An intelligent authoring model for subsidiary legislation and regulatory instrument drafting within construction and engineering industry

McGibbney, Lewis; Kumar, Bimal

Published in:
Automation in Construction

DOI:
10.1016/j.autcon.2013.04.005

Publication date:
2013

Document Version
Peer reviewed version

Authors:

Lewis John McGibbney, Bimal Kumar

Contact Address:

School of Engineering and Built Environment, Glasgow Caledonian University, 70 Cowcaddens Road, Glasgow, Scotland, UK, G4OBA

E-mail addresses: lewis.mcgibbney@gcu.ac.uk (L. J. McGibbney), b.kumar@gcu.ac.uk (B. Kumar)

An Intelligent Authoring Model for Subsidiary Legislation and Regulatory Instrument Drafting within Construction and Engineering Industry

Abstract

Of primary importance within the domain of open data and more specifically open legislation, lies the essential central requirement for data to be available in a user oriented manner; whereby public and professionals alike can consume, share, reproduce it upon request and utilise it when time demands. Focusing specifically on subsidiary legislation (SL), current drafting workflows fall far short of addressing this vision. Whilst a significant amount of recent research has focused on less technical issues such as the actual definition of open data itself, the case for open data as to its management within a legal context, etc. due to the domain specific nature of producing robust and technically accurate open datasets, little work has been done on techniques for the drafting of legislative resources (in particular SL) as open data. To address this problem our work exercises a use-case driven from the domain of sustainable design and construction. As a validation vehicle, we select Scottish building regulations, which govern 32 local authorities across the country as a typical example of such legislation. Our work focusses on three areas of particular importance, (i) observing ongoing practice within the government organisations responsible for the drafting and publication of the aforementioned texts (ii) understanding the means and methods utilised by local authorities which have been tasked with ensuring that standards of compliance are met within all cases of design and construction, and (iii) reviewing and understanding how construction stakeholders actually execute their activities with respect to the texts in question. The outcome of our study has resulted in a methodology and subsequent production of an intelligent XML authoring workflow model (DROID-SL) for such documents which displays how legal texts of this nature can be better consumed within our society. We demonstrate that by adopting a user-oriented drafting vision, it is possible to produce high quality, user oriented, linked open datasets which wholly embrace the fast moving area of open legislation.

Keywords: Building Regulations, XML, Authoring, Crown Legislation Markup Language, Akoma Ntoso, Ontology, SPARQL

1 Introduction

The formal representation of UK SL including UK Statutory Instruments (UKSI's), Scottish Statutory Instruments (SSI's), Welsh Statutory Instruments (WSI's) and Northern Ireland Statutory Rules (NISR's) currently creates huge barriers for building professionals to properly and accurately engage with legal knowledge embedded within the construction related documents which fall within these legislatures[1]. This practice is not limited to the United Kingdom legislative system; other prominent examples include United States administrative law (Legal Information Institute, 2010), which can be further refined to the executive, legislative and judicial branches of the US Federal government (Breyer et al., 2011), where "It is Congress that grants general and specific powers to various Federal agencies through enabling legislation as well as the general laws for their fair and orderly administration. These executive powers are often quasi-legislative in nature (via rules and regulations applicable to a class of persons or organizations) or quasi-judicial in nature (via orders, adjudications and decisions involving particular persons or organizations)" (McKinney, 2010), and the legislature of the European Union; in which "One of the defining features of the community and, to a certain extent the Union, is the scope and level of power given to its institutions." (Steiner & Woods, 2009), where legislative power is disseminated among the Institutions of the European Union. In the latter example, an SL can be produced by either of the European Commission, Council and Parliament which includes acts and agreements by the legislature of the EU. One must, therefore, consider the unique nature of such legislatures, accompanied by the underlying inherited complexity of their historically developed working ecosystems when attempting a contextual study of their legislative outputs. Only then can one observe the dated production level, drafting work-flows, highly dependent on technologies which are ignorant (or unaware) of advances made in the domains of legal reasoning, knowledge and ontology representation (Sowa, 1999; Hoekstra, 2009), deontic notions, normative modalities, rights, factors, values and legal rule mark-up and argument making.

An overarching problem is that these administrative work-flows, proliferate society with huge amounts of unstructured data, subsequently creating systematic problems which filter themselves throughout many legal, ethical and social aspects of our modern states.

[1]Throughout this work, we refer to the term legislature as we intend to denote a parliament, assembly and/or elected body of people's representative's. Usually one would expect such an entity to comprise of a group(s) of person(s) with the ability to debate, edit, amend (or have amended) criminal or civil law within the state to which they act. Finally we refer to the relative output (legislative output) as legislation.

Unfortunately, efforts to mitigate against these problems have resulted in minimal measurable impact to date with the construction industry being no exception. Numerous studies (Salama & El-Gohary, 2011; Eastman, Lee, Jeong, & Lee, 2009) within the fields of applied computing and automation in construction have focused on improving compliance checking of building codes and regulations, with excellent research published on the topic of semi/fully-automating design and compliance checking (Garrett and Fenves 1986, Kumar 1989). This, however does not detract from the underlying debate that current administrative workflows are

- Inefficient; this is the result of the archaic nature of legislative drafting and the dated authoring procedures prevalent throughout legislative ecosystems.
- Highly inconsistent between parliaments and organisations responsible for dealing with the delegated tasks of producing legislation, and finally
- The primary cause of the fact that we, as humans, communicate with, infer from and use legislative documents which are represented in static forms differently every time we put them into application. Ironically, with this in mind, from a political perspective the parliamentary decisions (which include proposals, amendments, approvals, etc.) concerning ever growing quantities of legislation fail to acknowledge or consider the scale of problem the information overload phenomena has on the compliance process overall.

With such topics of international as well domestic importance e.g. Energy Conservation, Environmental Responsibility and Climate Change, we must address the requirement for a more evenly balanced enforcement model which acknowledges factors relating to compliance assurance instead of the existing narrowly directed, bureaucratically dictated model which is heavily biased towards obtaining (on many occasions) politically driven results. Subsequently there exists a requirement to look towards innovation, towards the use of emerging as well as established technologies from across the domains of Artificial Intelligence (AI), Informatics, the Semantic Web, etc. in an attempt to uncover possible solutions to rebalancing the enforcement of legislation as stated above. We have reached a stage within the research community, where a common consensus exists that the age of artificial intelligence within the domain of construction did not materialise into everything the initial hype promised as argued by (Tomiyama et al., 2012), “The dream in the early days was that it would eventually

become possible to develop a computer system of human level intelligence that would automate many of the tasks of engineers (in other words, completely replace human engineers). Unfortunately, this quickly turned out to be impossible, and the AI boom died before long. However, it also triggered diversified research efforts in the search for 'better paradigm AI techniques' to improve and increase information processing capabilities"

Our research focuses on improving professional and the public access to UK SL with a specific focus on construction and building regulations. Having witnessed first-hand, the current practice relating to the everyday use and application of such documents within several Scottish local authority building control departments, our vision to formulate a clearer and more comprehensive platform to enable efficient and consistent use of the SL for design checking and compliance control was initiated focusing not on systems "that facilitate the browsing and retrieval of regulations by industry practitioners" (Cheng et al., 2008), but instead on the formal representation of the underlying data model itself. Previously our work made a comparative study (McGibbney and Kumar, 2012) between two such computational data models e.g. the Crown Legislation Markup Language and Akoma Ntoso for the suitable representation of UK Construction Regulations.

