

# A Submission to the Government 2.0 Taskforce

**Andrae Muys <andrae.muys@gmail.com>**

Senior Software Engineer

Metadata and Informatics

## Executive Summary

*The Taskforce is charged with finding ways of accelerating the development of Government 2.0 to help government **consult**, and where possible actively **collaborate** with the community, to **open up** government and to **maximise access** to publicly funded information through the use of Web 2.0 techniques.*

*(Towards Government 2.0: An Issues Paper, Page 3, emphasis added)*

This submission interprets this to mean:

The application of Web 2.0 techniques to

1. Provide mechanisms to improve government policy formulation by permitting earlier, broader, and better informed participation by the general public.
2. Provide mechanisms to improve governance and oversight through greater transparency into the formulation and administration of government policy
3. Through increased access to publicly-funded information, to allow individuals, organisations, and other agencies to benefit the public good.

On the first two points this submission raises two issues for consideration by the Taskforce:

1. The Web 2.0 view of even traditional documents as dynamic 'living records' with a transparent revision history; and, the implications of this for transparency, accountability, and as focii for collaboration and community formation.
2. The need for a re-evaluation of the legislative, regulatory, and cultural norms relating to the participation of public servants in the public sphere. Specifically a need to alleviate the unreasonable level of jeopardy they face though participation with Web 2.0.

On the third point this submission wishes to contribute expert advice on two points:

1. A model to assist with reasoning about data interoperability; specifically, the different levels of interoperability, the various questions each raises for consideration, and the technologies and standards available to address them.
2. The distinction between inverted and structured indices, and the ramifications for the development of search functionality by Government.

## On Public Access to 'Living Records'

**summary:** *The elimination of publication as the dominating cost to information dissemination requires rethinking the definition of public record. The existing FOI/RTI reforms envisaged by Australian Governments may not go far enough. The policies, guidelines, and regulations currently treated as 'publications' should be progressively transformed into living documents, subject from their initial inception to perpetual beta. We*

*need to change our perception of the drafting process from a process of drafting and subsequent publication to a process of curation and moderation.*

One fundamental limit on the mechanisms of government accountability is the cost of communication. The benefit of providing greater transparency into the processes and deliberations of government is weighed against the costs associated with providing the access to the associated records. Our traditional governance and oversight policies and procedures originate in the costs of physical publication and duplication, and a record's value in providing transparency and accountability in government.

As communication has become cheaper, government has responded by providing increased access to public records. Records that were once only available at great cost, or in-person at various libraries, archives, and relevant government offices are now routinely available at no cost online. But while access has improved, the records and their drafting processes are to a significant extent a reflection of technological constraints that no longer apply.

The past 20 years of communications technology developments have successfully reduced the essential cost of many-to-many publication to zero. The remaining costs are those associated with the research, collection, and analysis of data; the solicitation, assessment, and collation of contributions; and the drafting, revision, and approval. This shift undermines the existing rationales that structure present governance and transparency. Instead a new balance should be sought from the cost of curation and mediation.

In principle citizens should not have any restriction on their ability to examine and understand the decision-making processes of the institutions that regulate and provide for them. It is now feasible for what were static records, subject to a static linear process of drafting, editing, approval, and publication, to become 'living documents', published and publicly available from the first draft and in 'perpetual beta'.

It is the opinion of this submission that at the heart of Gov 2.0 is a process of cultural change; one in which the general public are not seen as stakeholders to be consulted, but rather collaborators. Where government experiments with, and builds the skills and process required to surrender its role as the author of policy, and instead becomes a policy curator. For this to happen public records would need to be released as transparent revision histories from the very first conceptual outline, to the final approved publication.

## **On the Role and Regulation of Public Servants**

**summary:** *In an environment of open consultation and perpetual beta, errors and omissions become matters of public record. As such public servants need to be provided room to fail, if they are not to be forced into paralysis or subversion of the access policy. To operate successfully Gov 2.0 must accept the existence of errors and implement tight corrective feedback loops seeking a trajectory of increasing accuracy. It cannot work if public servants are in constant fear of criticism and rebuke for the errors and omissions that are a natural part of any drafting or problem solving process. It is also worth noting here that a shift from being authors of policy to public curators frees public servants to collaborate as citizens in the public contemplation of policy.*

The internet pioneer Roy Fielding has observed that at internet scales, systems can only operate if "As the number of consumers increase, the per-consumer cost of the overall

system must decrease in order to sustain the system.”<sup>1</sup> Dr Fielding applies this principle to the technical architecture of the internet, however the same principle applies to social-structures that wish to accommodate internet-scale. If it is expected that the internet will provide us the ability to increase the scale collaboration, we will need to ensure that the mechanisms we use distribute the costs of collaboration amongst the participants. One of the costs that must be addressed is the cost of validating the specific acts of public servants as they participate in online collaboration.

