The Use of XML to Express a Historical Knowledge Base

Katsuko T. Nakahira
Nagaoka University of Technology
1603-1 Kamitomiokamachi
Nagaoka, Niigata, Japan

katsuko@vos.nagaokaut.ac.jp

Masashi Matsui
Nagaoka University of Technology
1603-1 Kamitomiokamachi
Nagaoka, Niigata, Japan

065367@mis.nagaokaut.ac.jp

Yoshiki Mikami
Nagaoka University of Technology
1603-1 Kamitomiokamachi
Nagaoka, Niigata, Japan

mikami@kjs.nagaokaut.ac.jp

ABSTRACT

Since conventional historical records have been written assuming human readers, they are not well-suited for computers to collect and process automatically. If computers could understand descriptions in historical records and process them automatically, it would be easy to analyze them from different perspectives. In this paper, we review a number of existing frameworks used to describe historical events, and make a comparative assessment of these frameworks in terms of usability, based on "deep cases" of Fillmore's core grammar. Based on this assessment, we propose a new description framework, and have created a microformat vocabulary set suitable for that framework.

Categories and Subject Descriptors

H.2.3 [**Languages**]: Data description languages

General Terms

Experimentation

Keywords

historical knowledge representation, XML, RDF, microformats

1. INTRODUCTION

There are many records related to history available on the Web, such as scientific papers and databases dealing with history or chronological tables. The descriptions in these records are not designed with the aim of allowing computers to understand or collect data automatically. If information about historical events could be collected and converted into a reusable form automatically, history could be analyzed from a wider range of perspectives than is possible today.

There have been attempts to develop frameworks or languages to describe historical knowledge. In Japan, the Technical Committee on Electrical Technology History of the Institute of Electrical Engineers of Japan, created a database on the history of electrical power system technology[1], and developed the Historical Space Modeling Language (HSML) and a GUI called Mandala for browsing historical information written in HSML[2].

The Historical Event Markup and Linking (HEML) Project[3] is developing its own XML schema designed for historical description, and a Web application that displays chronological tables and does mapping in a variety of formats.

Table 1: Fillmore's deep cases [6]

Case	Description
Agentive	Role of the person who causes a certain action
Experiencer	Role of the person who expreiences a psychological phenomenon
Instrumental	Role that is a direct cause of an event or that stimulates a reaction in relation to a psychological phenomenon
Objective	Moving object or changing object. Or, role that expresses the content of a psychological phenomenon, such as judgement or imagination
Source	The starting point from which the object moves, and the role that expresses the original state or shape when the state or shape changes
Goal	The goal that the object reaches, and the final state or result when the state or shape changes
Locative	Role of expressing the location when an event occures
Time	Role of expressing the time when an event occurs

The Dublin Core Metadata Initiative[4] has developed a framework for writing metadata for published documents. In the field of historical documents, there is an attempt[5] by the National Museum of Japanese History to map items in a history research database onto the vocabulary of Dublin Core. In this way, there have been a variety of attempts to describe history or historical events available on the Web.

Although a variety of languages and formats have been developed for historical description as mentioned above, there are few that can handle causal relations. Among those mentioned above, only the set of HSML and Mandala can handle them. However, since HSML cannot be used without an HSML interpreter, it is not as readily usable as the Resource Description Framework (RDF), which is now widely used. This paper describes an attempt to develop a generic description format.

2. ITEMS OF DESCRIPTION TO BE STUDIED

This paper uses deep cases in Fillmore's case grammar as our underlying framework for studying the items of description suitable for historical events. A deep case indicates the role an individual word plays vis-a-vis the verb. Any historical event can be associated at least with a subject and an action it takes. Therefore, we consider that the framework based on the concept of deep cases provides the widest range of description items regarding an event description. The eight cases given by Fillmore's deep case concept are shown in Table 1.