The research documented herewith presents our motive and methodology for intelligent authoring and subsequent drafting of such SL. Section 2 provides a summated history of applied computing techniques within the legislative drafting domain introducing relevant concepts covering semantic web research and XML-based drafting workflows. Section 3 refines this narrative focusing specifically on SL drafting within the U.K context by highlighting key limitations in current drafting workflows. Section 4 then introduces DROID-SL our XML authoring framework which specifically builds on two aspects of our methodology for intelligent processing of SL data. At the programmatic, data modelling level we detail the following (i) the intricate discrepancies presented by the underlying Scottish Technical Standards (STS) (The Scottish Government, 2012) data model as presented by, and so commonly encountered within documents in the umbrella SL category (ii) our approach to addressing these particular characteristics during mapping to a target data model through the use of automated text transformation, named entity recognition and term annotation based on industry standard internationally recognised schema and dictionary definitions. Section 5 then concludes this paper.

2 History of Applied Computing in Legislative Drafting Workflows

Earlier research (Sartor et al., 2011) bestows to us the utopian contrivance that "In the emerging framework of the semantic web (where information can be directly processed by computer according to its meaning), legal documents and in particular legislative documents, are undergoing a fundamental change. Being directed to the Internet, rather than to a print house, such documents need to be identifiable in the web, structured according to document models and enriched with machine processable meta-data." The underlying vision of the semantic web is one with which we are now familiar as researchers from several domains have been working to move from rhetoric to reality in order to achieve this paradigm shift. In the run up to the millennium we witnessed the fusing of several research phenomena (Hoekstra, 2009) e.g. description logics, formal ontology and knowledge representation ontology, which were themselves products of technologies which shared their roots with the anagogic fields of Philosophy, AI and Cognitive Science, and which had evolved and matured from the early days of knowledge representation (Sowa, 1999) and expert systems such as DENDRAL (Lederberg, 1987; Lindsay et al., 1993), MYCIN (Shortliffe & Buchanan, 1975; Buchanan & Shortliffe, 1984), etc. which we embraced in the 1970's. This amalgamation of the former (millennium era) technologies mentioned above, directly resulted in what we now refer to as the semantic web. Subsequently, focused research into the development and use of semantic web technologies has resulted in widespread adoption of such technologies within many aspects of our societies, with increasing and positive commitment being shown by government administrations globally. If we observe arguments from a variety of domains like legal informatics (Sartor et al., 2011), ontology representation (Hoekstra, 2009) and computing and automation within construction and engineering (Cheng et al., 2008; Salama & El-Gohary, 2011), we see the consensus that, "This is achieved using standards based on XML (the eXtended Markup Language) to express document structures and insert in the documents meta-textual information. XML standards can be supplemented with ontology languages (for specifying conceptual structures), and rule languages (for capturing the logical content of legal rules)." (Sartor et al. 2011)

2.1 Review of relevant studies concerning semantic web and legislation

The first of a dedicated book series providing cutting edge rain commentary which falls on

the umbrella topics of artificial intelligence and law[2] openly forecasts that technologies relevant to the domain of law and IT include such topics as "(i) Legal Information Retrieval (LIR), (ii) Electronic Data Discovery (E-Discovery), (iii) Collaborative Tools (e.g. Online Dispute Resolution platforms), (iv) Metadata and XML Technologies (for Semantic Web Services), (v) Technologies in Courtrooms and Judicial Offices (E-Court), (vi) Technologies for Governments and Administrations (E-Government), (vii) Legal Multimedia, (viii) Legal Electronic Institutions (Multi-Agent Systems and Artificial Societies), (ix) The Socio-legal Web (Blawgs) and Web of Data (3.0)." (Casanovas et al., 2011) We take comfort from the authors' foresight when they indicated "that this fast development may be followed up and monitored by research focused on regulations, regulative devices and behavioural patterns" (Casanovas et al., 2011); the appurtenant fields of research only becoming admissible once the foundational subject areas within the territory somewhat mature. However, history teaches us that "difficulties and shortcoming of current choices become visible and require collective reflection" (Casanovas et al. 2011) in due course. Originating from a use-case driven approach, drawing parallels with (Casanovas et al., 2011), our research is "driven by the need to develop new computer applications...", (and in our case new data model(s)), "...to better meet the demand of practitioners and citizens, in a framework characterized by an accelerated technological development". Crucially, Hoekstra's MetaLex Document Server (Hoekstra, 2011) displayed work having a noticeable impact on the direction of our own research, enabling us to engage with foundational work, firmly established and widely quoted which presented in all its entirety a production quality implementation, the content and subsequent findings of which opened our access to a myriad of thoughts and innovation. Additionally we find it wholly relevant to mention Propylon's Legislative Workbench (Propylon Inc., 2011) and the Bungeni[3] suite of applications, which embrace the widespread use of legislative XML to support the running of legislatures. Until this stage, our own attempts (McGibbney & Kumar, 2011a; McGibbney & Kumar, 2011b) at creating xml-based representations of SL, specifically Scottish energy performance construction and building regulations, adopted an intelligible ontology engineering practice based on perception that "The choice of methodology behind ontology design is very much dependant on the nature

[2]Expanded sub-topics of absolute importance are driven overwhelmingly by "recent developments in semantic technologies, Natural Language Processing (NLP), ontologies, Information Retrieval technologies (IR) and the Web 2.0 and 3.0..."

[3] See http://www.bungeni.com

and characteristics of the targeted domain and its various applications, as well as the resources and development time available and the required depth of analysis of the ontology" (Rezgui, 2007). Sometimes it is not fully sufficient to manually work towards "establishing semantic relationships between the terms selected from legal sources" (Saravanan, 2009) as this in itself leads directly down the path to inconsistency.

