

# Cataloguers may yet inherit the earth

by

Lloyd Sokvitne

Manager (Information Systems Development)

State Library of Tasmania

Email: [lloyd.sokvitne@central.tased.edu.au](mailto:lloyd.sokvitne@central.tased.edu.au)

## Abstract

*The development of Service Tasmania Online, the Tasmanian government portal, is described. This required consideration of the specific methodologies that meet the special requirements of resource discovery on the World Wide Web. These special requirements lead the State Library of Tasmania to develop an user-centric model for the provision of an online government discovery service. The conceptual and pragmatic characteristics of the information architecture that underpins the Service Tasmania Online service are discussed, with an emphasis on the capacity to provide multiple avenues for resource discovery that assume no prior knowledge of content format and production structures. The use of XML, search engine technologies, and special software applications to provide a data driven government portal are explained. The role of libraries and information professionals in the production of quality metadata and discovery systems is explored.*

## Introduction

Metadata has become a popular topic and its importance to Web discovery has become an accepted axiom. Business managers within government look at metadata as a way to utilise minimal-cost options to realise the investments made in web content production (Office for Government Online, 2000).

The significant component missing from current metadata development is an understanding of the outcomes of the Web retrieval process as they relate to the needs, capabilities and searching behaviours of actual users. To some extent this is understandable: the web is so new that we don't in fact have a great deal of well-developed conceptual models or relevant research to drive discussions on Web retrieval. However, with metadata applications now rapidly emerging across Australia, it is important to develop metadata-based systems that address the outcomes that are produced for the end user rather than systems that simply focus on reducing the costs of metadata production.

This paper will outline the experience of the State Library of Tasmania in providing a metadata based information and retrieval system for the State Government of Tasmania. The development of the central online government portal in Tasmania, *Service Tasmania Online*, has required the development of an information architecture based on user needs, the provision of new web-orientated retrieval vocabularies, and an uncompromising approach to metadata quality. The experience gained from this process has highlighted the important role that libraries and professional information retrieval experts (librarians) can and should play in the metadata arena.

## Developing an appropriate information architecture

*Service Tasmania* was established in 1997 to provide a simple integrated way for the Tasmanian community to access government services. This integration of government services is provided through three channels: over the counter, over the telephone, and over the Internet. There are

now 24 *Service Tasmania* shops around Tasmania, each providing access to over 300 government services. *Service Tasmania* also provides a single phone number to contact government, and an integrated voice response service that allows over 20 payments for government services to be made over the telephone. In 1998, the State Library of Tasmania was contracted by the State Government of Tasmania to develop a web site that could deliver *Service Tasmania Online* over the Internet.

The process of building the *Service Tasmania Online* web site required the formal development and implementation of an information architecture. An information architecture is here used to describe the provision of an integrated navigation, searching and presentation environment (Rosenfeld, 2000). Formal information architectures are a way to counter the often haphazard use of menus, search engines, tables of contents, directories, and other retrieval options that are provided on web sites to allow the user to find what they are looking for. An information architecture ensures that retrieval is a major component of web site planning, working with the graphic design and content production components.

The underlying objective of the *Service Tasmania Online* information architecture was to allow the client to easily identify and find government content that was relevant to their needs and without prior knowledge as to the format of that content (static web pages, PDF documents, forms, online-transactions, etc) or the area of government (be it local, State, or Commonwealth) where that content would be found.

## **Understanding the web information environment**

The size and diversity of Web publishing is now almost taken for granted. It is an information resource of almost unbelievable size, but one without order or structure. Yet it gives the average user an illusion of order (through search engines and services such as Yahoo and AltaVista), through portal sites and services (AOL, Microsoft, ISPs), and through individual web sites that present themselves with great credibility and professionalism irregardless of the true value or purpose of the content.

Retrieval methodology on the web has developed within this open and unstructured environment, where tools have purposely developed to provide simple and easy to understand interfaces based on word searching and simple directories. This is an environment where retrieval services can produce results irrespective of the skills of the client, where the medium itself, hypertext linking, can provide the retrieval method and deliver the actual results, and where the provision of intuitive options and visual cues are critical to the effectiveness of retrieval tools.

The information architecture for *Service Tasmania Online* had to deliver a retrieval system that could match this expectation for simple and intuitive interaction. *Service Tasmania Online* would also require a web site that provided the user with online government content in an ordered and structured way, even if that underlying content came from uncoordinated government services and disparate structures.

