All else is just syntax. Case law is a bit like a unit test suite.

Is it possible to remove the tables completely from legal/para-legal documents? No the meaning is just too subtly inter-twined with the presentation. It is possible to remove the need for unit tests in software development? No, the meaning of source code is impossible to separate from its interpretant – an execution environment. A great way to see this is to look at static analysis tools and realize what it is about your code that static analysis can never tell you. Arguable the limitations of static textual analysis were established by the great Alan Turing back in 1936 with the halting problem.

So, if those tables cannot come out, what to do? I believe the most promising approach is to use the interpretant and stare decisis to remove as much ambiguity as possible. I.e. legally/socially binding exposition on what parts of tabular material contribute to meaning and what parts do not. That way, computer system designers like me would have guidance as to what needs to retained and what doesn't. Examples are things like fixed widths, tab leaders, vertical character alignments etc.

I honestly do not think it is possible to completely separate typesetting attributes into a nice binary "keep/optional" split but we won't know until we get the strare decisis process kicked off and let the interpretants in the Judiciary do their thing." McGrath closes in spectacular fashion, stating "I know of no jurisdiction that has attempted to grapple with this issue to date but it is becoming more and more pressing as the need for digital "authentic" legal materials grows and grows." (McGrath 2011)

4.2 The Crown Legislation Mark-up Language

In close proximity to the topics discussed in Section 4.1 is the CLML target syntax; this is the open data standard currently in implementation within legislation.gov.uk. Some documentation providing reasoning behind the use of this particular format for representing legislation states that "Public data will be published using open standards, and following relevant recommendations of the World Wide Web Consortium. Open, standardised formats are essential. However to increase reusability and the ability to compare data it also means openness and standardisation of the content as well as the format... The XML is published using an open XML Schema: the Crown Legislation Mark-up Language (CLML)." (The Stationary Office 2010) Although an in-depth dissection of the CLML and associated XSD's exceeds the scope of this research paper, it is important to describe the modular nature of the XSD make-up provides a very substantial underlying data model which is rich in metadata; built using various open standards such as Dublin Core, which have been specifically adapted to work with UK legislation. From the outset, it was our firm intention to *re-invent* as little of the wheel as possible with regards to the use of web standards during the mapping process. Our decision to base our methodology and use of the CLML standards around the proven production quality implementation used within legisaltion.gov.uk greatly enhanced not only the applicability of our research but also the resulting data mediation between the legacy XHTML STS data source and CLML destination.

4.3 Transubstantiative Mapping of Legacy Elements

Traditionally, when we refer to a data mapping problem we can be referring to a number of possible data integration tasks including ,but not limited to (i) data transformation between a source and target syntax, which usually includes the identification of data relationships as part of the lineage analysis (ii) data extraction and the discovery of hidden and possibly sensitive data which we can derive some business logic from (iii) data consolidation, usually comprising the amalgamation of multiple data sources which may reside in multiple distributed systems or databases into any one given data store, this would also involve some sort of data de-duplication to identify redundant data for elimination. In the above situation however we refer predominantly to (i). To date the most challenging aspect of the mapping process is the lack of metadata contentType elements within the XHTML source. Inversely, the CLML XSD's specify a rich and detailed metadata model, meaning that any XSL transformation is not so much of a mapping problem as a substantiating problem, where we find ourselves adding to the existing STS model substantially, this is reflected in the adequate naming of this subsection. Subsequently, it is important to note that this marks a large difference in the way we envisage an RDF/XML-based STS to be used. In discussions with various individuals working on legislation.gov.uk, we discovered that the metadata is key to the way in which the data models were being used, instead of attempting to infer information such as "When did sub-section 3.2.2 of the Building (Scotland) Act 2003 actually come into force?", we instead find verifiers asking questions more like "What U-value must a double glazed window have in an external conservatory which has a floor space larger than 15m²?" Both of these queries require a fundamentally different underlying graph data model for us to ascertain the correct information; the former meta-oriented and the latter substantially more domain and content specific.

4.4 Annotation of Terminology Semantics

The novel aspect which this research offers over the data model offered by legislation.gov.uk relates to the prevalent, interconnected inclusion of domain specific terminology semantics. These stem from two sources; (i) the ifcXML2x3 schema definition which itself consists of two parts; the common schema which annotates the STS with definitions for the header section and the general data types, part of ISO 10303-28 ed2, and the IFC2x3 specific unit of serialization, uos, which contains the XSD definitions of all IFC specific classes, relationships, attributes and data types, and (ii) the BSD defined terms including comprising comprehensive descriptions of regulations specific objects, their relationships and usually function. This helps to not only produce a fully annotated data model, creating new semantic relationships in the process, but also greatly reduces uncertainty caused by ambiguous terms, synonyms, antonyms and hyper/hyponyms which we so often see peppered throughout legislation of this nature. All mapping, schema validation and terminology annotation is done using custom embedded GATE processing resources running in XMLCalabash, an open-source Java-based XProc implementation.

