www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

A COMPARATIVE STUDY TO DETERMINE A SUITABLE REPRESENTATIONAL DATA MODEL FOR UK BUILDING REGULATIONS

PUBLISHED: February 2013 at http://www.itcon.org/2013/2
EDITOR: Amor R.

Lewis J McGibbney, PhD Researcher
School of Engineering and Built Environment, Glasgow Caledonian University, Scotland, UK
lewis.mcgibbney@gcu.ac.uk

Bimal Kumar, Professor
School of Engineering and Built Environment, Glasgow Caledonian University, Scotland, UK
b.kumar@gcu.ac.uk

SUMMARY: *The notion of advancing levels of adequate regulatory compliance within the domain of construction and engineering is by no means a new phenomenon. An extensive degree of effort has over a number of decades previously focused on semi/fully automating design checking against regulatory building codes. After experiencing somewhat of a dip in popularity within academia, this research topic is once again of great interest due to the on-going adoption of Building Information Modelling (BIM) within the construction domain among other drivers like frequent updates to building regulatory documents particularly in relation to sustainability related issues like energy efficiency of buildings. This interest will only increase as we witness further widespread adoption of this new BIM paradigm across until now untouched disciplines. Unfortunately design checking within the remit of BIM still faces major problems with respect to national and local regulations as no sustainably sensible or scalable method has yet been implemented to ensure compliance rules are consistent with current legislation within these contexts. This work questions a suitable representational data format for UK construction and engineering regulations focusing on machine processable formats built entirely on open data standards which can then set the basis for addressing the aforementioned issues. We compare two existing legislation data models, namely the Crown Legislation Markup Language (CLML) and Akoma Ntoso, whilst in the process grading each on its suitability for most accurately accommodating and expressing the domain and profession specific nature of typical building regulations. The study conclusions indicate that the design, management and current availability of the CLML has resulted in a scenario where barriers exist to widespread adoption, this includes building out community in support of the language. Akoma Ntoso on the other hand has focused on building community around the proposed standard. This increases the likelihood of phased organisational transition towards the standard for achieving improved representational data modelling.*

KEYWORDS*: Open Data Standards, Building Regulations, Open Government, Data Model, Akoma Ntoso*

REFERENCE: *Lewis J McGibbney, Bimal Kumar (2013) A comperative study to determine a suitable representional data model for UK building relations, Journal of Information Technology in Construction (ITcon), Vol. 18, pg. 20-39, http://www.itcon.org/2013/2*

1. INTRODUCTION

Within the remit of subsidiary, supplementary or delegated legislation[1] there exist very few efforts to make the information more openly accessible to the citizen and more processable to the average machine. Currently we are witnessing a fast growing number of movements[2] from across the globe engaged in active monitoring of such topics as parliaments, public finance, transport, environment and political accountability, etc. However, very little has been done to engage within and monitor subsidiary legislation with a specific focus on actual design checking of construction work from a consumer's point of view. Although there is now an overwhelming argument backing the publication of legislation as open data, organisations which have the delegated responsibility of producing adequate data within the facet of subsidiary legislation still seem to be neglecting this task entirely. Unfortunately, such artifacts of subsidiary legislative data within the construction sector such as building regulations seem to be on the whole ignored. Recent documents such as the declaration of parliamentary openness (OpeningParliament.org, 2012) set clear guidelines for not only governments but also for authorities and organisations wishing to contribute data into the public aspects of our society to promote, educate, make available and publish data in such a way as to make it easily accessible for citizens and business alike. As an example, although the separate regional Governments within the United Kingdom are aware of the requests for open data as a direct consequence of requirements for increased transparency, existing legislative drafting workflows which affect legislative and regulatory data of this kind fail to acknowledge the movement to publish data in machine processable, standardised formats. In particular, the drafting workflows concerning the production and subsequent publishing of regulatory material within the UK suffers from the failure of central government to acknowledge the highly fragmented nature of the construction industry and the consequences this has on the authoring of such guidance. Over a number of years we have observed the issue of the so called *compliance gap* edging its way into government commissioned reports (Optimal Economics Ltd *et al.* 2011a), (Pye Tait Consulting, 2012) and (Optimal Economics Ltd, 2011b) focused on construction regulations in an attempt eventually to improve the mechanisms for checking of adequate and sustainable development of our built environment. Although such reports make clear acknowledgement that the current situation needs to be improved, there is a huge amount of work to be done to accurately understand the best strategic methods required to achieve this utopian vision. Within this work we pursue the argument that the production of open legislative data will make significant progress in meeting these targets within the construction domain. If one were to briefly consider the authoritarian method by which the aforementioned legal documents are currently authored, amended, published and disseminated, one will quickly discover that the very principles founding the entire work-flow are based on a series of dated, historically rendered events which entirely favour those at the authoring end of the spectrum. This is in opposition to a more mutually weighted model designed to satisfy

[1] We classify these documents together and refer to the names interchangeably throughout this work. Their defined meaning within the context of this work covers the permission for governments to make changes to the law using powers conferred by an Act of Parliament. In the UK the overwhelming majority of delegated legislation falling within this legislative facet is formally referred to as a statutory instrument.

[2] Examples include geographically dispersed groups such as Parliamentary Monitoring Organisations, Academia, Hacktavists, and Political Scientists etc.

stakeholders such as regulatory building control officers tasked with ensuring reasonable levels of enquiry are achieved with regards to the legislation and of course the citizen end users who regularly engage with an array of changing legislative data in an attempt to meet ever changing levels of regulatory compliance. This is in stark contrast to specific government guidelines on the publication of such data (The Scottish Government, 2012) which (as a an example) is not alone in creating optimistic and loosely enforced directives covering this topic.

Very recently a set of guidelines established within the Declaration of Parliamentary Openness (OpeningParliament.org, 2012) aims to establish a basis upon which we can begin to bridge the gap mentioned above in a structured, realistic manner. The guidelines aim to develop the way in which government organisations publish data into the public domain by making it more sustainable and mutually competitive meaning that not only (for example) legal experts are able to obtain and decipher various government artifacts. Within the remit of this paper we discuss some of the sets of issues covered within the Declaration on Parliamentary Openness with a specific application to artifacts of UK subsidiary legislation such as building regulations. This paper reviews two available data models, and considers issues involving transparency of information, accessibility, potential for citizen participation, sharing of good practice, the provision of the models to accommodate open structured formats for data representation and the mitigation of technological barriers for adoption.