2.2 Rationale behind the paradigm shift to XML-based drafting work-flows for subsidiary legislation

From an administrative, historically rendered point of view, it is little surprise that the argument for publishing and representing legislation as interoperable, open standards compliant, linked datasets, is initially not an appealing one. Additionally it should be noted that such an argument becomes even less appealing as it is neither a trivial nor generally approachable vision to achieve either. One aspect of utmost importance is that in any authoring process, regardless of the target domain, the content meaning contained within the source text must not be altered in any way, shape or form between source and intended target. That is to say it is absolutely imperative that at every stage in any transformation pipeline, the underlying semantics of legislative texts remain intact. To put this into context, one should consider the following argument where (McGrath, 2010) argues (on a similar topic) that the paradigm shift towards legislative XML (which in our case is a precursor to open, linked SL) "is not happy hunting ground for the blind application of standard IT architecture patters from the document management/content management/publishing space". McGrath continues on this thesis teasing us with uniform pro legislative-XML commentary "The first, often quite glaring architectural non-sequitur goes like this 1) legislatures/parliaments are full of very structured documents, bills, resolutions, journals, calendars, statutes, annotations… which all have readily apparent structure, 2) XML is all about handling very structured documents, 3) Therefore classic XML approaches fit legislatures/parliaments." McGrath then advances to present the twist in his tale in the most simplistically structured but paradoxically convoluted manner detailing barely enough for us to accede his viewpoint whilst in parallel craving a more comprehensive exploration of the underlying agonies presented by the legislative XML rationale. He continues "There are a variety of reasons why this analysis is, in my opinion, wrong[4]… a) the centrality of line/page number citation in amendment cycles, b) the complex

[4]It is always of critical importance that any of our citations are presented within the correct context, therefore we find it appropriate to substantiate on this particular citation stating that McGrath continues "Before I start, let me

nature of amendatory cycles, c) the critical nature of fidelity with historically produced renderings d) the fluid nature of work-in-progress legal assets e) the complexity of amendment cycle business rules that often pre-date computers and cannot be changed to make life easier for modern software f) the subtle inter-play between legal content and legal content renderings g) content aggregation and derived document types". We simply close this section by stating that it is imperative to take as many of these articulate facets into consideration before moving to XML-based SL drafting workflows. Of utmost importance is consistency. Consistency is key.

3 Current Snapshot of the U.K. Subsidiary Legislation Drafting Process

Hands-on drafting of documents within the legislative ecosystem is a tremendously difficult process to model within any legislature regardless of its location around the globe. Before being able to accurately appreciate the intricacies of such a complex model we are also required to acquire understanding and familiarity with the sub-processes of reasoning, judgement and decision making, which are all central components (albeit involved in activities which occur post drafting) within the legislative life-cycle. "Reasoning is concerned with making inferences, judgement is about the formation of beliefs about the likelihood of uncertain events, and decision making is about choosing between alternatives. These three aspects of cognition are overlapping, and interlinked: we may reason when attempting to form a judgement and our judgements may inform our decision making" (Hardmann & Macci, 2005). In stating this we temporarily explore a reversal of roles, questioning the traditional commissioning and publication of SL (which currently restricts our understanding and accurate application of such documentation within some given domain) by opening it up, and by representing it within a linked, user oriented data model to achieve better overall compliance measures. The argument (and accompanying methodology) we present stems from a feeling of frustration at the way SL texts are drafted, commissioned and published, of how we 'expect' users of these documents to merely be in compliance, taking little consideration for the compliance process involved both at a professional and end-user level. In our early work, we highlight the regulation loophole, where a significant gap exists between how those drafting such legislation *envisage* stakeholders and end users to actually

point out that XML *has* an enormous role to play in the legislatures/parliaments but it cannot be simply applied blindly per the standard XML value model without causing significant problems". This comment also represents our own particular viewpoint on this topic.

be complying with the relevant legislation, and how the end users *actually* are. In some cases we have found these two formalities to be quite literally worlds apart. Figure 1 displays an abstract perspective of the intricacies involved in a typical authoring workflow. Authoring (which usually combines efforts from numerous departments and many individuals) begins on the extreme left, after which data is stored and made available for human interpretation over the Web. Currently regulatory documents are served to end-users over HTTP and are made available in either PDF or HTML. Of critical importance is the publication and availability of static documents at the storage layer. The failure to create documentation in machine processable manners leads to inevitable confusion, represented by the question mark which lingers between various human-to-document communications within the Application/Access Layer.

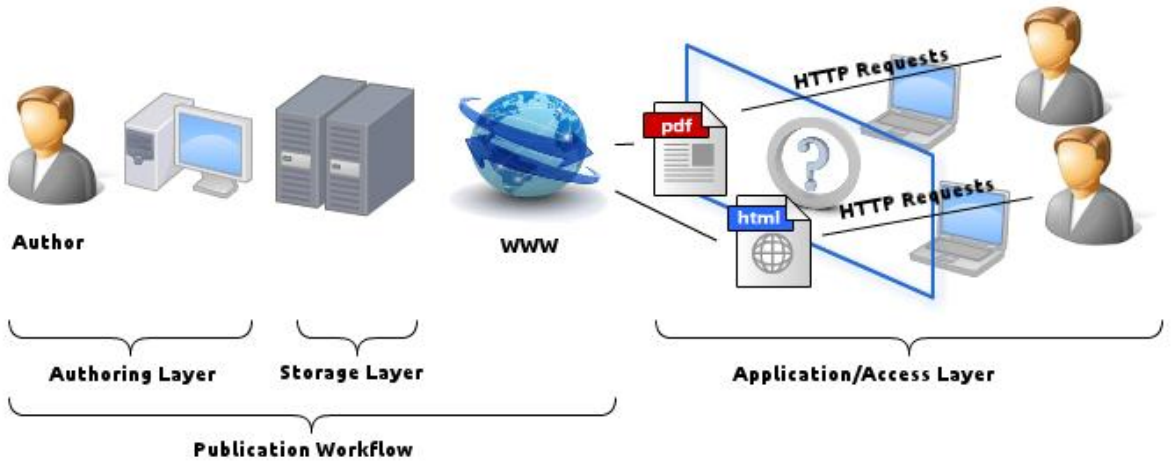


Figure 1 – Typical Authoring and subsequent Publishing workflow for Subsidiary Legislation

We envisage that a far greater understanding of legal texts could be achieved if they were communicated to their intended audience in a more appropriate manner. We therefore find it important to detail some aspects of typical authoring and publication workflows which as a result produce legislation, the nature of which we are concerned with. With a specific focus on SL, some workflows incorporate little or no level of provenance into the underlying legislative resource. This makes tasks relating to provenance tracking of document revisions through time and subsequent decision making a significantly more difficult task than it need be. This argument becomes more weighted if we consider that any artefact of documentation produced throughout an amendatory process is itself considered supplementary documentation within its own right. By nature such supplementary documents are often structured using the notion of line and page numbers (which is neither a machine friendly nor scalable option, obfuscating the picture further), rather than structured using domain specific

URI's (which uphold their identity even if the underlying semantic meaning changes, whilst in parallel withstanding the tests of time). We therefore finally progress to propose an authoring framework which accommodates the existing drafting processes of SL, which at its core, provides a supplementary means for publishing SL as open access, linked data. We consider this of primary importance to the overall value of our approach to the production of such data, as otherwise literally thousands of documents, from a myriad of domains persist to remain in static formats with little relative use.