The current review and approval processes that ensure the accuracy of government communications to the public impossibilise the full cost of validation on the public service. When the volume of public sector information, public records, and public collaboration and discussion implied by Gov 2.0 is considered it becomes obvious that we must address Dr Fielding’s “Economies of scale” principle for sustainable internet systems.

The result must be a reappraisal of the legislation, regulations, and social-norms governing the performance of Public Servants. Prior review and approval by suitably trained and authorised officers, of all publications, releases, and communications by public servants, will be impossible in Gov 2.0. In fact the act of moving from authorship to curatorship preempts the very reviews and other procedures that currently identify the mistakes, errors, and omissions, natural during any drafting process, prior to publication.

Gov 2.0 needs public servants to have room to fail.

The current legal and career jeopardy public servants face for such failures is severe. This will undermine any attempt at implementing Gov 2.0 until it is removed. In place of censure and rebuke must be an acceptance of ambiguity. A frank acknowledgement that various governmental sources of information, whilst public, will be pre-curation/pre-validation. In these cases it will be necessary to liberate public servants of many of their obligations and instead of punitive frameworks, institute rapid feedback paths from the public to the public service to allow the work of validation to be shared by the users of the information. By distributing the cost of validation in this manner, Gov 2.0 will be permitted to scale.

## **On Communication and Data Interoperability**

**summary:** *While the essential cost of modern communication is now effectively zero, there remain risks that accidental costs may remain. The most serious of these is the use of proprietary data formats for public records and the public data necessary to inform policy debates. A matter of concern is the rather crude models of interoperability used in discussions of open-data. A richer model of data interchange would permit a more nuanced discussion of the issues involved. We can define data-interchange as a process operating at five levels: medium, encoding, syntax, denotation, and connotation. Much of the discussion appears to focus on syntax technologies, yet syntax is relatively unimportant. Several recent technologies, including RDF, address the more significant issues of denotation and connotation. It is very useful to think in the terms of the five level ‘semiotic stack’ when trying to understand where these technologies fit; how they interact with existing standards; and what problems they don’t solve.*

Data Interoperability is an act of communicating meaning between automatons. It has been long recognised that there are different levels of meaning that can be communicated

---

<sup>1</sup> <http://roy.gbiv.com/untangled/2008/economies-of-scale>

between humans—the same is true of computers. Between computers, these levels of meaning can be classified into five distinct levels.

The most basic level of digital communication is the bit, transferred via a shared medium. While this may seem trivial, the need for a shared medium does lead to serious questions. It is this requirement that forces us to consider accessibility for the disabled; access to internet facilities for the disadvantaged; and, access to network bandwidth of sufficient power to participate in web 2.0 in regional areas. Questions all too easily drowned out in the multitude of issues to be addressed at higher levels. I would urge the taskforce to adopt a systematic approach to interchange to ensure room is made for the consideration of these important issues.

The second level is the grouping of bits to encode symbols from a shared alphabet<sup>2</sup>. In an image these may be pixels; in a sound file, samples. In a textual document the bits are grouped to encode characters. Fortunately almost all human languages can now share the same encoding through the use of unicode. Prior to this, encoding used to present significant challenges. This still reminds us to consider how Gov 2.0 is to address non-English speakers.

Connotation
Denotation
Syntax
Encoding
Medium

The third level is the recursive structuring of symbols into basic units of uninterpreted meaning using a syntax. This process is called parsing, and an analogous comparison in prose is the grouping of letters as words, then phrases, sentences, paragraphs, etc. In human communication this task often cannot be performed without reference to the meaning of the various words. When the communication is intended for computer interpretation, requiring interpretation to complete a parse will often lead to ambiguity, inefficiency, and error and so is to be avoided.

Discussions about “data formats” often take place at this level. Scanned image, vs. PDF, vs. spreadsheet, vs. CSV, vs. XML, vs. JSON, ... the debates are endless; and ironically mostly wasted time. The only real distinction that matters at this level is the degree of interpretation required to extract the intended structure of the data. So images are bad, but a regular, tabular PDF file may well be better than a messy spreadsheet. CSV, JSON, and XML are similar in this regard, and only better or worse in the context of a specific task.