2. EXPERIENCES OF THE DOMAIN SPECIFIC NATURE OF UK SUBSIDIARY LEGISLATION

A dominant aspect of this work concerns the underlying irregularities and anomalies presented to us within Subsidiary Legislation. It is clear that for domain specific, highly specialized tasks such as the checking of building design the mere provision of a means to execute domain specific text based search over a document corpus (the web or local intranet) is not sufficient to ensure consistently accurate levels of compliance in agreement with the relevant regulatory material (McGibbney & Kumar, 2011a). During the early stages of our work, we concluded that although obtaining a thorough understanding of both focused web crawling and index-based search architectures were valuable to the progression of our methodology, we learned that our error was within naively considering them as primary areas of focus. We do acknowledge that the aforementioned methods of obtaining very specific information provided significantly enhanced search accuracy in terms of search precision over commercially motivated, generic search engines; however obvious gaps remained, particularly with respect to the lack of any structured query functionality (McGibbney & Kumar, 2011b). To elaborate, although one was 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, one was still unable to infer information from the document in a consistent and structured manner. Besides, the dependence upon traditional text-based queries proved the failure to accurately ‘get inside’ clauses within documents in a replicable manner, resulting in a situation where we found ourselves addressing a problem of significantly minor concern[3], which in turn

[3] We classify these problems within the category of minor importance as our perceived overall impression of the fundamental underlying problem of legislative data existing and being persisted in inefficient, static

augments the major problem further by providing an even stronger case for a more direct solution. Subsequent reworking of our problem brief, therefore, chose to make a direct switch from addressing the typically traditional problem of document search to a more strategically oriented content search approach. In order to address the content search scenario, our data would first need to undertake a physical mapping process whereby structured datasets were created directly from existing subsidiary legislation. This would empower individuals with the functionality of structured search, allowing the datasets to act as a hub for inference, reasoning and communication with the vision of providing a platform for integrity-based enhanced decision making within the domain. Key to progressing with this vision was for us to understand firstly what levels of granularity and expressiveness are required respectively within such a data model from an end user perspective before reviewing existing models which embrace the facilitation of a two way communication work-flow whilst adhering to the use of non-proprietary representation formats improving and encouraging levels of reuse.

2.1 Pre-requisites for a Suitable Data Model for Building Regulations within the UK

Until 2010, the Scottish Government published the Scottish Building Regulations online in both PDF and HTML formats. Upon initial inspection one would expect raw HTML to be a far superior point of reference to begin from over its plain text cousin. However, as is so often the case, the inherent difficulty when starting with HTML is that it is designed for presentation and written to display data, the underlying focus being on how it looks through a browser rather than how the data represented within it is physically structured or what the data actually is or means. One must therefore begin with a meticulous process of content analysis subsequently undertaken across the ~830[4] documents in our particular document corpus; this began to provide interesting but entirely predictable findings. For example, we found that (i) the source and target (we provide detail on the latter entities in forthcoming sections) data models bear few similarities. This is not as trivial as mismatching elements or tags, here we lay emphasis more on the actual nature of the data itself; the inherited characteristics of defining features such as nesting of headings, subheadings, paragraphs, the use of bullet points over numbering, etc. (ii) the source data model lacked many basic features you would expect from well-constructed HTML or XHTML, namely that section headings should be identified under the appropriate tags, similarly with subheadings. Unfortunately, the source HTML did little to define between paragraph beginnings and subsection headings, which if present, would have enabled us to identify, design and construct generic pattern matching expressions further down the

representational formats was considered as the major problem which has to be addressed. Not the fact that individuals couldn't find the documents they required... although ironically this is still quite a commonly acknowledged problem encountered by many individuals tasked with complying with the ever changing legislation!

[4] This figure subsequently breaks down into component parts and on some occasions includes additional document types present to aid with presentation e.g. CSS and JavaScript. This figure also represents documents from across both the regulatory documents for Domestic and Non-Domestic construction within Scotland, however it does not include any additional accompanying documentation which goes hand in glove with the actual regulatory corpus, e.g. procedural guidance, industry recommendations, environmental guidance, etc. It is however also important to first recognise, then acknowledge that the task of ensuring adequate compliance is seriously hindered should the aforementioned supplementary documents be absent from the compliance checking process.

mapping transformation when progressing to constructing XML Path Language (XPath) (W3C, 2010), expressions for XML Stylesheet Language Transformations (XSLT) (W3C, 1999) mappings between source and target data models (iii) defining characteristics inherent within such building regulations include tables, headers, diagrams and equations to name a few, the latter used for fire and energy performance calculations and such like. Unfortunately, correctly constructed mark-up for these features, unique to documents of this nature was in the significant majority of cases absent. In the worst of cases, this was consistent for several consecutive subsections meaning that a large section of work concentrated on the untangling of numeric data which should be correctly attributed to the mark-up of HTML table elements from normal paragraphs, which when combined make up subsections and document sections respectively. The preceding example is by no means an exhaustive description of the type of modelling challenges we encountered during the untangling (commonly referred to as "data cleaning") aspect of the early content analysis carried out over the source data corpus. However, we think it provides a fair indication of the scale of the data analysis required to ensure that consistent mappings can be achieved across an entire document corpus of similar type.

As was briefly indicated in earlier sections, we were able to construct a hierarchy of the data model based on one huge positive being that we were provided with schemas illustrated in Figure 1 produced by a previous research project[5], which described the entire structure the Scottish Technical Standards (STS) must adhere to. Again unfortunately, the 6 XML Schema Definitions (XSD) (W3C, 2012) within the schema suite, suffered from a similar lack of attention to detail as the source HTML itself, resulting in several inconsistencies which were allowed within the STS. One prominent example of such a lack of attention to detail was the tagging of two or more (apparently distinct) Uniform Resource Identifiers (URI's) with the same namespace declaration. As per the World Wide Web (W3C) XSD 1.1 specification document this is classed as illegal resulting in inherent inconsistencies being littered throughout the source HTML. Such inconsistencies permitted the illogical inclusion of handbook descriptions within sections and section descriptions within chapters when in fact the inverse should be permitted. A correctly compliant suite of XSD's would flag such relationships as illegal and prevent such ambiguities littering themselves throughout the HTML.

[5] Unfortunately no resulting literature was produced from this research, with the Scottish Government's website stating the project was a "Proof of concept for a software system to automate formatting of the Technical Handbooks. The prototype system uses extended markup language (XML) and MSInfopath. The only product of the research is the prototype system."