4 DROID-SL - DRafting of Open Intelligent Data for Subsidiary Legislation

Throughout this work, the development of our argument for improved drafting of SL artifacts has been the direct result of a substantial degree of observation and understanding of the correlation between parties drafting and consuming such legislative materials. By this we refer directly to the entrusted (government) offices relayed with the specific delegation of power, whose duty concerns the assigned responsibility for the drafting of such legal texts, the numerous local authorities distributed throughout geographical boundaries who must ensure compliance is upheld in accordance with the most recent governing regulations, and finally the thousands of individuals on the other side of the fence tasked with ensuring their day-to-day work meets adequate levels of expected compliance. In this section, we focus solely on the design and implementation of an intelligent model for authoring and subsequent drafting of SL. Throughout our work we draw upon the STS as a vehicle not only for validation but for driving our interests and passion with regards to improving compliance checking within the domain of sustainable design and construction.

4.1 Design Methodology and Strategic Rationale

Although one is able to locate a specific document out of an entire corpus of closely related, interlinked web documents relating to a particular field/topic of interest, few systems exist to facilitate consistent inference of legislative resources. The common misguided dependence upon traditional text-based queries consistently fails users when they wish to ‘get inside’ complex documents which naturally contain a plethora of clauses. The primary diver behind minimal provision of explicit information can usually be attributed to a lack of structure embedded within such documents. Our method for addressing this problem therefore chose to make the conventional switch to a more strategically directed content oriented approach. In order to address the content search scenario, our data would first need to undertake a physical

mapping process whereby datasets were created directly from existing SL. This would empower individuals with the functionality of structured search, allowing the datasets to act as a hub for inference and reasoning, communication, with the vision of providing a platform for integrity-based enhanced decision making within the domain. We encapsulate the step-by-step execution of this physical mapping transformation in our DROID-SL authoring framework. The basis and extent of our methodology is therefore explained throughout the remainder of this section.

4.1.1 Intriguing Source Data Model Characteristics

One interesting and completely ironic feature of the STS data model which plays both a crucial role in our subsequent work and further discussion throughout this thesis was the mark-up of expressions which are denoted 'popular or defined terms'. By making such terms explicit both in the structure and meaning they could be utilised to act as a direct means of term disambiguation. In this context, by definition 'defined terms' were terms which referenced highly significant domain specific entities, each entity was then linked to an industry recognised description (definition) of what that entity actually meant within the context of the STS. An example would be 'Access deck', which means 'a structure having a surface in the open air suitable for ingress and egress of persons to a building'. Coincidently, this trivial example also makes reference to the entity 'building', which is itself, a defined term meaning 'any structure or erection, whether temporary or permanent, other than a structure or erection consisting of, or ancillary to:

a) any road (including any bridge on which the road is carried),

b) any sewer or water main which is, or is to be, vested in Scottish Water,

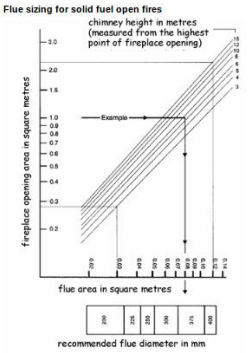
etc…

This, therefore, in effect creates two things, viz. (i) complete disambiguation of the term itself and (ii) a direct approach to embedding semantic meaning within the data model. Anyone with any interest and knowledge of linked data can surely see the benefits of such an approach to authoring artefacts of legislation regardless of the domain within which it governs. From our previous work, we know that term disambiguation does not only improve results of precision with regards to term queries, but also provides a platform to link in resources creating an interlinked web of knowledge as opposed to static SL documentation (iii) the previous two bullet points provide basis for direct, more focused interaction with the

legislative texts themselves. Although this final point shares significant overlapping interest with the body of this paper, it begins to move towards application layer specific discussion which we consider as meriting separate discussion in addition to this research[5]. This type of term disambiguation, weighting emphasis on technically important terms, means that people not only engage more with the data source but also gain a fundamental understanding of what certain entities mean in certain contexts within the text.

Other anomalies[6] peppered throughout the STS include numerous diagrams of varying natures, styles, relevance and complexity, mathematical equations (used for addressing the regulation of combustion appliances, energy performance calculations, etc.), large tables containing assortments of numeric and lexical data, headers, footers, mandatory sub-section standard summaries, etc.

	Main space heating system fuel [1] [2]				
Element or system	**Gas (Package 1)**	**LPG (Package 2)**	**Oil (Package 3)**	**Electricity (Package 4)**	**Biomass [3] (Package 5)**
Walls	U = 0.19	U = 0.19	U = 0.19	U = 0.19	U = 0.19
Floors	U = 0.15	U = 0.15	U = 0.15	U = 0.15	U = 0.15
Roofs	U = 0.13	U = 0.13	U = 0.13	U = 0.13	U = 0.13
Openings [4]	U = 1.5	U = 1.5	U = 1.5	U = 1.5	U = 1.5
Allowance for thermal bridging [5]	0.08 x total exposed surface area	0.08 x total exposed surface	0.08 x total exposed surface	0.08 x total exposed surface	0.08 x total exposed surface
Open *flues*	None	One	One	None	One
Heating system (pump in heated space)	Gas boiler *room sealed* fan *flued*, 90.2% efficiency	LPG boiler *room* sealed fan *flued*, 90.2% efficiency	Oil boiler *room* sealed fan *flued*, 93% efficiency	Air to water heat pump [6]	Wood pellet boiler HETAS approved



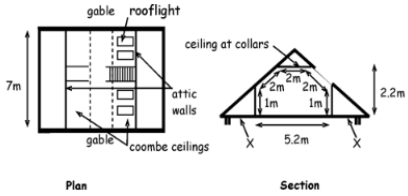


Figure 2 – Brief assortment of source data model characteristics. From left to right: Table including numeric and lexical data, graph data relating to fire regulation calculations for combustion appliances and finally plan and section elevation diagrams

4.1.2 Substantiation of Legacy Elements

Traditionally, when we refer to a data mapping problem we may 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 from which we can derive additional business logic, and (iii) data consolidation, usually comprising the amalgamation of multiple data sources which may

[5] More commentary is however focused on this topic in the concluding section.

