# Government Ontology and Thesaurus Construction: A Taiwanese Experience

Chao-chen Chen[1], Jian-hua Yeh[2], and Shun-hong Sie[3]

[1] Graduate Institute of Library and Information Studies,
National Taiwan Normal University
cc4073@cc.ntnu.edu.tw
[2] Depart of Computer and Information Science,
Aletheia University
au4290@email.au.edu.tw
[3] Department of Library and Information Science,
Fu Jen Catholic University
modify@ms37.hinet.net

**Abstract.** Due to the quantity and the diversity involved in e-government presentations and operations, traditional approaches to web site information management have been found to be rather inefficient in time and cost. Consequently, the necessity of establishing a government knowledge management system, so as to speed up information lookups, sharing, and linkups, naturally arises. Moreover, this knowledge management system would in turn enhance e-government effectiveness as it helps to store and transmit information, be it explicit or implicit in nature. The first step in creating this knowledge management system is to build up the government ontology and thesaurus. Upon the completion of the ontology and thesaurus needed, semantic searching can be conducted, which in turn kickstarts other mechanisms required for effective information management.

Our research team has been commissioned by the Executive Yuan of Taiwan to establish the draft of government ontology and thesaurus and to design a framework for multiple-layered information management systems upon which the ontology and thesaurus can be constructed. The goal of this paper is to present the government ontology and thesaurus which our research team has come up with as well as the related infrastructure and function of the multiple-layered information management system.

**Keywords:** government ontology; government thesaurus; ontology editor; semantic interoperability; knowledge management.

## 1 Introduction

The number of web sites of Taiwan government increased quickly in recent years, which made Taiwan a popular e-government country. Due to the quantity and the diversity involved in e-government presentations and operations, traditional approaches to web sites information management have been found to be rather inefficient in time and cost. Consequently, the desire to speed up information lookup, sharing, and linkup created a need to establish a government knowledge management

E.A. Fox et al. (Eds.): ICADL 2005, LNCS 3815, pp. 263 – 272, 2005.

system. Moreover, a knowledge management system would in turn enhance e-government effectiveness as it helps to store and transmit information, be it explicit or implicit in nature. The first step in creating this knowledge management system is building up the government ontology and thesaurus. Upon the completion of the ontology and thesaurus needed, semantic searching can begin to function properly, which in turn would kickstart mechanisms required for effective information management. Three major issues in creating the ontology and thesaurus were analyzed and are as follows:

1. Extend construction and processing level of government web site information
   Most government information is presented in the form of web pages, with one web page devoted to each topic. However, the high number of web pages complicates management of pages and maintenance of topic crosslinking. A better way to efficiently manage government information is to use a database for data storage, a high performance search engine for query processing, and a dynamic catalog system for subject browsing.
2. Enhance classification efficiency by using ontology and thesaurus information
   Automatic information classification plays an important role in both information retrieval systems and subject catalog systems. A high-quality web site should provide multiple ways for information retrieval and browsing while not requiring tedious manual data classification. Thus the automatic classification function is critical for government knowledge management. Currently, due to the limitations of automatic classification technology, it is common practice to use ontology with thesaurus information to do automatic classification work.
3. Define a semantic exchange standard for government information
   Ontology and thesaurus information are the bases for efficient information retrieval and knowledge management of government information. With well-formed subject terms and knowledge hierarchy, the government information can be processed and categorized into a systematic structure, thus becoming a useful knowledge and semantic exchange standard.

## 2 Related Work

Having a common semantic expression among government departments is becoming more and more important in recent years because of the increase in information exchange among departments. The construction of government ontology and thesaurus information becomes a critical mission in many countries. For example, the Portal Thesaurus Project of New Zealand government, which creates the New Zealand Government Locator Service (NZLGS) Thesaurus; the Australian Governments' Interactive Functions Thesaurus (AGIFT), which provides Australian Government Locator Service (AGLS) a standard thesaurus terms for metadata; others such as UK Pan-Government Thesaurus (PGT), Government of Canada (GoC) Core Subject Thesaurus (CST), ETB Thesaurus for European Schoolnet (EUN), and so forth. These projects have a common feature: to provide a united semantic expression for information exchange, making possible efficient information processing and retrieval.