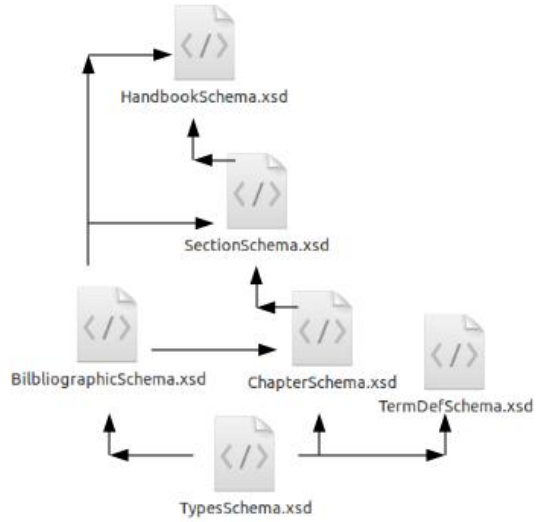

FIG. 1: Hierarchical Breakdown of the Scottish Technical Standards Schema Model

Our experience working with the source HTML, including the content analysis task undertaken, provided an insight into the mountain of problems we (as publishers of open data) are plagued by if we do not initiate a concise, measured approach to representing this data as structured open legislation. In this paper we aim to implant and underlying argument which makes it an important requirement for any formal representation of such legal texts to acknowledge and embed sufficient mark-up (which comprehensively structures and describes the document content from an objective perspective) into the source data regardless of its representational format. If we truly wish to move towards a more open, user oriented, data intensive way of ensuring that compliance is met with regards to building regulations, it is essential that from the outset the drafting and publication methodology e.g. how laws/legislation is developed and eventually communicated to the consumer, needs to start as it intends to move on, which is to make more effort up front (regarding what regulations objectively comprise), allowing us to reap the benefits further downstream throughout an improved approach to utilising digital data. Having commented somewhat on content analysis from an abstract level so far, reasoning behind the attention to detail being provided on this topic is attributable to the culture within which the source data model was required and is currently used. We now however make significant headway into this topic not by discussing traditional channels of thought relating to the philosophical topics of philology, hermeneutics and semiotics, but instead through a discussion centered on authenticity, permissibility and restrictive governance (topics key to legislative representation). With the prerequisites covered we now progress to elaborate on the suitability and applicability of two target data models for suitable representation of UK Subsidiary Legislation.

3. AKOMA NTOSO

The Akoma Ntoso[6] initiative attributes its roots to the requirement to establish a schema that could deal effectively with different legal document types, the different legal traditions and their possibly unique and probably varied exceptions that as we know within independent legislatures are referred to as legal rules. The aim of the effort is solely focused on modelling legal document structure (e.g. parliamentary debates, committee briefs, journals, legislation, judgements, etc.), the legal metadata connected to the document(s) (e.g. the language, roles, processes, publication, etc.) and the versioning of specific categories from within the legislation over time. When referring to the nature of this metadata one will observe its objective nature, this is to say that the metadata we wish to encapsulate within the markup is factual in nature which can be leveraged and fed directly into other legal processes to improve judgement and decision making within the legal work-flow. Working hand-in-glove with Akoma Ntoso (or more accurately LegalDocument ML) is LegalRuleML (OASIS, 2012(b)) which aims to model the *content of legal norms* e.g. obligations, rights, permissions, etc. in order to permit legal reasoning over the top of the underlying legislative data representation layer. "It is positioned between the Deliberation rules and the Reaction rules facilitating the modelling of either norms or business rules. This approach provides support for the implementation of reasoning engines combining both norms and business rules" (Palmirani *et al.* 2011). Akoma Ntoso grasps concepts and lessons learnt from very early work which focused on legal document modelling and document interchange (FORMEX, 2004) right through to newer tecnology standards[7] which focus solely on standardizing the way in which sources of law and references to sources of law are to be represented in eXtensible Markup Language (XML). Akoma Ntoso aims to tie the layers from many efforts together in an attempt to "*Fill the Gap in Legal Knowledge Modelling* " (Palmirani *et al.* 2009).

3.1 Approach to Legislative Localisation

Typically a significant barrier to previous widespread adoption (which can in many cases be centred around the pessimism of skeptics who would rather the situation remain *de facto* rather than invest in improvement) of legislative mark up models (of any kind) within Government work flows is the topic of localisation e.g. how does such a model accommodate the intracacies evident throughout our local/county/borough legislation. Within the Akoma Ntoso data model currently the schema is separated into two parts: first the general (abstract) schema, a vocabularly and a minimal set of constraints that all Akoma Ntoso documents must comply with and a set of stricter schemas: -- the Akoma Ntoso custom schemas, exist in parallel providing more constraints over the same vocabularly of elements to enforce rules of specific document types in specific legislatures. A fundamental requirement for any document satisfying one of these localised schemas is that it also conforms to the abstract schema. Localised schema models would usually also includes optional, mandatory and partially optional rules for compliance with customised schemas (capturing very specific aspects of some legislature such as the

[6] Soon to be OASIS LegalDocument Markup Language as it is currently engaged in the OASIS standardisation process. For more information see (OASIS, 2012(a))

[7] See (metalex, 2010), This is just one of many other legal document efforts used as inspiration for Akoma Ntoso.

acceptance that some aspect of the law is enforceable in), common and custom metadata sets (which in this case could easily relate to design and/or construction concepts such as key terminologies or entities with industrywide recognition/importance) and local extensions (localisation) respectivley. We now progress to substantiate on some of these.

3.2 Overview of Metadata and Document Lifecycle

Within the schema, encapsulation of metadata is separated from content within each Akoma Ntoso document. As one would expect, metadata are split, categorised and structured in subsections to retain integral aspects of any document such as passive change data relating to legislative provenance over time, information relating to publication etc. The objective nature of this metadata is optionally defined within custom and/or common metadata schemas and in some instances ontologies providing discrete or unique structure for metadata additions. The Akoma Ntoso metadata model has been carefully designed to permit extensibility in this fashion in order to make best (re)use of existing metadata vocabularies such as the Functional Requirements for Bibliographic Records (FRBR) (IFLA, 2009) and Dublin Core (DCMI, 2012) to name a few. Another functionality embedded within the metadata model includes a structured naming convention for core metadata which must be adhered to. The adoption metadata naming constructed using namespace prefixes pointing to the country codes of the relevant emanating body ensures that the potential for ambiguity or duplication of naming is removed between common and custom metadata content

As we also discovered in our forthcoming analysis of the Crown Legislation Markup Language (CLML) provenance tracking within legislative documents has become of central importance within the remit of legal document representation. It greatly aids our understanding of certain aspects of legislation and of course improves transparency with regards to legislative access for both humans and within legislative informatics. Akoma Ntoso therefore models provenance through amendments with emphasis on the following two amendment planes: (i) amendments and events in the document lifecycle which can be captured objectively as they happen at a precise moment in time e.g. original approval, coming into force, final repeal, official publication, etc. and (ii) amendments and events within the lifecycle which are due to the enactment of a specific, individual document which can be objectively traced back and identified with a URI e.g. some more authoritative artifact of legislation which amends the particular legal document we are concerned with, we should be able to directly trace this amendment back to the authoritative entity.