When the eight cases are applied to the description of his-

Table 2: Proposed microformat vocabulary set

Classification	Case	Element	Attribute	Value	Item	Location	Data type
Event	Objective	Not specified	class	event	event name	title attribute	character string
Time	Time	Not specified	class	start_date	Starting date name	within element	character string
					Starting date name	title attribute	date
		Not specified	class	end_date	ending date	within element	character string
					ending date	title attribute	date
Location	Location	Free	class	location	location name	within element	character string
					ending dat	title attribute	date
Participant	Agentive	a	rel	participant	participant's name	title attribute	character string
					participant's URL	href attribute	URL
Evidence		a	rel	evidence	evedence name	title attribute	character string
					evidence URL	href attribute	URL
Cause	Instrumental	a	rel	cause	cause name	title attribute	character string
					cause URL	href attribute	URL

torical events, they can be described as follows. The agentive case refers to the entity that caused an event. This entity may be an individual or an organization. The instrumental case corresponds to the cause of an event that occurred. The objective case applies to the object to which an event occurred, but its nature depends on the nature of the event. For example, when a new technology is being developed, the technology is the objective case. The location case corresponds to the place where an event occurred while the time case indicates the time when an event occurred. Since an event can be said to be expressed in direct discourse, the experiencer case does not exist. It is more difficult to interpret what the source case and the goal case represent. Unlike time or place, whether there are starting and goal points depends greatly on the nature of event (verb). Therefore, in this paper, we assume that no roles exist that correspond to the source or goal case for an event.

Consequently, we consider that a historical event may be defined by five elements: person, cause, object, location and time.

3. CONSIDERATION OF THE DESCRIPTION METHOD

We use microformats[7] for the description of metadata within an item of content. Microformats give a certain name to an attribute of the XHTML text to enable metadata to be extracted from a document. Microformats make use of the attribute values provided in the XHTML grammar. An advantage is that, even when our approach is applied to the existing content, almost no change to its structure is required. This advantage led us to choose to use microformats because we are considering the application of our approach to existing descriptions of historical events.

Based on Table 1, we have created the microformat vocabulary shown in Table 2. In parallel with this effort, we have created a microformat XSLT that extracts metadata and converts it to a Resource Definition Format (RDF).

4. PROTOTYPING

We have marked up documents on the history of technology using the microformat vocabulary mentioned above, and displayed marked-up events on a SIMILE Timeline. Specifically, we have marked up "the Development of Electronic Computers in the World in their Early Days" and "the Development of Electronic Computers in Japan in their Early Days", both reports published by the Institute of Electrical Engineers of Japan.

First, we marked up the documents using the defined vocabulary. Then, using the XSLT associated with these documents, we extracted events from the documents, and obtained RDF data that describes the events. Finally, we converted the RDF data into the data format of the displaying application used (Timeline in our case), and displayed the data. Timeline[8] is an Ajaxy Widget for displaying a timeline, developed by the SIMILE Project. The data of Timeline is written in XML using its own vocabulary. Therefore, if we develop an

XSLT that directly converts microformats into the Timeline format, it would be possible to derive Timeline data from the documents. We did not adopt this approach and instead converted microformats to RDF formats and then into the Timeline format. This is because we did not want to tune our method to a specific display format, since we were seeking to allow a variety of display formats.

This prototyping has confirmed that the extracted events can indeed be mapped on the timeline.

5. CONCLUSION

In order to create a framework for describing historical events, we have developed a unique description framework based on the deep cases in Fillmore's case grammar. We have created microformat vocabulary suitable for this framework. Our future studies include the development of a tool that displays, in an intuitional manner, events marked up using the above vocabulary.

6. REFERENCES

[1] *Technical Committee on Electrical Technology History*, Institute of Electrical Engineers of Japan Technical Report, **991**, pp. 9-11(2004).

[2] Y. Matsumoto, A. Yamada, *An association-based management of reusable software components*, Annals of Software Engineering, **5**, pp. 317-347(1998).

[3] B. G. Robertson, *Visualizing an Historical Semantic Web with HEML*, Proceedings of the 15th International Conference on WWW, pp. 1051-1052(2006).

[4] *Dublin Core Metadata Initiative[Online]*, Available at: http://purl.oclc.org/dc/ (January 2007)

[5] Fumio Adachi, *Study on Mapping of Historical Research Databases to Dublin Core Metadata*, SIG Technical Report, **2006**(112), pp. 15-23(2006).

[6] Makoto Nagano, *Natural language processing*, Iwanami Shoten Publishers (2005)

[7] Rohit Khare, Tantek Celik, *Microformats: a Pragmatic Path to the Semantic Web*, Proceedings of the 15th International Conference on WWW, pp. 865-866(2006)..

[8] *SIMILE timeline. [Online]*, Available at: http://simile.mit.edu/timeline/ (January 2007).