# 3 The Proposed Method

The goal of this research is to create a management system for government ontology and thesaurus information. Besides the construction of the system, the government ontology and thesaurus are also specified in this research. The processing steps of this research are shown below:

1. Content analysis

   The content analysis process contains several major steps, including subject term extraction, synonym construction, thesaurus construction, and ontology creation. The subject term extraction process in this research contains both manual and automatic term extraction. The manual term extraction utilizes existing subject category, thesaurus, and related web sites as references to create a basic subject term set. The automatic term extraction applies phrase segmentation and statistical methods to related government web sites to generate candidate terms. After these terms are generated, the synonyms for these terms are created by using Google to search for related information. These terms generated from manual and automatic term extraction are then revised by domain experts to generate final versions of subject terms.
2. System construction

   To support ontology creation from subject term generated previously, the research team creates an ontology management system which contains both ontology and thesaurus maintenance features. This system is able to maintain and present government ontology along with related thesaurus information. The ontology created by this system can be further rendered into RDF-based ontology and XTM-based topic maps.

   The entire processing sequence is shown in Fig. 1.

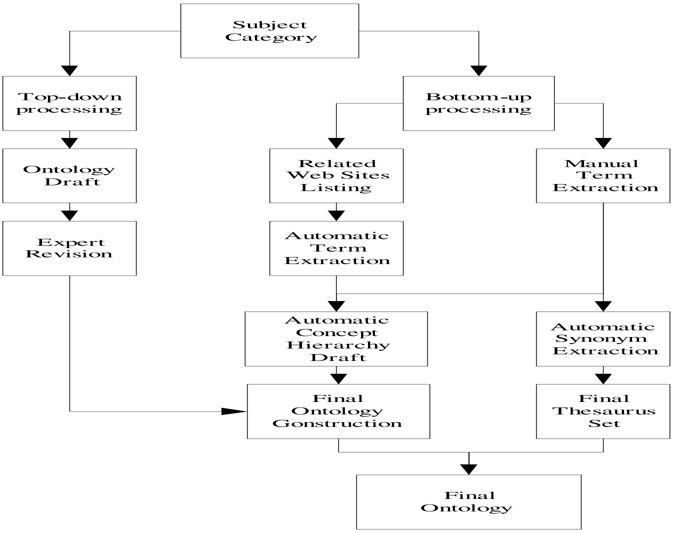


**Fig. 1.** The ontology processing steps

## 4 Creation of Government Ontology and Thesaurus

It is quite difficult to generate government ontologies and thesaurus information in a short time. But it is also difficult for government to refine these information without a draft. During the course of this research, we generated an ontology and thesaurus draft in a short time for the government and experts to refine. The achievements of this research are described in the following section.

1. Automatic term extraction

   The related information from various web pages was fetched for later use. A web robot fetched more than 1,000 related web sites and generated over six million possible term fragments. Table 1 shows the automatic term extraction result of this research.

**Table 1.** Automatic term extraction result

| Sites fetched | Possible term fragments | Useful term selected | Usefulness ratio |
|---|---|---|---|
| 1,107 | 6,851,165 | 40,531 | 0.592% |

1. Association build-up of terms

   This research adopts a statistical approach along with document feature to generate term associations. The associations suggested by this approach are for human references only.

2. Additional terms from existing thesaurus

   Discounting subject terms already found in the automatic term extraction process, the related thesaurus information is also useful to provide additional meaningful subject terms. Table 2 shows the statistics of subject terms from related thesaurus and subject catalog.

**Table 2.** The statistics of subject terms from related thesaurus and subject catalog

| Category | Authority term | Anonymous | Broader term | Narrower term | Related term | Scope note |
|---|---|---|---|---|---|---|
| Count | 16,073 | 984 | 1,401 | 2,210 | 6,389 | 404 |

3. Top-down generation of government ontology

   As mentioned earlier, this research uses top-down generation for government ontology drafts. Currently, there are 27 categories (ontologies) drafts and 8,135 sub-categories, and over 100 domain experts were involved in draft revision.

4. Association creation between government ontology and thesaurus information

The ontology and thesaurus are both tools for concept or subject presentation, but they traditionally are used separately. Since the features of ontology and thesaurus are in different layers, it is suitable to combine ontology and thesaurus to create more powerful concept representation. The ontology is able to express subject hierarchies, with each node contained in the hierarchy represents a single concept, but only one representation for one concept. The thesaurus contains a set of small term hierarchies, which expresses only one level of term relationship at a time, but the related terms and synonym shows the multiple representation possibilities of a concept. So it is clear that the ontology and thesaurus are complements in our scenario, as shown in Fig. 2.