3.3 Akoma Ntoso URI Naming Convention

In Akoma Ntoso URI naming is approached from a generic direction with the intent to provide a standardised referencing mechanism of legislative concepts embedded within the document whilst providing flexibility in scope. Flexibility required within Akoma Ntoso URI naming relates to the wide scope of documents which the data model has been designed to represent e.g. acts, bills, amendments, debates and Hansard, judgments, minutes, official reports, etc. Documents with legal status hold most important within the naming scheme (this concept being adapted from FRBR) and subsequently descriptions of such resources are classified as having a type relating to the concepts of Work, Expression, Manifestation or Item.

Work URIs in Akoma Ntoso relate to an abstraction of the legal resource containing the necessary identifying features we would expect from a publishing organisation. Following the domain name (in Fig 2(a)) the top URI comprises the country code, type of legislative document (classification within jurisdiction), explicit specification of document subtype (e.g. Scottish Statutory Instrument as oppose to Statutory Instrument), emanating actors (connected within the flexibility agenda, consider the example where an act or bill would not usually require and actor whilst a ministerial decree would), original creation date, final unique characteristic deemed important as a means to disambiguate the document URI.

Expression URIs concern the any expressive characteristics associated with some particular version of a document e.g. its content with respect to another version of the content of the same artifact of legislation. An example relating to the use of expression within content may relate to changes to the specification or language of a document. Typically Expression URI's (as can be seen in Fig 2(b)) mimic the Work type URI adding a human readable language code (according to ISO 639-2 alpha-3), the @ character, zero or more comma separated version identifiers e.g. the version date of the expression and an assortment of optional identifying features which provide further detail on the expression of the docume.

We now come to Manifestation-type URI's, the sole purpose of which is to characterise the process by which a generated document came to be in its current format, this can on a large part usually be assigned as the document MIME type suffix e.g. .xml, .rdf, .doc, etc. Fig 2(c) (which mirrors the Expression URI adding specifics) displays the appropriate URI pattern.

Finally we arrive at the Item type URI's. Fortunately rules for this type URI are simple and justified, namely that attributes or distinctive characteristics relating to physical storage for recording document manifestations are not included within the URIs. The logic behind this stems from our requirement to access permanent URIs for legal documents. In adhering to best practice for referencing and making reference to legal texts one should never point to physical files in a specific repository under a specific filename as this immediately introduces the intermingled problems associated with static identifiers. Distinguishing between an abstract idea of the destination document and its concrete representation as a computer file allows us to accommodate technological evolutions to our systems, tools and files without a corresponding complete redesign of our referencing mechanisms.

a) [http://www.${nameoforganisation}]/${country}/${legislationType}/${subType}/${actor}/${creationDate}/

   ${uniqueChars}

b) [http://www.${nameoforganisation}]/${country}/${legislationType}/${subType}/${actor}/${creationDate}/

   ${uniqueChars}/${languageCode}/@${versionNumber(s)}/${optionalFeatures}

c) [http://www.${nameoforganisation}]/${country}/${legislationType}/${subType}/${actor}/${creationDate}/

   ${uniqueChars}/${languageCode}/@${versionNumber(s)}/${optionalFeatures}/${manifestation}

FIG 2: Akoma Ntoso URI Naming Convention: From top to bottom: (a) Work-type URI, (b) Expression-type URI and (c) Manifestation-type URI

Although the above does not embody a complete account of URI naming in Akoma Ntoso we bring this section to a close switching to a cross examination of the CLML against the criteria established at the end of the introductory section.

4. THE CROWN LEGISLATION MARKUP LANGUAGE

We begin by introducing the second target data model, the CLML, based on the UK National Archives' http://www.legislation.gov.uk, which aims to encapsulate a governance model for the representation of all UK legislation regardless of its nature. We unearthed several rather specific document features which we noted were missing from the semantics of the CLML target model. We like to think of this section as a rather curious, unintentional by-product of our own research, an unpredictable outcome of our research if you wish, considering the aforementioned problems we encountered with the source data model itself. Finally we move towards the novel aspect of our work, covering areas pertaining to our own additions to the target data model as described.

Since President Obama's ground-breaking announcement during his first day in office, which subsequently led to the eventual launch of the U.S. Government's data.gov project in late May 2009 (data.gov, 2012), we have witnessed many government initiatives from across the global community aimed at opening up public sector, government collected data to wider audiences as pressure behind the open data movement eventually made it to the mainstream. In the UK, data.gov.uk has paved the way for individuals as data consumers, and organisations as data producers to come together to share, consume, link and utilise the ever-growing mountain of publicly funded government data, empowering innovation within business and increased levels of transparency within the communication of public sector data in general terms. This stems directly from the project's mantra which states that it "aims to promote innovation through encouraging the use and re-use of government datasets" (data.gov.uk, 2012). Directly attributable to these two keynote projects was the widespread buzz which eventually led to the use of nouns such as *openness* and *transparency* edging their way into the board rooms of other public sector organisations. It will come as no surprise then that one of these areas, being of key importance to ourselves, was legislation. We, therefore, continue with the launch of the National Archive's Legislation.gov.uk, described by Lord McNally (The National Archives, 2009) as a platform which "presents complex information in a clear and intuitive way. This ground-breaking work puts democracy at the heart of legislation and makes a major contribution to the government's transparency agenda." As the subsection heading indicates, part of that contribution which we wish to discuss in detail are areas of the CLML such as the representational format standing key within this entire project, the promotion of open data standards which follow relevant recommendations of the W3C, interoperability awareness and the governance model which controls all of these elements. The CLML, in combination with other tools and, customised to fit the stringent requirements of UK legislative data provides not only a means by which we can access legislation, whether that be via HTML, PDF, XML or RDF (Resource Description Framework), but more importantly, the bones, the structure to which we attach the fleshy artifacts of legislation as produced by parliament and subsequently drafted within our legislative workflows. For reference we have included a pictorial overview of the CLML