## **A user centric model**

The provision of an user-centric information architecture for *Service Tasmania Online* required an understanding of the behaviour of users as they search for content within the government sector. However, the range of government online content available, especially across all three tiers of government, is extremely variable and undergoing rapid but uneven development. This makes it difficult to analyse user needs in terms of identifying specific content outcomes that could meet

any given need. This led us to instead identify the types of outcomes that would meet a wide range of user needs. These types of outcomes were identified as follows:

- 1) **Known item searching:** the user is looking for a specific resource that they know exists online, e.g. the home page for the consumer affairs office, the Tasmanian Dangerous Goods Act of 1997, an application form for a business license, etc.; users will expect to find the specific item easily in their search
- 2) **Known resource searching:** the user is looking for a limited set of specific resources which they expect to find online if they exist: eg. consumer affairs support, legislation related to dangerous chemicals, business license information; users will expect to find all relevant resources easily in their search
- 3) **Known topic searching:** the user is looking for information on a particular subject but has little knowledge as to the breadth and nature of resources that exist: e.g. consumer affairs, legal issues, establishing a business; users will expect to find all relevant resources easily in their search but will also expect to contribute to the process of selecting and evaluating relevant resources. The difference between Known Resource and Known Topic searching is largely the specificity of the search, but this distinction is important because it affects user behaviour.

The outcome of this model was that the *Service Tasmania Online* information architecture had to produce retrieval sets with high relevance and with low recall. The integrated government online service would have to bring together those resources that precisely matched what the client expected to find, in the way that they searched for those resources, and without forcing the client to sift through large result sets. This is a major challenge as most web retrieval mechanisms produce large result sets and rely heavily on the client to scan and identify the resource that matches their query.

## **Providing a variety of tools and retrieval mechanisms**

The need to accommodate user behaviour is made more challenging because users do not necessarily exhibit a single approach when adopting retrieval strategies on the web. Instead it must be assumed that users may act differently when searching for different types of content, or when assuming contextual persona. Their behaviours will reflect different states of knowledge in a given subject area, different assumptions about content, and the different social and emotional contexts that generate specific user needs. The solution to this issue is to provide a variety of discovery tools and to design the web site so as to present them all as equally accessible alternatives.

*Service Tasmania Online* therefore set out to develop an information architecture that could provide a wide range of retrieval options. All of these options should allow easy navigation and browseable capability, but must yield high precision and relevance to all types of choices with low recall. The structure also had to be easily adaptable, so that as more government content develops and as more knowledge about user behaviours is gained, retrieval options and navigation structures can be changed.

The retrieval methods to be provided on the *Service Tasmania Online* web site were then defined as follows:

- Capacity to search, using browseable structures, by the general class of resource sought, the target audience, and by subject area
- Capacity to use an easily navigable topical structure that matches the scope of government resources
- The ability to deliver alternative options such as:
  - Service Packs (groupings of government services to provide a specific outcome, eg. housing assistance)

- Life Events (groupings of government services related those key events, eg having a child)
- A-Z government listings (lists of government organisations by department, portfolio, tier of government)
- A-Z subject listing (alphabetical listing of subject terms)
- What's New (new government resources available)
- What's Topical, etc. (popular government resources)
- The ability to provide free text searching access to both the descriptive metadata and the free text of the actual resources described
- Capacity to provide synonym-matching for common free text search terms
- Free text searching across the documents related to a described resource (eg searching an annual report would include all the contents, rather than just the entry heading)
- Free text searching across all content located on government web servers (separate harvester), not just those with metadata

## Metadata as the key

An information architecture was created for *Service Tasmania Online* that relies on a central metadata repository to describe government content, to manage the Web site, and to provide dynamic web site navigation that constantly displays the variety of retrieval methods available. This is a surrogate-based architecture that separates the real resource from the description of the resource, and that allows the development of additional functionality based on those surrogates.

The *Service Tasmania Online* metadata repository is maintained by librarians at the State Library of Tasmania and is able to provide high quality, consistent, and comprehensive web results that are independent of the extent and quality of metadata available from various jurisdictions and online service providers. This also allows *Service Tasmania Online* to select government content for inclusion in the metadata repository, and meet a core objective to produce high relevance and reduced recall.

A significant enabling component of RDS has been the development of a natural language, browseable thesaurus that guides navigation through the Web site. Reflecting both government services and customer needs, the thesaurus allows users to quickly choose the information they are seeking. The AGIFT thesaurus and the Community Information Thesaurus were used as a basis for this online thesaurus.

## XML as the building block

The *Service Tasmania Online* development project created special software to allow the metadata repository to itself drive web site navigation and content retrieval. This system, the Resource Discovery System (RDS) uses the individual resource descriptions contained in the metadata repository to deliver content through standard presentation templates. The metadata records contain Australian Government Locator Service (AGLS) metadata as well as other data descriptions need to meet specific *Service Tasmania Online* requirements.