The fourth level provides a mapping from the basic units of meaning identified by the syntax, to the canonical definitions they denote. Colloquially, the “dictionary definition”. For example, denotation allows interpretation of “The window of glass”, but not “The window of opportunity”. Similarly denotation allows a computer to correctly interpret the statement

```
<doc:1234> <dc:creator> <id:AMuys>  
as
```

---

<sup>2</sup> I am using the word ‘alphabet’ here as it is used in computer science automata theory: A finite set of symbols. The distinction made in linguistics between alphabets, syllabaries, and logographies is not made here.

The entity named 'id:AMuys' is the 'entity primarily responsible for making' the resource named 'doc:1234'.

While a computer cannot understand the import of this statement, it can be instructed, by a human who does, to react in an appropriate way. This might include adding an entry to an index, or a search result; or determining the legal status of the document.

The most promising collection of standards supporting computerised denotation is that associated with the Semantic Web. It is important to recognise that most of these standards are independent of syntax, operating at a different level of communication to the XML standards to which they are often compared. The core standards are:

- The Resource Description Framework (RDF) defines an abstract data model that describes what a syntactic unit of meaning must (and must not) contain to be denotable.
- RDF Schema (RDFS) defines a vocabulary for providing canonical definitions for properties. It is using RDFS that the term "dc:creator" is defined as "the entity primarily responsible for making the resource".<sup>3</sup>
- Simple Knowledge Organization System (SKOS) defines a vocabulary (using RDFS) for defining the controlled vocabularies used in thesauri, classification schemes, subject heading systems, and data dictionaries.

The Dublin Core Metadata Initiative, the Australian National Archives, the various State Archives and records agencies, and various other organisations and individuals all use these standards to define interoperable sets of terms supporting interoperable denotation. AGLS, AGRkMS, DCMI, FOAF, DOAP, vCard, iCal, all have RDF/RDFS/SKOS descriptions that permit improved interoperability. Naturally these standards can also be used to provide denotation within other data models<sup>4</sup>; such as traditional relational schemas, and the containment-hierarchies often used with XML based standards.

The fifth and final layer captures the implications of the communication in the context of the receiver: the connotation. Here the forces of operational, cultural, and ontological context that limit the ability of any to entities to achieve perfect communication of information. To the extent we can denote canonical descriptions of ontology and operation we can mitigate the effects of context<sup>5</sup>; however, we are dealing with people and paradox, there is no perfect solution. Hence consideration of connotation returns us to the need to acknowledge error and omission, and the feedback cycles guide us ever closer to an unattainable goal.

---

<sup>3</sup> See <http://dublincore.org/2008/01/14/dcelements.rdf#creator> ; or, <http://dublincore.org/documents/dcmi-terms/#elements-creator> for a human readable version.

<sup>4</sup> Microformats provide explicit support for a few of these, as microformats don't have any concept of a namespace. Due to this shortcoming affecting government's ability to support existing vocabularies, it is recommended that the use of microformats be deprecated in favour of RDFa.

<sup>5</sup> The Web Ontology Language (OWL) can define various levels of ontological context; Business Process Execution Language (BPEL) is a commonly used standard that can be used to provide a degree of operational context.

## On Search Facilities and the Implications for Government Web 2.0 Procurement

**summary:** *There are different types of search facility. Some lend themselves to federation —where the search is distributed across agencies and the results recombined; others require centralisation to be effective. Inverted Feature Queries, normally full-text search, is easy to federate; Structured Queries, against the relationships between and attributes of records, requires greater centralisation. Hence the complete centralisation of metadata indices forms a part of the infrastructure envisaged for Gov 2.0. Yet government should resist the temptation of the grand gesture. The scale and scope of this final goal is such that no-one knows how to achieve it, yet. It must therefore be developed incrementally, with feedback processes, and without fear of duplication of effort. It will take multiple attempts in parallel exploring the problems and the potential solutions to reach the goal.*

It is hardly insightful to note that the volume of data already provided by government, let alone that contemplated by Gov 2.0, defies browsing. While this may suggest that the development of a comprehensive search facility is a critical prerequisite for the success of Gov 2.0, Web 2.0 permits cooperation at a scale previously unimagined. Crowdsourcing is capable of harnessing tens of thousands of individuals to achieve what no centralised facility could ever do, whether analysing radio signals for extra-terrestrial life<sup>6</sup>; combing through thousands of pages of evidence<sup>7</sup>; or, searching for prior art to invalidate a patent<sup>8</sup>.