XML Schema module hierarchy in Fig 3[8]. Although one may initially struggle to fully interpret and therefore appreciate the component parts of the target CLML model, we progress to provide a comprehensive description of it in the following sections. One will notice that the modular nature of the CLML schema model make-up provides a very substantial underlying data model (especially with respect to the source model) which is rich in metadata; built using various open standards such as Dublin Core amongst others, the relevant assemblies from within each of these standards being used to express the particular characteristics of the data we encounter within UK legislation. If one were to browse XML mark-up of any legislative artifact once it has been mapped to any of the machine processable formats within the CLML suite, one would find only minimal inclusion of general legislative interchange formats such as CEN Metalex. Paul Sheridan, Head of Legislation Services at the National Archives provides apt reasoning behind this, stating that these formats "lack the expressive power we need for UK legislation, but could easily be wrapped around the XML..." (Sheridan, 2010) This is also reflected in the RDF representations of legislation which are produced within the ecosystem, "the RDF from legislation.gov.uk is limited to largely bibliographic information. We have made use of the Functional Requirements for Bibliographic Records (FRBR) and the Metalex vocabularies, primarily to relate to the different types of resources we are making available." (Sheridan, 2010)

Consequently, before going any further, we wish to make it clear that based upon a detailed examination of the drivers behind the use of the CLML within legislation.gov.uk, it's suitability for representing source legislation similar to building regulations, and finally it's advantages over other general legislative data interchange formats, we were pleased to see the CLML has succeeded in *re-inventing* as little of the wheel as possible with regards to the use of web standards.

[8] N.B. This diagram is intended only to illustrate the point that the CLML data model is a much more expressive model over our source model, which by nature underpins many addition layers of complexity. We allegorize the two models as much akin throughout this section of our thesis in a bid to ensure we communicate a fuller, more comprehensive description of our work

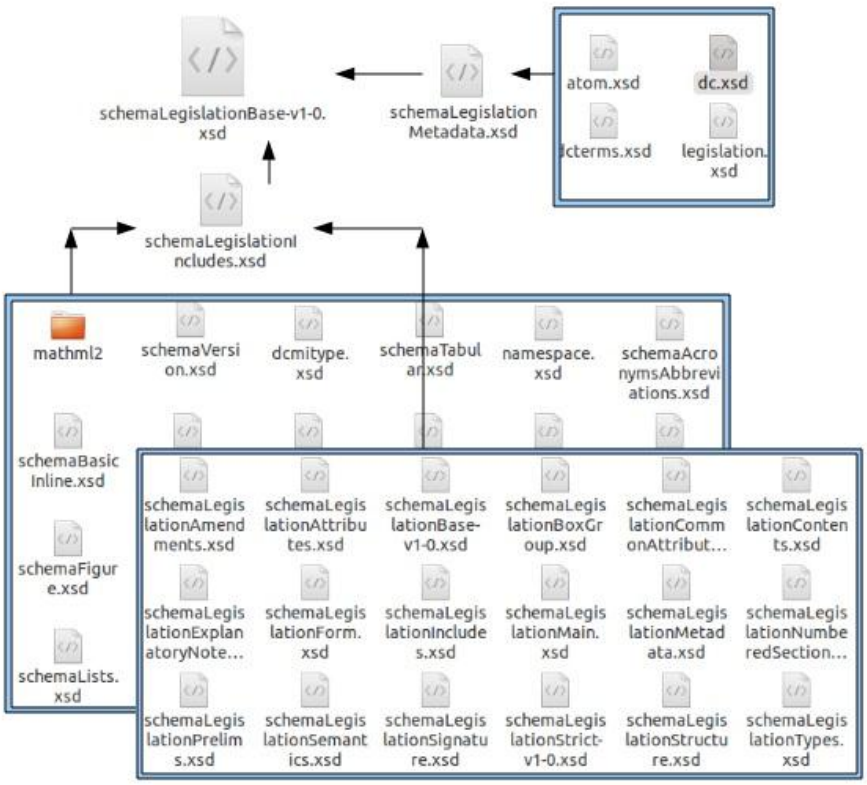

FIG 3: Hierarchical Breakdown of the Crown Legislation Mark-up Language Schema Model

If we briefly cast our thoughts towards quotes relating to the interwoven nature of presentation and content (McGrath, 2010), we will see that by designing the CLML model in this way the underlying problems herewith described are substantially eradicated. These two branches (presentation and content respectively) can therefore be thought of as extension points, from which we can create domain specific customisations within our own schema implementations, which would in turn inherit the governance model we see present across all of the generic legislation. This not only makes the CLML model extremely flexible with regards to changes in legislation, but also provides a consistent approach to governance whilst upholding the integrity and applicability of the model as a whole. As we traverse the model we see increasing evidence of the vision and design methodology adopted throughout its creation, namely that there has been a common acknowledgement that more dynamic legislative artifacts becomes dated, deprecated, and subsequently obsolete on a regular basis, which has the knock-on consequence of increasing the burden on those attempting to use the legislation as they have the additional overhead created by ensuring that legislation they have in their possession is the most recent for their requirements. As previously explained, this is an extremely common problem within building design and construction. To address this situation, the CLML model incorporates versioning and facilitates the notion of expressing changes over time which helps us to understand the underlying metadata semantics behind the surface content. Additionally, we see a strong emphasis being placed upon a suitable approach to expressing semantics. Each artifact of legislation maintains its own URI, as does each heading, subheading, part, and so forth. This not only establishes a common meta-representation of crucial aspects of the legislation itself, but also enables us to easily disambiguate, identify and reference very specific, unique elements of data represented within the model.

4.1 Schema Conformance

As with Akoma Ntoso, a defining characteristic of the CLML model is its noticeable weighting towards metadata-oriented representations of the legislative artifacts it describes which differs drastically from the source regulation model documented previously. Everything within the target model itself stems from a trunk *schemaLegislationBase* (at the top left of Fig 3), which contains elements of schema restrictions specific to all UK legislation regardless of whether categorised as common or criminal in nature. This base module represents the most abstract layer within the entire model. Branching from this trunk Fig 3,, we see a typical example of software engineering, where two further schema abstractions detail intrinsic, more specific component parts of the entire model as it currently stands (of course should any other module be required, another module of abstraction could simply be bolted on to the model). The *LegislationMetadata* and *LegislationIncludes* schema modules (to the right and bottom of Fig 3, respectively) are designed specifically to both substantiate upon, and separate from one another, document independent metadata and document dependent content respectively.