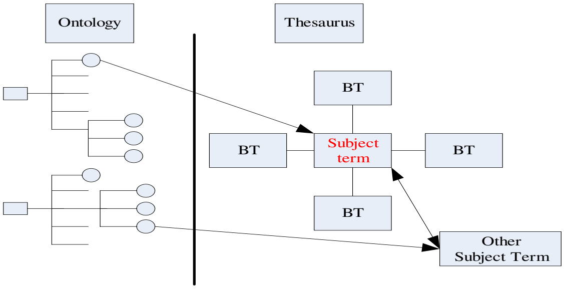


**Fig. 2.** The role and relationship of ontology and thesaurus in this research

Except for the relationships between ontology and thesaurus information, the concepts contained in ontologies can have associations also. In this research, we define two relationships: interlinks for concept associations across different ontologies, and intralinks for concept associations inside the same ontology. Fig. 3 shows the intralinks and interlinks in this research.

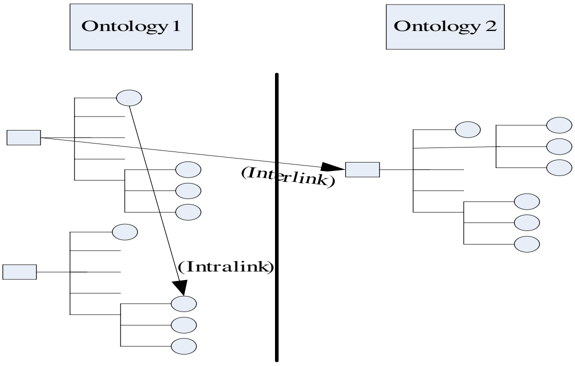


**Fig. 3.** Intralinks and interlinks

## 5. Ontology maintenance system

In order to manage the ontology and thesaurus that we constructed, we designed the X-ontology system. The service architecture of this research is shown in Fig. 4:

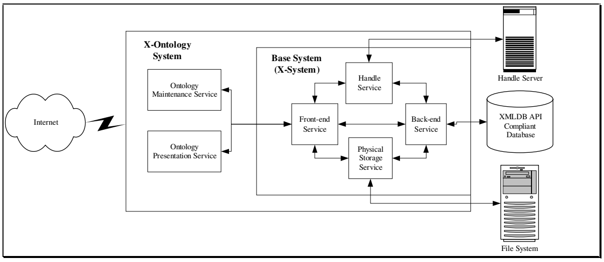


**Fig. 4.** The service architecture

In Fig. 4, the service architecture of X-Ontology system contains a base system called X-System (Yeh, 2003) and an ontology application extension (X-Ontology); the base system provides all digital archive content management functions while the X-Ontology is capable of processing knowledge network contents.

Since X-Ontology is built on top of the X-System, it certainly uses the functions of the base system to provide advanced ontology related features. In X-Ontology, the ontology structure can be stored to and retrieved from the base system. X-Ontology provides the user a working area for ontology maintenance, association maintenance, and additional thesaurus information. There is also a user interface for ontology content authoring in X-Ontology. These facilities will be discussed in the next section.

## 6. Ontology contents processing

As we mentioned earlier, the ontology contents created in X-Ontology are stored and retrieved through the base system, implying that the ontology contents are stored in native XML format. The ontology processing functions in X-Ontology support not only a single concept hierarchy, but also links information between ontologies. The result is that the ontologies created by X-Ontology contain not only information on a single concept but also associations between concepts. The processing functions are discussed as below:

### 6.1 Ontology hierarchy maintenance

The major component in X-Ontology is the ontology hierarchy maintenance service. The X-Ontology provides a tree-based hierarchy maintenance interface for the user to manipulate one concept hierarchy at a time. In addition, associations between hierarchies (ontologies) can also be maintained in this interface. Fig.5 shows the tree-based hierarchy maintenance applet. In the upper-right part of the applet shows the basic information form of the selected concept along with inter- and intralinks. The lower-right part of the applet maintains the additional information, which comes from the thesaurus saved in the system. These features will be discussed in the next paragraph.