[6] See Figure 2

reside in multiple distributed systems or databases into any one given data store, this could also involve some sort of data de-duplication to identify redundant data for elimination. We find it appropriate to coin the term *mapping* as our primary adjective of reference, as it captures the challenging process of traversing the gap between the limitations offered by the source model and the expressiveness of the target. Inversely within the scope of this work our desired target model(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[7], where we find ourselves bulking up the source STS HTML model to meet the requirements of the target model.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
     <title>Building Standards Domestic Handbook - General
</title>

     <meta http-equiv="Content-Type" content="text/html;
charset=UTF-8"/>
    <meta name="keywords" content="Building Standards Scotland,
Technical Handbooks, Domestic, 2010" />

<meta name="revised" content="2010-2-15T14:50:7+0"/>
     <!--[if lt IE 7]>
     <style type="text/css">div.page img.alpha{display:none}
</style>
     <![endif]-->
     <style type="text/css">
          html,body,div{margin:0;padding:0}
```

Figure 3: <head> element of the Scottish Technical Standards source HTML snippet

An example of this is the extensive use of the Dublin Core vocabulary (DCMI, 2012), where in the target mapping model we require several important object relations such as <dc:author, organisation, title, etc.>. In many cases the values for such attributes simply do not exist in the source markup. To further enable the reader to visualise and appreciate this problem, we graphically illustrate identical excerpts of content from the STS modelled in source and target representational formats within Figures 3 and 4 respectively; these provide insight into the stark differences between the models and the depth of content analysis and document structure required before accurate mappings can be achieved. Figure 3 displays a snippet from the source STS <head> HTML element. Generic useful data subsequently utilised within the mappings includes content from within the <title> tag, http-equiv, content and

[7]For clarity however we draw upon the widely recognised nature of the former, i.e. mapping.

charset attributes within the <meta> tag, the last revised date from within the second content <meta> attribute, etc. These fields map directly to the Dublin Core <dc:description>, <dc:content> and <dc:modified> elements respectively.

Figure 4: Crown Legislation Markup Language snippet output after phase 1 of mapping

Subsequently, it is extremely important for us to note that this marks a large discrepancy in the way we envisage such documents as the STS to be used when mapped to the more semantically rich target models. One will note that on a typical basis legislative metadata models are purposely suited to non-technical artefacts of legislation. In specific terms, it is fundamentally central to the way in which the data artefacts were being used, where it caters well for users attempting to infer information such as “When did sub-section 3.2.2 of the Building (Scotland) Act 2003 actually come into force?”, but goes no real length to accommodate typical queries we may find verifiers seeking to infer from the STS such as “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 data model for us to ascertain the correct information from. The former requiring a vaguer provenance related information source and the latter substantially more domain and content specific. Based upon this we decided to implement more focused content processing enabling us to incorporate additional levels of expressiveness in order to meet the domain specific query requirements the data would typically be subject to.

4.1.3 Driver behind Annotation of Terminology Semantics

Returning to the domain specific terminology noted earlier the purpose behind the inclusion

of interconnected domain specific terminology semantics was twofold; (i) to ease navigation between relevant articles or section of interest when the STS were viewed through a browser, and (ii) to disambiguate as well as define specific terms which bore significant weight either within one particular subsection, section or of course the entire handbook itself.

```
 <p><strong><a name="Uvalue" id="Uvalue"></a>U-value</strong> (or thermal transmittance co-efficient) is a
measure of how much heat will pass through one square metre of a structure when the temperatures on either
side of the structure differ by 1 degree Celsius (expressed in W/m2K).</p>

  <p><strong><a name="v" id="v"></a><a name="Ventilator" id="Ventilator"></a>Ventilator</strong> means a
window, rooflight, grille or similar <em><a href="#Building">building</a></em> component (and in the case of
a <em><a href="#Dwelling">dwelling</a></em> includes a door) capable of being opened to provide ventilation.
</p>

  <p><strong><a name="w" id="w"></a><a name="Wastewater" id="Wastewater"></a>Wastewater</strong> means water
that is contaminated by use and normally discharged from a watercloset, shower, bath, bidet, washbasin, sink,
washing machine, floor gully and similar facility and also includes rainwater when discharging in a <em><a
href="#Wastewater">wastewater</a></em> drainage system. </p>
```

Figure 5: Snippet of example domain specific ‘defined terms’ occurring throughout the Scottish Technical Standards

With regards to ease of navigation, it is correct for us to reaffirm that in producing the output datasets, within the remit of our work, we are not remotely interested in how the data appears when viewed through a browser.. From our previous work we recognised the importance of having a standardised approach of identifying crucially important terms and keywords from within a document corpus. Once an initial identification process like this has been undertaken, we can begin to utilise tools to annotate every occurrence of such terms with their corresponding definition, hence embedding an additional layer of semantic understanding within the dataset. Secondly we utilised the ifcXML2x3[8] schema definition; the primary, globally recognised vocabulary and data exchange format relating to the communication and exchange of construction and engineering information. Again for clarity typical snippets of these resources are detailed in Figures 5 and 6 accordingly. Figure 5 shows three HTML paragraph elements containing name and id attributes with values of the corresponding defined term. These terms are populated throughout the transformed regulations whenever the term is matched. A similar principle is adopted to populate the terms encountered within the IfcXML XSD’s as displayed in Figure 6.

[8] See http://www.buildingsmart-tech.org/specifications/ifcxml-releases/ifcxml2x3-release/