4.2 CLML URI Naming Convention

The naming convention used in the CLML designates three levels to the topic of URI's, namely Identifier, Document and Representation. Fig 4(a) displays the typical characteristics of a Identifier URI; namely that a legislation type (determined initially by classification as primary, secondary, subsidiary or draft type) proceeds the http://www.legislation.gov.uk domain (as all CLML URI naming does), which is then followed in logical order by year, number and section (if one is present). There are exceptions to this naming rule which take into considerations artifacts of legislation which may not have section/chapter numbers however these are not particularly relevant for those interested in using the convention for drafting purposes. Within the scope of Identifier URI's we should mention that each constituent part of the resulting URI does deconstruct into a fine grained logical representation which reflect the expressiveness required to disambiguate particular types of legislation from within the UK. Fig 5 shows a plain text example of such an artefact whereas Fig 6 displays the relative CLML representation. In this case the Identifier URIs clearly define that (i) the artefact of legislation is an Act of the Scottish Parliament (represented by asp) (ii) it was passed in August of 2003 (iii) and Part 1 is the section of relevance.

Secondly Fig 4(b) displays the Document URI's convention which exists to reflect document versioning enabling differentiation between document versions as published on the web. The additional identifier elements appended to the URI refer to the legislation authority (initial document source e.g. the statue law database), the extent to which any given piece of legislation is governing (e.g. Scotland, England, Wales or Northern Ireland) and finally versioning which further resolve to enacted/made, dated and prospective versions of the document in question. The Document URI included within Fig 6 the XML <Pblock> element begins to display this granular approach to expressiveness.

Finally Representation URI's in Fig 4(c) concern the physical representation the document is persisted in the National Archives data store and available to view or process. On large part representations accommodate .pdf, .html, .xml and .rdf as well as explanatory notes (if present).

http://www.legislation.gov.uk/id/${type}/${year}/${number}[/${section}]

a) http://www.legislation.gov.uk/id/${type}/${year}/${number}[/${section}][/${authority}][/${extent}][/${version}]
b) http://www.legislation.gov.uk/id/${type}/${year}/${number}[/${section}][/${authority}][/${extent}][/${version}]
/data.ext

FIG 4: CLML URI Naming Convention. From top to bottom: (a) Identifier URI, (b) Document URI and (c) Representation URI

PART 1

BUILDING REGULATIONS

Power to make building regulations

1 Building regulations

(1) The Scottish Ministers may, for any of the purposes of—

(a) securing the health, safety, welfare and convenience of persons in be affected by buildings or matters connected with buildings,

FIG 5: Plain Text Example Snippet of Part 1 of the Building (Scotland) Act 2003

```
– <Primary>
  – <Body DocumentURI="http://www.legislation.gov.uk/asp/2003/8/body" IdURI="http://www.legislation.gov.uk/id/asp/2003/8/body" Nu
    – <Part DocumentURI="http://www.legislation.gov.uk/asp/2003/8/part/1" IdURI="http://www.legislation.gov.uk/id/asp/2003/8/part/1"
      – <Number>
          <Strong>Part 1</Strong>
        </Number>
        <Title>Building regulations</Title>
      – <Pblock DocumentURI="http://www.legislation.gov.uk/asp/2003/8/part/1/crossheading/power-to-make-building-regulations" IdUI
        NumberOfProvisions="2" Status="Prospective" Match="false" id="part-1-crossheading-power-to-make-building-regulations">
        – <Title>
            <Emphasis>Power to make building regulations</Emphasis>
          </Title>
```

FIG 6: Example CLML URI Naming Convention for Part 1 of the Building (Scotland) Act 2003

5. MODELLING PITFALLS AND CONFLICTS OF INTEREST

Having analyzed both target models this section focuses more specifically on individual internal elements which were identified as crucially aspects of relevance from within the source building regulations and therefore included herewith. In this process we provide commentary close to a descriptive table of mappings displaying why mappings of this nature are by no account simple to achieve – providing a realistic outlook of the problems we faced. By this account our relative goal was to establish whether a justified degree of quality was attainable when mapping elements between both models. For reference, our criteria behind determining the measure of quality, is fully documented in further sections.

Theoretically speaking, the process of identifying relative mappings between source and target models with regards to the document corpus should have been a trivial task; first, both the source and target data interchange formats shared the concept of mark-up, which provides a strong argument for the use of XSLT to achieve the desired output. Secondly both document models share a very similar hierarchical form (as many legislative

resources do), resulting in similar solutions becoming available when the structuring of data becomes an issue. There was further agreement to be found as similarly we did not regard the actual presentation of such data a dominating factor within the scope of our work. We had a use case driven approach to achieving the correct mappings and understood the essential requirement for the semantic content of the data to remain absolutely intact between source and target mappings. Finally, we also maintained a thorough understanding of both the source and target models themselves. So based upon that, what could possibly go wrong? Here are some of the issues that arose during our attempts at mapping our source data from STS source into the target models:

1. We began with a rather bare, poorly marked up representation of the source data. In addition to some of the more trivial discrepancies mentioned earlier, we were unable to manufacture generic XPath expressions in a consistent and generic fashion meaning that we not only relied upon the type of HTML element for guidance, but also heavily relied upon certain element attributes to be able to distinguish between fundamentally crucial aspects of the data. We therefore quickly arrived at *neither here nor there* idiom, where we frequently encountered situations where the model was either too detailed meaning that meticulous XPath expressions had to be manufactured, and in some cases over a hundred cases per document, or else the source was way too vague, we again draw upon the example where it made no distinction between heading, subheading, or paragraph beginnings. This meant that a typical XPath expression aimed at matching a paragraph would also pick up numerous other undesired elements within the document structure. In many situations the model lacked the sufficient mark-up to distinguish the elements we wished to retrieve, meaning that a *data cleaning* process quickly snowballed into a complete re-write of growing sections of HTML. We add that although this is not a sustainable or scalable method to obtain the correct mappings, it was the only method available to us under the circumstances, as we could either start with badly formed HTML, or plain text. After obtaining encouraging results based upon our initial experiments, it was determined that we would progress with the former.