5 PRACTICAL EXPERIENCES

To date we have found that consistency is a key factor here. Throughout our research we have made the utmost efforts to ensure that the methodology for producing open datasets from STS as described above upholds the notion that the meaning of the content within the underlying data semantics should not change even though we are

communicating it to users through different metrics. In the case of ensuring that the STS are being used effectively, we argue that verifiers should infer data from the underlying data model programmatically rather than visually when attempting to determine whether building warrant applications or practical construction outcomes are in compliance with the regulations and subsequently the Act, however this in turn raises the question of consistency and subsequently the integrity within the entire process of using these documents for decision based outcomes. Although it may seem obvious, it is important to recognise the key importance of both the way data is structured as well as within this context. In light of this, one way to think of our approach is to consider that we should focus on input driven outputs taking precedence over output driven inputs rather than the inverse which is what we witness in practice.

6 CONCLUSION

In this research we have documented a methodology detailing the initial stages of a semantic annotation process with the ultimate goal of opening up the information within Scottish Technical Standards as semantically rich open datasets which will make their usage much more transparent and meaningful. To date our experiences demonstrate that historically constructed legislative drafting workflows are not only extremely complex, with the resulting legislation shadowing this description, but also that improving this situation is a far from trivial task. Even though we were able to model our work on the best global implementation of open legislation currently available, the many intricate anomalies identified within the Scottish Building Standards System, which subsequently lead to "the regulation loophole", mean that there is still a huge amount of work to be done in this field. Future work involves focusing on the query model and expanding the linked data aspect of the data model to allow a furthering of regulatory compliance.

7 ACKNOWLEDGEMENTS

REFERENCES

Hoekstra, R. "The MetaLex Document Server", VOXPOPULII, Legal Information Institute, Cornell University Law School, 2012, accessed 02/02/2012 via http://blog.law.cornell.edu/voxpop/2011/10/25/the-metalex-document-server/

McGrath, S "Visual Tables and Meaning", Sean McGrath's Weblog, 2011 accessed 29/01/2012 via http://seanmcgrath.blogspot.com/2011/09/visual-tables-and-meaning.html

McGrath, S "XML in legislature/parliament environments", Sean McGrath's Weblog, 2010 accessed 29/01/2012 via http://seanmcgrath.blogspot.com/2010/05/xml-in-legislatureparliament.html

Optimal Economics, BRE, Liz Sgiel Associates "Research Project to Identify a Future Model for Reasonable Enquiry", Final Report, October 2011, accessed 23/01/2012 via http://www.scotland.gov.uk/Resource/Doc/217736/0122124.pdf

Penrose, R "The Road to Reality: A Complete Guide to the Law of the Universe", Vintage Books, 2007, ISBN: 0-679-77631-1

Sheridan, J. "Legisaltion.gov.uk", VOXPOPULII, Legal Information Institute, Cornell University Law School, 2010, accessed 02/02/2012 via http://blog.law.cornell.edu/voxpop/2010/08/15/legislationgovuk/

Statute Law Database "Building (Scotland) Act 2003", Statute Law Database, 2003, accessed 23/01/2012 via http://www.legislation.gov.uk/asp/2003/8/contents/data.xml

The National Archives "legislation.gov.uk", accessed 23/01/2012 via http://www.legislation.gov.uk/

The Scottish Government "The Scottish Building Standards Procedural Handbook 3rd Edition", Version 1.1, Building Standards Division, Directorate for the Built Environment, May 2011, accessed 23/01/2012 via http://www.scotland.gov.uk/Resource/Doc/217736/0116340.pdf

The Stationary Office "Opening up legislation for re-use", 2010, accessed 07/01/2012 via www.tso.co.uk/documents/LegislationBrochure.pdf

Working Group on Open Government Data "The Open Government Data Film: What is open government data? What is it good for?", Open Knowledge Foundation, 2012 accessed 23/01/2012 via http://opengovernmentdata.org/

W3C. (World Wide Web Consortium). *XProc: An XML Pipeline Language.* Available: http://www.w3.org/TR/xproc/. Last accessed 01/02/2012.