```
<xs:element substitutionGroup="ifc:IfcBuildingElement" name="IfcWall" nillable="true" type="ifc:IfcWall">
</xs:element>
<xs:complexType name="IfcWall">
        <xs:complexContent>
                <xs:extension base="ifc:IfcBuildingElement">
                </xs:extension>
        </xs:complexContent>
</xs:complexType>
```

Figure 6: Excerpt from ifcXML2x3 XSD

The actual annotation of domain specific terminology was achieved using GATE[9] (General Architecture for Text Engineering), a java-based text engineering framework including a suite of libraries for natural language processing (NLP), named entity recognition (NER), and other linguistic engineering tasks. This comprised of custom processing resources (PR), which combined NER with look-up mechanisms ensuing match and annotate sequences run from within a PR pipeline. This level of processing results in an enriched dataset fully annotated with each occurrence of a defined term as specified within either the STS defined terms, or the ifcXML XSD's.

4.2 The Intelligent Authoring Model

The DROID-SL authoring model (see Fig. 7) can be thought of as a pipeline where a sequence of synchronized operations is executed on XML documents in order to achieve some desired target outcome. The pipeline implementation is managed using XML Calabash, an open-source pure Java implementation of XProc[10] (XML Pipeline Language). Within the pipeline, the custom GATE PR and XSD validation steps are executed, the latter being achieved using a commercial version of Saxon, a pure Java XSLT processor. The first step in the model is to simulate a typical revision to an artifact of legislation contained within a database. For the purpose of this research, we have been using the popular MarkLogic Server as the primary storage tier for the resulting XML, reasoning behind this is two-fold:

- After the final XSD validation step, we can simply call the relevant insert document procedure from within XML Calabash. This allows us to directly send the fully annotated XML output directly to the storage tier, enabling us to execute queries against document updates in near real-time
- Generally speaking MarkLogic Server appears to have been very well accepted into

[9]See http://gate.ac.uk/

[10]See http://www.w3.org/TR/xproc/

the XProc and XML communities. The XML Calabash project also has excellent support with many convenience procedures specifically for MarkLogic server.

Throughout the remainder of this section we now elaborate on the specific component parts of the model as briefly described above.

4.2.1 Primary Propagation of Defined Terms

Immediately following the initial mapping of legacy elements to target syntax, (and first of three XSD validation intermediates[11]), comes the execution of a simple ANNIE[12] list lookup gazetteer, the source of which is provided as one of the core plugins for the GATE framework. The principal functionality provided by a gazetteer of any nature is to identify entity names in some given text based on input lists. Within GATE, the data within one of these lists is represented in plain text, with one entry per line (similar to comma separated value data). The primary list example provided by GATE concerns a set of names, such as the names of cities. For our specific requirement, each line contains a defined term from within the vocabulary as set out in the STS, to which a gazetteerFeatureSeparator (in our case this is set to a semi-colon ';') is directly appended.

To accompany each term we provide its corresponding definition. The functionality executed by the gazetteer consists of identifying the presence of each term in the list from within the input text, every time the defined term is encountered within the text (or as a series of occurrences scattered throughout the document corpus), it is then automatically annotated with a *Look-up* term definition tag which contains the definition text itself. Figure 8 displays subsequent text output after primary term propagation, everything excluding the black text is the result of primary term propagation. In this instance three occurrences of two terms have been discovered namely *building* and *constructed* respectively.

[11]based on the assumption that the relative output XML conforms with the target XSD

[12]The acronym given to a cluster of processing resources by GATE developers and stands for A Nearly New Information Extraction system

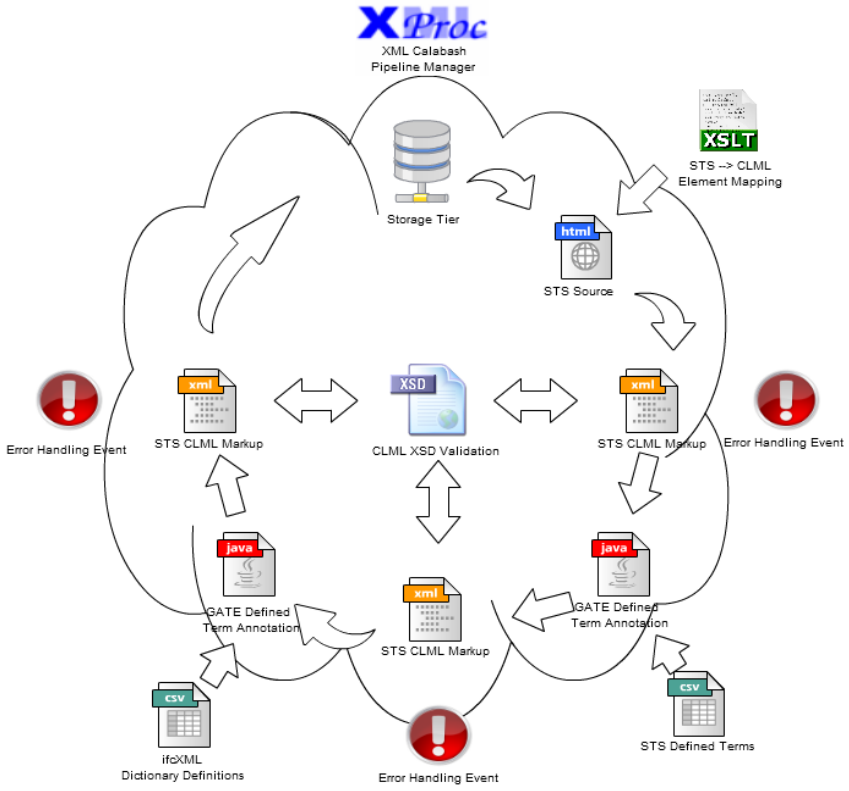


Figure 7 – The Intelligent Authoring Model: A custom XProc XML pipeline, incorporating XSLT, staged XSD validation, error handling and embedded GATE terminology annotation.

4.2.2 Schema Validation and Error Handling

In addition to the main steps we discuss above, one will notice the inclusion of error handling support within the XProc pipeline. In all, error handling is executed after every schema validation iteration enabling us to recognise errors within the pipeline before a full cycle completes. We found this incremental approach to ensuring the resulting XML complies fully with the XSD's provided better output results[13], especially when we again revisit our earlier comments on consistency within the mapping process as a whole.

[13] It also saved a significant amount of time during testing and development