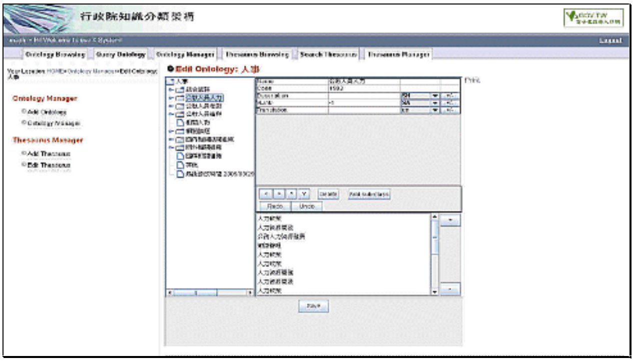


**Fig. 5.** The tree-based ontology maintenance interface in X-Ontology

6.2 Concept associations and additional information maintenance

The goal of concept associations in X-Ontology is to describe the relationships of two concepts "within" or "between" concept hierarchies. These "links" show a number of possibilities of semantics: it can be explained as a related relationship, or it can be treated as an "equal" relationship. However, the semantics of association is provided and used by the user, which will not affect the system implementation. Besides the association maintenance feature, X-Ontology also provides a way to enrich the concept created by the user: the thesaurus information. Fig.6 shows the maintenance interface, the lower-right part if the interface contains a list of additional information selected from thesaurus by user. The thesaurus contents can be picked from a query window (see Fig.7) and then attached to the current editing concept node, this feature gives the user a way to select additional information to enrich the target concept, which will be useful in future applications, such as document classification.

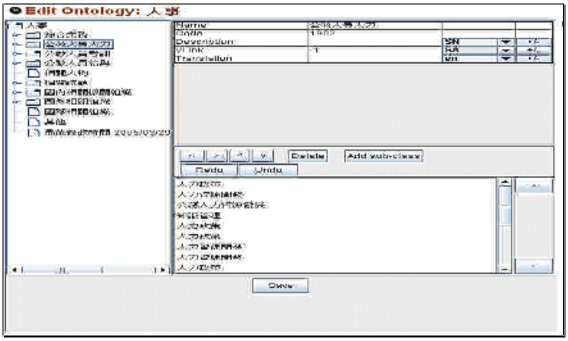
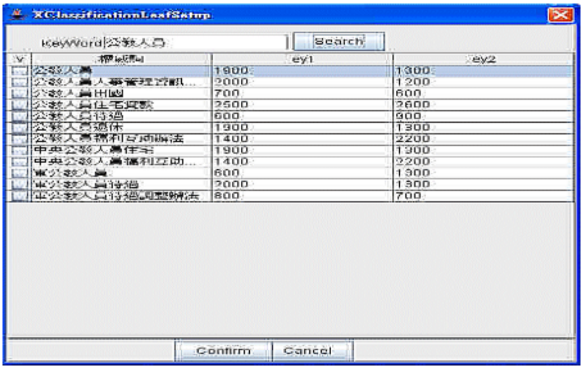


**Fig. 6.** The hierarchy maintenance interface with association and thesaurus information

6.3 Ontology presentation

The ontology presentation service in X-Onotology is quite straight-forward: information about the selected concept is presented. Fig. 8 shows a tree-based web



KeyWord 公教人員 Search

| V | [illegible] | ey1 | ey2 |
|---|---|---|---|
| ☐ | 公教人員 | 1900 | 1300 |
| ☐ | 公教人員人事管理資訊... | 2000 | 1200 |
| ☐ | 公教人員出國 | 700 | 600 |
| ☐ | 公教人員住宅貸款 | 2500 | 2600 |
| ☐ | 公教人員待遇 | 600 | 900 |
| ☐ | 公教人員退休 | 1900 | 1300 |
| ☐ | 公教人員福利互助辦法 | 1400 | 2200 |
| ☐ | 中央公教人員住宅 | 1900 | 1300 |
| ☐ | 中央公教人員福利互助... | 1400 | 2200 |
| ☐ | 軍公教人員 | 600 | 1300 |
| ☐ | 軍公教人員待遇 | 2000 | 1300 |
| ☐ | 軍公教人員待遇調整辦法 | 800 | 700 |

**Fig. 7.** Thesaurus query window for concept hierarchy maintenance interface

page for ontology presentation in X-Ontology. The ontology presentation uses a javascript-based component to perform tree presentation. When a concept is chosen to be displayed, additional information, including thesaurus and association information created by ontology maintenance interface, is also presented. The user can interactively traverse the hierarchy, check the thesaurus information, and jump to other concepts through the associations created by ontology maintenance interface.