2. Although both target models deal with artifacts of legislation, they both also inherit significantly different logical hierarchies and approaches to embedding structure. They do however share consensus when building from some notion of a *LegislationBase* (in CLML) or *AkomaNtosoGeneralSchema* (in Akoma Ntoso); which provides an interface to common elements present throughout all legislative texts, from which all subsequent documents add domain specific characteristics as appropriate. The source model considers the building regulations as one unique document failing to acknowledge the presence and indeed importance of any other texts. To expand, the source model provides a composition hierarchy comprising of a handbook, which contains sections, containing chapters and bibliographies, finishing with tools to describe data types and term definitions. This in itself displayed stark differences between both governance models which would have significant relative impacts further into our analysis.

3. We also note that the source model works to a page by page principle, whereas both the CLML and Akoma Ntoso do not, instead preferring to utilise the expression, flexibility and power offered within

the URI naming convention as previously described. In this particular case, reasoning and justification behind the shift to the use of URI's holds very strong over its line and page number relative. When we consider document amendments, a common area for confusion is exactly what and where some amendment affected a document, this can be traced back to the line and page approach to authoring documents of this nature. If instead we switch to a notion of URI's for naming, identifying and communicating relevant and important parts of such documents, we completely remove the aforementioned area of confusion entirely.

4. Although we did not consider the presentation aspect of any target data model as an important feature of the overall process, in some cases it became physically impossible to avoid the problems presented by this notion as it persisted to continually plague the mapping processes. It should be noted here that the ecosystem of applications[9] which have been developed around the Akoma Ntoso project helped exceedingly here. Building on point three directly above, within the source model each page subsequently has a header; defining which section the page belonged to, the name of the chapter and the year of publication. No other artifact of legislation available via legislation.gov.uk displayed these characteristics; however this is not the only anomalies we discovered present within the source STS. Additionally the source model includes a rather puzzling obscurity, very specific to technical documents of this nature (possibly unique to this particular document corpus) where every chapter, apart from the first in every section includes a featured regulation description. This description details whether the chapter is of type standard or regulation, the number and whether it is mandatory or obligatory. Additionally the feature includes a summarised description of the entire chapter laying emphasis on the contents which are deemed to be of high importance and which make up essential reading. As far as we know the only way for addressing this anomaly was to produce custom Akoma Ntoso schemas, similarly with the CLML. The schemas are then required to validate against the base/common schema.

5. Finally, as the proverb goes, 'the proof is in the pudding'. One would be naïve to think that complex legislative mappings of this nature would produce fruitful results based solely upon the basis that we are fundamentally mapping from source legislation to target. A key aspect of determining our quality criteria (as explained shortly) could be gathered during the process of validating our output against the base schemas. When facing such a scenario as we were, with the firm requirements that none of the source data was to be removed, edited, altered or missing, during, after or throughout the overall mapping process between representational formats, it is of utmost importance that we safeguard the integrity of the process entirely. Should at any stage we discover that either the output is not in compliance with the base CLML or Akoma Ntoso XSD's, or that some content did not properly map meaning that the semantic integrity is not upheld, the process would stop, solutions would be crafted and when satisfied, we would continue.

[9] See (Bungeni, 2012) and the accompanying Akoma Ntoso OpenOffice plugin

6. DISCUSSION

The design of CLML model appears to stem from a requirement to model common legislative artifacts e.g. (mainly) UK primary legislation and other type classifications of legislation sharing similar structural characteristics. Unfortunately the suite of schema representations is currently not available for public download which immediately makes the model less accessible and hence less appealing than Akoma Ntoso. From a management point of view the most noticeable disadvantage (and disincentive) for organisations operating within the legislative ecosystem who currently author and publish subsidiary and/or supplementary documentation concerns the conformance control and version tracking for the schema sources. As no relative governance control is possible (from a data producer point of view aspiring to use the CLML) it may well be the case that document markup could be superseded by new changes to the schema, this would have the knock-on effect of legacy document makrup becoming obsolete. We primarily consider this an issue of transparency however it also feeds more directly into the nature and scope of the community surrounding the CLML and the subsequent ramifications this may have to successful adoption of the CLML outside of its current field. Additionally there is very little guidance concerning sharing of good practice across potential organisations (e.g. regulation drafters and data producers) who may potentially adopt the CLML for their data modelling requirements. Further the potential for citizen participation remains extremely low[10] resulting in a scenario whereby a large number of the individuals currently interested in adopting the CLML reside within the commercial organisation currently participating largely to the CLML itself, one can't help but envisage a vendor lock in scenario. With regards to technological barriers again there seems to be little or no tools which facilitate the task of producing CLML compliant data. It therefore becomes a completely unrealistic scenario where organisations are expected to somehow develop their own in house workflows to produce CLML schema compliant, interoperable linked legislative data. It should be mentioned that there is a public API (data.gov.uk, 2010) which enables developers to build applications around the data currently hosted within the UK Governments National Archives however again this does little to provision potential adopters with the correct tools necessary to build upon and leverage the CLML for their own modelling needs in a more standardised, openly accessibly manner.

On many levels the Akoma Ntoso suite of schema's and data model originated from an entirely different perspective than the CLML, focused instead on what producers and users of legislative data require within document structure and representation rather than how legislation is perceived to be required from a development point of view. As we explain previously this has resulted in an easily configurable highly extensible model suited perfectly to localized data modeling requirements. As the data format is currently undergoing standardization issues concerning information transparency and governance are greatly reduced simply due to the open standards the data format will comply with combined with the scope for community participation and citizen uptake. An important aspect of the standardization process relates to sharing of good practice both in terms of producing Akoma Ntoso compliant legislation resources as well as strategic modeling for specific localization of document and associated markup. The community and existing technologies which facilitate the

[10] This situation may or may not change, however we expect the former to come to fruition as developers become more interested.

use and adoption of Akoma Ntoso feature high on our ranking criteria. Currently a suite of tools (Akoma Ntoso, 2012 and Bungeni, 2012) exist which include a Akoma Ntoso subschema generator; permitting users to produce custom subschemas which are fully compliant with the full standard, Bungeni; an open source parliamentary and legislative information system which incorporates Akoma Ntoso for data modeling, an Open Office editor plugin for facilitated drafting of Akoma Ntoso compliant texts, a convertor to convert to and from traditional legacy documents and Akoma Ntoso XML, a name resolver to guarantee access to document content and metadata regardless of the storage options and architecture and various post-editing tools.