```
<Text>Every <definition tag="means any structure or erection, whether temporary or permanent, other
   than a structure or erection consisting of, or ancillary to: a) any road (including any bridge
   on which the road is carried), b) any road, c) any sewer or water main which is, or is to be,
   vested in Scottish Water, d) any aerodrome runway, e) any railway line, f) any large raised
   reservoir within the meaning of the Reservoirs Act 1975 (c23), g) any wires and cables, their
   supports above ground and other apparatus used for telephonic or telegraphic communication. Any
   references to a building include references to a prospective building. Any references to a
   building, structure or erection include references to a part of the building, structure or
   erection. In relation to the extension, alteration or conversion of a building, references to a
   building are to so much of the building as is comprised in the extension or the subject of the
   alteration or conversion">building</definition> must be designed and <definition tag="includes,
   alter, erect and extend, and construction and related expressions are to be construed accordingly"
   >constructed</definition> in such a way that any component part of each fixed combustion appliance
   installation will not cause damage to the <definition tag="means any structure or erection,
   whether temporary or permanent, other than a structure or erection consisting of, or ancillary to:
   a) any road (including any bridge on which the road is carried), b) any road, c) any sewer or
   water main which is, or is to be, vested in Scottish Water, d) any aerodrome runway, e) any
   railway line, f) any large raised reservoir within the meaning of the Reservoirs Act 1975 (c23),
   g) any wires and cables, their supports above ground and other apparatus used for telephonic or
   telegraphic communication. Any references to a building include references to a prospective
   building. Any references to a building, structure or erection include references to a part of the
   building, structure or erection. In relation to the extension, alteration or conversion of a
   building, references to a building are to so much of the building as is comprised in the
   extension or the subject of the alteration or conversion">building</definition> in which it is
   installed by radiated, convected or conducted heat or from hot embers expelled from the appliance.
</Text>
```

Figure 8: Resulting ANNIE entity recognition and term annotation

4.2.3 Secondary Propagation

Not only do the ifcXML dictionary schemas describe components peppered throughout technical documents such as the STS, but they also exist as open data which greatly improves interoperability between client applications which use them. The ifcXML XSD's comprise of two parts; the common schema which annotates the STS with definitions for the header section (metadata) and the general data types, part of ISO 10303-28 ed2, and the IFC2x3 specific unit of serialization, which contains the XSD definitions of all IFC specific classes, relationships, attributes and data types. As was stated previously, our intention was to reuse as many existing standards within the target mapping as possible, this would improve the provenance of the resulting dataset and would also allow us to adopt and monitor a more granular governance strategy as well. Due to the native representational format for the ifcXML XSD's, it became apparent that a simple gazetteer would not suffice to deliver the functionality we require in order to obtain similar results as those achieved above. One will immediately deduce the differences between the snippets provided in Figures 5 and 6 respectively, with the latter encompassing the more complex data source. In addition to the obvious nested structure of the additional *Type* data (present throughout the entire ifcXML schema specification), one will also see that it is also permissible for any element to have a number of attributes. This therefore poses two problems when we consider implementing the formal annotation of such terminology throughout any text; (i) due to the relationship-based

properties ifcXML terms share with their relative classes, data types, etc. the definitions we are interested in quickly become cumbersome, therefore the population of terms and annotations throughout gazetteer lists does not scale well (ii) leading directly from this point we therefore assume that a more expressive annotation algorithm is required to successfully achieve adequate term annotations other than a simple look up list. We are therefore required to implement a more technically complex gazetteer processing resource which allows us to load large vocabularies and which can also operate at scale in parallel with the size of the ifcXML data as it grows. We utilise some of the functionality from a customised Large Knowledge Base (LKB) gazetteer in order to satisfy these requirements. The LKB Gazetteer works on the principle that expressive vocabularies are represented as RDF, either locally or persisted in an RDF database(s). These vocabularies are loaded into the gazetteer to obtain lookup annotations which have two annotation features, namely *instance*[14] and *class*[15] URI's, which can then be used within the annotation process. A far greater description of the type of URIs we design and refer to can be sought in guidance focused specifically on the design of URI sets for the UK Public Sector (UK Government, 2009). Unfortunately, however we have not automated the ontology construction process therefore we construct and persist the RDF locally. Some additional configuration required to ensure the LKB gazetteer operated correctly entailed the configuration of a stored SPARQL[16] query[17], used to obtain a subset (specific term) from within the RDF ontology based on term matching within the legislative text. Upon a particular successful term matching instance, the term is populated into the text similar to the ANNIE gazetteer introduced in the first annotation phase.

4.2.4 Data Persistence and Storage

This section merely acts as a beginning point for our future work; focused predominantly on a suitable query model for the resulting XML which is persisted to a data store of choice. As briefly described in the opening text of this subsection, for reasons of convenience we chose to utilise MarkLogic Server; in part due to the generous academic license offerings but predominantly due to the close connections and apparent unofficial adoption of this product

[14]Which contains the URI of the ontology instance

[15]Containing the URI of the ontology class that the ontology annotation instance belongs to.

[16]See http://www.w3.org/TR/rdf-sparql-query/

[17] A sample query can be seen in Figure 9

from within the XProc community. Without going into the specifics regarding the MarkLogic data model, functionality offerings and benefits within the big data marketplace, we will mention that the relative support and functional suitability was very well suited as an apt storage tier for XML pipelines of this nature.

```
PREFIX  dc:   <http://purl.org/dc/elements/1.1/>
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>
PREFIX  owl:  <http://www.w3.org/2002/07/owl#>
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  IFC2X3_subset: <http://www.buildingsmart-tech.org/ifcXML/IFC2x3/FINAL/IFC2X3_subset.owl#>