The metadata descriptions are stored in eXtensible Markup Language (XML) format. XML was chosen for its simplicity, the flexibility it provided, the functionality that could be leveraged through standard search engine technology, and the long-term application transparency and independence of the data.

The RDS system is designed to take a number of different elements within the XML records that describe a given resource and to interactively use those elements on the Web site as navigation and retrieval options. These options are context-sensitive and appear wherever relevant across all searches. The use of a number of distinct descriptive categories reflects a faceted classification approach, with each of those categories hereafter referred to as a facet. *Service Tasmania Online* currently uses three key facets to describe government content on the Web: main topic, type of task, and target customer group. It is possible to change these facets or add additional ones as new needs emerge.

RDS uses the facets contained in the resource descriptions to generate a post-coordinate web navigation system, whereby access to the resources is provided on the web site according to the permutations of the facet combinations. In other words, it is not necessary to access a specific facet by following a pre-defined hierarchy, the choice of one facet leads to the inclusion of alternative relevant facets on any results page where there are resources that match those alternative facets.

It is possible to search on the *Service Tasmania Online* web site by main topic, task, or customer, choosing those facets in any order. For example, it is possible to start with a task, then choose a main topic, then customer, or alternatively choose the target customer group first, then the topic, then the task, etc.

The use of post-coordinate faceted navigation within *Service Tasmania Online* means that if all three facets are applied, the site will provide a maximum of :

$$((F1 \times F2 \times F3) \times 3!) + ((F1 \times F2) \times 2!) + ((F1 \times F3) \times 2!)$$

facet-based pathways for resource discovery (where F1 is the number of possible values for Facet 1, F2 is the number of possible values for Facet 2, and F3 is the number of possible values for Facet 3. Note that certain facet combinations are excluded from the formula because main topic (F1) within *Service Tasmania Online* is required to display the results, which means that some combinations such as (F2x F3) and (F3x F2) yield no resources. With the addition of topic-derived pathways, *Service Tasmania online* provides over 7000 possible pathways that are delivered in real time on the web site.

## **Underlying software development**

Special software processes were developed to meet the performance requirements of delivering all the required navigation options on the live web site. The actual indexing of the XML records is performed via a search engine and the high performance of modern search engines allows *Service Tasmania Online* to treat the choices made by the users as they navigate the web site as search queries and to deliver those results as if static pages had been chosen.

Additional software utilises the off-line construction of tables that describe all the relationships between the facets that exist in the data. The consequent loading of these tables into memory allows them to act as online high performance databases delivering real-time relationship information as the client navigates the web site.

The actual text contained in the resources that are described by the XML records in the metadata repository is also gathered by the same search engine to enable composite free-text searching. A full-harvest of all government sites is also performed regularly, allow free-text searching of government content that is not included within the central repository.

## **Quality assurance and standards adherence**

The reliance on metadata to drive the entire *Service Tasmania Online* web site has meant that the metadata must be reliable and accurate. Quality control and internal metadata standards adherence are fundamental to the operation of *Service Tasmania Online*. The RDS system has developed web based data entry forms with data validation processes and controlled vocabularies to ensure consistent metadata creation. Wherever possible, data entry into field elements is controlled against authority lists, and the topic field carries validation against an online version of the thesaurus. Librarians from the State Library of Tasmania have been contracted to indexing *Service Tasmania Online* resources and to ensure that the metadata is consistent and accurate.

Quality control also extends to routine batch processes that verify the existence of the resources described in the metadata repository. These processes remove the resource from the results display if that resource is no longer available.

At present approximately 3000 metadata records have been created, representing less than 5% of the estimated online resource available across government. However these 3000 records represent the demand points for online services covering all three levels of government in Tasmania, and provide comprehensive and quality results for Tasmania users. The user is presented with the best resource known that addresses the query, although at times this may be an entry point to a fuller range of detailed resources provided on a government agency Web site.

## **Technical components and standards**

A summary of the RDS' technical components and standards are outlined below.

- Data storage  
data files kept in XML format.
- Data input  
Web-based entry screens, Version 4 browsers with javascript capability.
- Data quality management  
data input verification as specified in the Australian Government Locator Service (AGLS), Education Network Australia (EdNA), and internal standards.
- Search Engine  
SiteServer from Microsoft
- Web site presentation  
HTML based graphics for Web site presentation.
- Web site navigation  
mixture of ASP controls using SiteServer display syntax controls.
- Access by other services and government jurisdictions  
HTML harvestable lists and HTML visible AGLS metadata.
- Hardware  
Microsoft NT Operating System.  
Two tandem servers linked for redundancy.