Yet crowdsourcing does have one critical weakness: it requires a substantial community of interest. So while a search facility is not required for Gov 2.0, it does provide individuals and small constituencies the ability to navigate the sea of public sector data in support of less engaging objectives. One cause of confusion is that there are three distinct types of search, each with very different properties.

1. Structured Search
2. Metadata Search
3. Inverted Feature Search

Structured Search applies where the data itself has a defined regular structure, with a readily accessible denotational semantic. These include traditional databases, many spacial databases, much tabular data, and the metadata component of Metadata Search (see below). The defining property of these queries is that they focus on comparisons between different records. This requires that structured records be centralised if a comprehensive and efficient search facility is to be provided. The query may then be applied directly to centralised indices that form the data itself—this is a **database**.

Metadata Search applies where the data requires human intervention for interpretation, but the individual records have properties that can be described in a defined regular structure. A common example of this is an email archive: the email bodies themselves require human interpretation; however, they also contain extensive set of well defined

---

<sup>6</sup> <http://setiathome.ssl.berkeley.edu/> remains to this day the worlds most powerful super-computer.

<sup>7</sup> See [http://tpm.apperceptive.com/muckraker/2007/03/tpm\\_needs\\_you\\_to\\_comb\\_through.php](http://tpm.apperceptive.com/muckraker/2007/03/tpm_needs_you_to_comb_through.php) for an example where crowdsourcing was able to read, in detail, 3000 pages of emails and other evidence in less than 9 hours.

<sup>8</sup> <http://linuxdefenders.org/> a crowdsource based response to software patents by the free software community.

headers amenable to structured search. Metadata queries share many of the properties of structured queries, including the need for centralisation; however, only the metadata need be centralised, as the data itself is not searched, only the metadata. The data itself is often distributed across the network, while queries are applied to a central metadata index—this is a **catalogue**.

Inverted Feature Search applies where data, while requiring a human to extract meaning, also has a computer-processable syntax. This allows automated extraction of features, which are then indexed. The most common example of this style is full-text search; where, the syntax describes a text document as a sequence of words, and an index is built mapping words to the documents that contain them. One important property of inverted feature queries is that whether a record matches a given query (or not) can be determined by examining the record in isolation, without reference to any other. As a result this style of query is imminently distributable, in fact, gains no benefit from centralisation. The query is therefore distributed to an arbitrary number of indices distributed across the network. Another name for this kind of search facility is **concordance**.

Due to their distributed nature, full-text search engines will be the easiest to develop and deploy. A centralised metadata and structured search facility introduces complexities of scale that are staggering. Consider a conservative estimate of the scale involved: 150,000 public servants × 5 records/day × 10 attributes/record × 200 days/year = 1,500,000,000 attributes/year just for federal business-of-government recordkeeping. This does not include spacial, statistical, or regulatory data collected by government. Nor does it include any of the data associated with state and local governments, or government corporations.

The simple fact is that we do not yet know how to manage data at this scale. One serious risk is that the government will be tempted to announce a sweeping vision, the grand gesture that will establish a massive project to develop a centralised catalogue and search facility. The success of inverted feature search facilities at this scale, such as Google or Yahoo, is deceptive. For reasons given above, these facilities are only nominally centralised, in practice they succeed only because they are able to exploit the distributed nature of this kind of search. The techniques they use will not work when applied to metadata and structured search. Instead of a single program, we should focus on building capabilities and solutions iteratively by:

1. building small-scale metadata query facilities within departments, agencies, and even workgroups.
2. launching small-scale integration projects, to merge existing metadata catalogues.
3. an iterative approach of medium-scale integration, continually integrating ever larger catalogues.
4. a mediator and curator, most likely in the OIC, that doesn't dictate, but rather observes and records the various projects and is responsible for disseminating best-practices, standards, and technology platforms, amongst the various levels of government.

As with most other aspects of Gov 2.0, this approach works by celebrating duplication of effort, experimentation and even failure. Even where we know where we are going, we have little or no idea how we will get there—it is only by parallel, redundant attempts that we can explore and map the paths available to us. This is a recurring theme. To succeed government will need to give public servants permission to experiment, permission to innovate, and therefore permission to occasionally fail. Instead of insisting on success, insisting on improvement, and building into every plan feedback mechanisms to learn from success and failure alike.