DESCRIBE ?x
WHERE
  { ?x IFC2X3_subset:hasName IFC2X3_subset:IfcBeam }
```

Figure 9: A simple SPARQL DESCRIBE query to obtain any individual(s) from the RDF whose name is IfcBeam

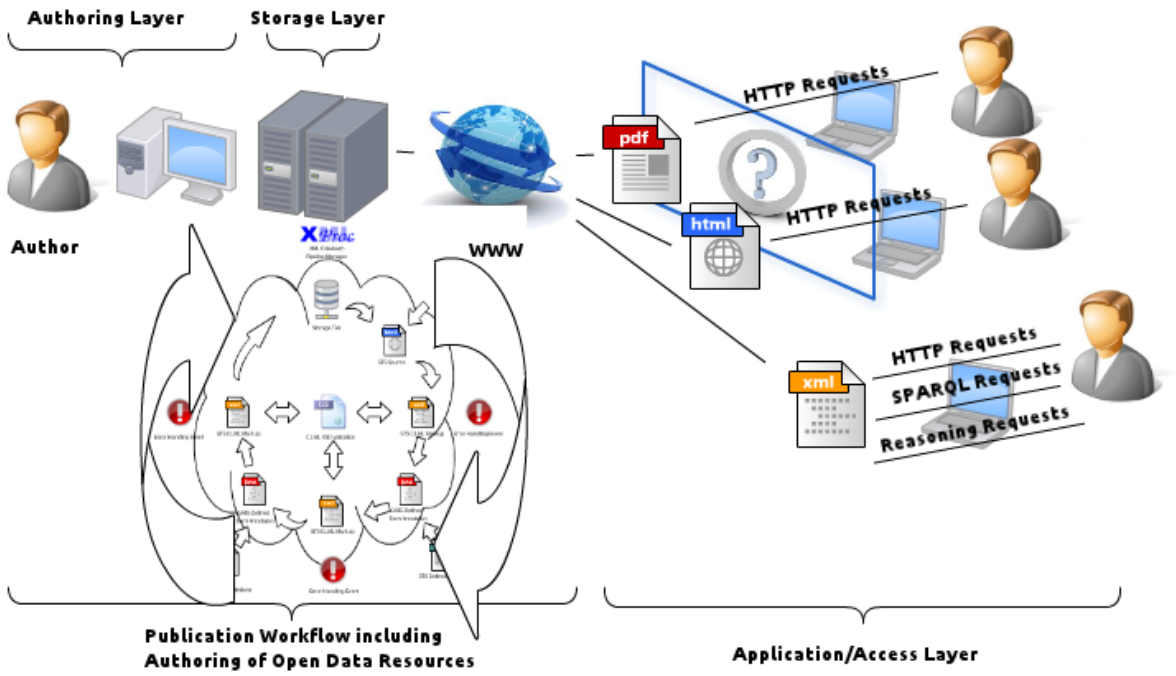


Figure 10 – Lifecycle Authoring and subsequent Publishing workflow for Subsidiary Legislation including the DROID-SL open data workflow

5. Conclusions

During the course of this work we have demonstrated an applied approach to producing open legislation without affecting existing authoring and publication workflows for such texts. Using subsidiary legislation within the domain of building, design and construction as our target field of focus we display that it is possible to embrace the fast moving domain of open data by producing high quality, consistent subsidiary legislative data in a user oriented manner; whereby public and professionals alike can consume, share, reproduce it upon request and utilise it when time demands. As it currently stands we are unaware of similar efforts in production elsewhere, especially within our target domain, therefore we see a high potential for the applicability of our work within other domains reliant upon, and regulated by dynamically changing subsidiary legislation, regulatory guidance and compliance documentation.

By adopting a user-oriented approach to addressing the scenario, our methodological rationale fundamentally shifted from traditional document search to content-specific data modelling. The direction change is evident within the decision to branch from the less practical issues surrounding the production of linked data instead opting for a more technically oriented analysis of textual content. In doing so, we have uncovered a myriad of modelling pitfalls frequently encountered when considering the production of open data from SL enabling us to adapt our workflows to suit such occurrences. For completeness we include a final graphic in Figure 10 displaying an abstract overview of how the DROID-SL framework would accompany existing drafting workflows. Neither data publishers nor data consumers need to be aware of the supplementary drafting framework being executed *under the hood* in order to produce corresponding updates to legislation as open linked data. As well as the actual framework, Figure 10 uncovers alternative document interactions, namely that data users now have the means to submit structured queries, reason and infer data which they would have been unable to do previously.

The DROID-SL framework has at its core a firm belief that by embracing a paradigm shift towards supplementary intelligent drafting it is possible to produce high quality, user oriented, linked open datasets which wholly embrace the fast moving area of open legislation. This research establishes a basis for improving the representation (and subsequent inference of) legislative documentation with the long term aim of further facilitating the use of open linked legislation within society.

References

Breyer, S. G., Stewart, R. B., Sunstein, C. R., & Spitzer, M. L. (2011). *Administrative Law and Regulsatory Policy: Problems, Text, and Cases* (Vol. 7Ed). Portland, US: Portland Book News, Inc.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* (Vols. ISBN 978-0-201-10172-0.). Reading, MA: Addison-Wesley.

Casanovas, P., Sartor, G., Biasiotti, M. A., & Barrera, M. F. ((2011). Introduction. Theories and Methodology in Legal Ontology Engineering. In G. Sartor, P. Casanovas, M. A. Biasiotti, & M. F. Barerra, *Law Governance and Technology Series. 1.* Springer.

Cheng, C. P., Lau, G. T., Law, K. H., Pan, J., & Jones, A. (2008). Regulation Retrieval Using Industry Specific Taxonomies. *Artificial Intelligence in Law, 16*(3), p277-303.

DMCI. (2012). *Dublin Core Metadata Initiative.* Retrieved 10 29, 2012, from http://dublincore.org/

Eastman, C., Lee, J. M., Jeong, Y. S., & Lee, J. K. (2009). Automatic rule-based checking of building designs. *Automation in Construction, 18*(8), 1011-1033.

Garrett, J. H., Jr., Fenves, S. J. (1986) A Knowledge-based Standards Processor for Structural Component Design, Report No. R-86-157, Department of Civil engineering, Carnegie-Mellon University, Pittsburgh, U.S.A.

Hardmann , D., & Macci , L. (2005). Introduction. In *Thinking: Psychological perspectives on reasoning, judgement and decision making* (Vols. ISBN 10: 0-471-49457-7). England: John Wiley & Sons Ltd.

Hoekstra, R. (2009). *Ontology Repsresentation: Design Patterns and Ontologies that Make Sense.* IOS Press.

Hoekstra, R. (2011). The MetaLex Document Server: Legal Documents as Versioned Linked Data. *Lecture Notes in Computer Science,* (pp. 128-143).

Kumar, B. (1989) Knowledge Processing for Structural Design, Ph. D Thesis, Department of Civil Engineering and Building Science, Edinburgh, UK.

Lederberg, J. (1987, 11 05). How DENDRAL WAs Conceived and Born.

Legal Information Institute. (2010). *Administrative Law*. Retrieved 03 18, 2012, from http:www.law.cornell.edu/wex/Administrative_law

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1993). DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence, 61*(2), 209-261.

McGibbney, L. J., & Kumar, B. (2011a). A Knowledge-Directed Information Retrieval and Management Framework for Energy Performance Building Regulations. *Proceedings of the 2011 ASCE International Workshop on Computing in Civil Engineering*, (pp. 339-346).

McGibbney, L. J., & Kumar, B. (2011b). The WOMBRA Project: A Web-based Ontology-enhanced, Multi-purpose, Building Regulation, Retrival Application for Scottish Technical Standards. *Proceedings of the CIB W78-W102 2011: International Conference.* Sophia Antipolis.

McGibbney, L. J., & Kumar, B. (2012). A Comparative Study to Determine a Suitable Representational Data Model for UK Building Regulations, submitted to International Journal of Data and Knowledge Engineering, Elsevier.

McGrath, S. (2010, 05). *XML in legislature/parliament environments*. Retrieved from Sean McGrath's Weblog: http://seanmcgrath.blogspot.co.uk/2010/05/xml-in-legislatureparliament.html

McGrath, S. (2010). XML in legislature/parliament environments : The centrality of line/page number citation in amendment cycles . Retrieved from http://seanmcgrath.blogspot.co.uk/2010/05/xml-in-legislatureparliament_30.html

McKinney, R. J. (2010). Presentation to LLSDC Librarians. *Federal Administrative Law: A Brief Overview*. Federal Reserve Board. Retrieved 03 18, 2012, from LLSDC Librarians: http://www.llsdc.org/attachments/wysiwyg/544/Fed-Admin-Law.pdf

Salama, D. M., & El-Gohary, N. M. (2011). Semantic Modelling for Automated Compliance Checking. *2011 ASCE International Workshop on Computing in Civil Engineering.* Miami: ASCE.

Saravanan, M., Ravindran, B., & Raman, S. (2009). Improving legal information retrival using an ontological framework. *Artificial Intelligence in Law, 17*, 101-124.

Sartor, G., Palmirani, M., Francesconi, E., & Biasiotti, M. A. (2011). *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management* (Law Governance and Technology Series ed., Vol. 4). Springer.

Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine . *Mathematical Biosciences, 23*(3-4), 351-379.

Sowa, J. F. (1999). *Knowledge Representation: Logic, Philosophical, and Computational Foundations.* Pacifica Grove, CA: Brooks Cole Publishing.

Steiner, J., & Woods, L. (2009). *EU Law* (Vol. 10th ed). New York, US: Oxford University Press.

The Scottish Government. (2012). *Technical Guidance.* Retrieved 04 12, 2012, from The Scottish Government: http://www.scotland.gov.uk/Topics/Built-Environment/Building/Building-standards/publications/pubtech/thb2011octdom

Tomiyama, T., Smith, I. F., Chen, C. H., & O'Brien, W. O. (2012). Editorial. *Advanced Engineering Informatics, 26*(1), p1-2.

UK Government. (2009). Designing URI Sets for the UK Public Sector. *1.0*. Retrieved from http://www.cabinetoffice.gov.uk/sites/default/files/resources/designing-URI-sets-uk-public-sector.pdf