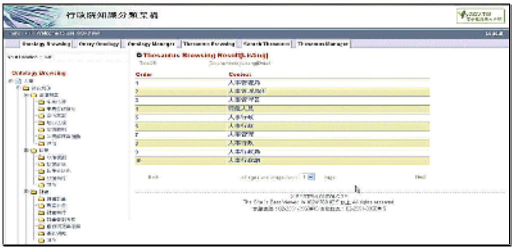


**Fig. 8.** Ontology presentation

### 6.4 Distributed ontology processing

Since ontology maintenance is conducted in a distributed manner, version control of ontology information is an important issue in our research. We introduced a dynamic resource locking mechanism to ensure that a sub-tree is edited by a single person at a time. Due to the stateless feature of HTTP protocol, the edit ontology action should assert an editing state which is recorded server-side. To avoid deadlocks, once the editing client idles for over 10 minutes, the assertion is dismissed and the lock no longer exists. The lock process flow is shown in the following Fig.9.

Another concern relates to authorization. In this research, different users are able to maintain different sub-tree structure in a specific ontology. Each node contained in an ontology contains read, write, and delete privileges, and these privileges are granted to different users for different maintenance goals. The ontology privilege sample is shown in Fig. 10.

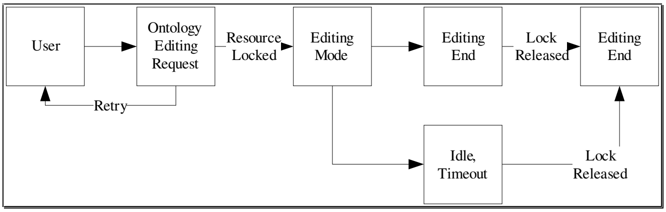


**Fig. 9.** Resource lock processing

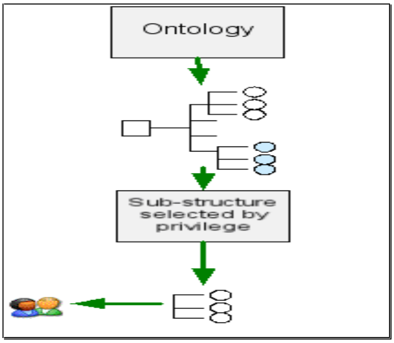


**Fig. 10.** Ontology sub-structure filtering by privilege

Besides the lock processing and sub-structure selection, our system also provides undo and redo actions to facilitate user maintenance efficiency. User activity logs are also kept for exceptional processing and error recovery. These features are all essential to asserting version control in this system.

## 5 Conclusion

This paper describes the experience of government ontology and thesaurus construction in Taiwan, and also describes the design of a framework for multiple-layered information management systems. This research produced not only a large amount of ontology and thesaurus content, but also tools which make use of the web to enable wide access and provide users with the ability to publish, browse, create, and edit government ontologies. Both automatic and manual term extraction for ontology creation are adopted, and our experience shows that the automatic process results in low extraction rate for qualified concept terms. In this research, we combine both concept hierarchy and thesaurus information to present a rich information ontology structure. This information will be useful in future applications such as document classification and portal subject browsing. In order to meet the challenges and needs of the e-government providers and users as well as to understand their demands and capabilities on dealing with government knowledge, our research team will continue to maintain Taiwan's government ontology and thesaurus contents as well as develop more supporting infrastructure.

## References

The New Zealand Government Locator Service (NZLGS) Thesaurus, http://www.e-government.govt.nz/docs/interim-thesaurus/index.html

The Australian Governments' Interactive Functions Thesaurus (AGIFT), http://www.naa gov.au/recordkeeping/gov_online/agift/summary.html

The UK Pan-Government Thesaurus (PGT), http://www.govtalk.gov.uk/documents/UK%20Metadata%20Framework%20v1%202001- 05.pdf

Government of Canada (GoC) Core Subject Thesaurus (CST), http://en.thesaurus.gc.ca/intro_e.html

ETB Thesaurus for European Schoolnet (EUN), http://www.eun.org/eun.org2/eun/en/etb/sub_area.cfm?sa=440&row=1

Yeh, Jian-hua and Chen, Chao-chen,(2003), The X-System: Design and Implementation of a Digital Archive System, Technical Report, Oct. 2003