7. CONCLUSION

Of utmost importance within the remit of this work lies the fact that we were easily able to see production ready examples of the CLML being used within the National Archives' http://www.legislation.gov.uk portal. Whilst on the surface it may seem that this may have made it easier to objectively evaluate the appropriateness of the CLML model as a suitable representation format for building regulations or similar subsidiary legislation, unfortunately this was simply not the case and did not feature within the conclusions of our study. In addition it should also be noted that a community pilot study concerning the evaluation and suitability of Akoma Ntoso for representing source data such as the STS was undertaken as part of ongoing standardization efforts within the OASIS Legal Document ML committee (OASIS, 2012(a)), which aims to advance worldwide best practices for the use of XML in legal documents. However again we consider the primary outcomes of this study as ones which mark basis for accurate and justified comparison between the CLML and Akoma Ntoso rather than to express some kind of bias towards one particular data model.

We mentioned early on in this work that our study would determine a suitability criteria based on concepts identified within the Declaration of Parliamentary Openness and benchmarks observed thereafter. In closing the Introduction to this paper we stated that issues involving transparency of information, accessibility, potential for citizen participation, sharing of good practice, the provision of the models to accommodate open structured formats for data representation and the mitigation of technological barriers for adoption would be considered within the review of both the CLML and Akoma Ntoso within the remit of representing UK subsidiary legislation and associated artifacts. Therefore we now substantiate on our findings.

Finally before concluding we close discussion by stating that although the comparative study presented in the paper acts as a precursor to tasks concerning automation and intelligent authoring of legislative resources, we cannot argue strongly enough that for any process of this nature to be executed an hence laying the path for a tailored solution to producing legislative datasets of this ilk, then it is absolutely essential the source data is produced with end user requirements in mind. Future research in this area should most certainly be based upon the strategic guidelines specified within the Declaration of Parliamentary Openness.

8. REFERENCES

Akoma Ntoso. (2012). *Akoma Ntoso Subschema Generator: XML for parliamentary, legislative & judiciary documents*, Retrieved 10 29, 2012, from http://generator.akomantoso.org

Bungeni. (2012). *Bungeni: open source parliamentary and legislative applications*, Retrieved 10 29, 2012, from http://bungeni.org

CEN Metalex. (2010). *CEN MetaLex: Open XML Interchange Format for Legal and Legislative Resources*, Retrieved 10 29, 2012, from http://metalex.eu

data.gov (2012). *Frequently Asked Questions (FAQ)*. Retrieved 10 13, 2012, from *http://www.data.gov/faq*

data.gov.uk (2010). *legislation.gov.uk API*. Retrieved 10 13, 2012, from *http://data.gov.uk/blog/legislationgovuk-api*

data.gov.uk (2012). *About us*. Retrieved 10 13, 2012, from *http://www.data.gov.uk/about-us*

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

FORMEX. (2004). *Formex Version 4, Formalized Exchange of Electronic Publications.* Retrieved 10 29, 2012, from http://formex.publications.europa.eu/formex-4/formex-4.htm

IFLA. (2009). *Functional Requirements for Bibliographic Records.* Retrieved 10 29, 2012, from http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

McGibbney, L. J., & Kumar, B. (2011). A Knowledge-Directed Information Retrieval and Management Framework for Energy Performance Building Regulations. *Computing in Civil Engineering (2011)* (pp. pp 339-346). Miami, FL: ASCE.

McGibbney, L. J., & Kumar, B. (2011). The WOMBRA Project: A Web-Based Ontology-Enhanced Multi-Purpose Building-Regulation Retrieval Application for Scottish Technical Standards. *Proceedings of the 28th International Conference of CIB W78* (p. 64). Sophia, Antipolis: CIB.

McGrath, S. (2010, 06 04). *Sean McGrath's Weblog*. Retrieved 05 12, 2012, from http://seanmcgrath.blogspot.co.uk/2010/06/xml-in-legislatureparliament_04.html

OASIS. (2012(a)). *OASIS LegalDocumentML (LegalDocML) TC*. Retrieved 10 29, 2012, from https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legaldocml

OASIS. (2012(b)). *OASIS LegalRuleML (LegalRuleML) TC*. Retrieved 10 29, 2012, from https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legalruleml

OpeningParliament.org. (2012, 07 10). *Declaration on Parliamentary Openness.* Retrieved 07 10, 2012, from OpeningParliament.org: http://openingparliament.s3.amazonaws.com/docs/declaration/Declaration%20on%20Parliamentary%20Openness%20-%20for%20Comment%20-%20062712%20-%20OpeningParliament.org.docx.doc

Optimal Economics Ltd. (2011). *Research Project to Support the Appointment of Verifiers from May 2011: Analysis to Establish a Baseline for the Future Operation of the "Reasonable Inquiry" Functions.* Livingston: Directorate for the Built Environment.

Optimal Economics Ltd, BRE, & Liz Sheil Associates. (2011). *Research Project to Identify a Future Model for Reasnable Inquiry.* Livingston: Directorate for the Built Environment.

Palmirani, M., Contissa, G., & Rubino, R. (2009). Fill the Gap in Legal Knowledge Modelling. *Lecture Notes in Computer Science*, pp. pp 305-314.

Palmirani, M., Governatori, G., Rotolo, A., Tabet, S., Boley, H., & Paschke, A. (2011). LegalRuleML: XML-Based Rules and Norms. *Rule-Based Modelling and Computing on the Semantic Web*, pp 298-312.

Pye Tait Consulting. (2012). *Development of Key Performance Indicators to support the building standards verification system.* Livingston: Directorate for the Built Environment.

Sheridan, P. (2010, 08 14). *VoxPopuLII*. Retrieved 03 15, 2012, from Legal Information Institute: http://blog.law.cornell.edu/voxpop/2010/08/15/legislationgovuk/

The National Archives. (2009, 07 29). '*Groundbreaking' legislation website launched*. Retrieved 03 15, 2012, from The National Archives: http://www.nationalarchives.gov.uk/news/478.htm

The Scottish Government. (2012, 07 11). *Data Publication*. Retrieved 07 11, 2012, from The Scottish Government: http://www.scotland.gov.uk/About/Information/FOI/datapublication

W3C. (1999, 11 16). *XSL Transformations (XSLT) Version 1.0.* Retrieved 10 29, 2012, from W3C http://www.w3.org/TR/xslt

W3C. (2010, 10 14). *XML Path Language (XPath) 2.0 (Second Edition).* Retrieved 10 29, 2012, from W3C http://www.w3.org/TR/xpath20/

W3C. (2012, 04 05). *W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures.* Retrieved 10 29, 2012, from http://www.w3.org/TR/xmlschema11-1/