## **Current results and future directions**

The initial useability testing of the *Service Tasmania Online* web site yielded good results in terms of the ability of users to retrieve nominated government content. The web site went live in late April 1999 ([www.service.tas.gov.au](http://www.service.tas.gov.au)) and has been well received by both users and government.

Preliminary basic statistics have demonstrated the value of providing a variety of access methods. Over the first four months of operation, the proportion of users adopting the various options on the *Service Tasmania Online* web site was:

- Freetext searching: 30%
- Subject/topic : 30%
- Task: 15%
- A-Z listings: 10%
- Target audience: 9%
- What's new, life events, other options: 6%

Specific software has recently been developed to track the actual pathways selected by users and the sites then visited. Together with face-to-face and focus group testing, it is hoped to develop empirical models of user behaviour over the next year to guide ongoing developments. This will allow a qualitative guide for improvements to the site structure and language, the topic thesaurus, and the free text indexing tools.

## **The role of libraries in online government**

Experience in Tasmania has shown that libraries and librarians can have a positive influence on the utilisation of retrieval technologies and web delivery processes within Government. Nowhere is there more potential than in the area of government metadata standards development and metadata creation.

### **Understanding metadata**

Librarians need to ensure that metadata applications are based on a complete conceptual model for retrieval management. Metadata schema devoted to resource discovery have so far only provided a definition of the elements or fields that constitute the metadata set, together with some rules for syntax and low-level data type definitions (eg date fields). What is still needed is the development of the three other key components to successful metadata implementation: data standards, data creation business rules, and sustainable but effective data creation processes.

### **Data standards**

The key government metadata schema (DC and AGLS) allow a great deal of flexibility in the actual data standards that can be utilised within component elements. This is particularly true of the fields for title, creator, contributor, and subject. The ability to share data in such fields requires common agreement on the formatting of the data and the type of vocabulary to be used. There is no point in searching for cats in a metadata schema that uses the term *felines*.

Libraries are well versed in the types of controlled vocabularies, thesauri, and authority listings that can bring order to this area. Librarians must meet the challenge of ensuring that such metadata standards are chosen, not on the basis of how easy they are to create, but on how effective they are in retrieval.

### **Data creation business rules**

The metadata business rules tell the metadata cataloguer how to apply the schema and data standards against real-world resources, and are necessary to ensure correct and consistent data values. Those inexperienced in the field assume that this is a simple exercise. As professionals, we know that the development of standards can be very complex and requires significant knowledge of the resource and considerable understanding of the retrieval process. Librarians have the capacity to help develop a unified and consistent approach to these business rules.

### **Metadata creation processes**

The data creation processes must be based on the integration of the defined schema, standards, and application rules. For discovery metadata, this is a complex interaction that requires an understanding of the resource to be described, the data standards employed within the elements, and the application rules. This interaction must be based on an actual understanding of the content of the resources to be described and an understanding of the structure, semantics and syntax of the controlled vocabularies or thesauri that are to be used.

It is important for librarians to ensure that government business managers appreciate the complexity of metadata and understand that it is counter-productive and wasteful to spend resources producing bad metadata. For example, effective web discovery cannot be provided on the basis of tools that automatically generate metadata.

### **Conclusion**

We can translate all of these requirements for web retrieval via metadata into a simple familiar concept: the need for good cataloguing. This means cataloguing based on appropriate and well-developed rules; cataloguing that is cost-effective and affordable; and cataloguing that is managed by professionals.

Experience at *Service Tasmania Online* has shown that libraries and librarians can adopt a key role in developing new services. Government does care that its content is available online and discoverable, and is prepared to invest in systems that ensure this happens. Metadata is widely seen as the key tool to achieve this, but it is up to librarians to ensure that the hype of metadata is matched by the reality.

The web can be an enormous opportunity for professional cataloguers, where they may yet inherit the earth (at least in a professional sense). But it is up to us to make sure they are librarians.

### **Endnotes**

Office for Government Online. (2000). **GovernmentOnline**. [Internet]. Canberra, OGO. Available from: < <http://www.ogo.gov.au> > [Accessed 30 October, 2000]

Rosenfeld, L. (2000) Making the Case for Information Architecture. **Bulletin of the American Society for Information Science**. [Internet]. Vol 26, No. 5, June/July 2000. Available from: < <http://www.asis.org/Conferences/Summit2000/rosenfeld/index.htm> > [Accessed 30 October 2000]