Internship University Library Groningen

Internship Report

Junte Zhang

juntezhang@gmail.com

Abstract: I have worked for the SWHi (Semantic Web for History) project. It is a research project about the Semantic Web at the University Library of the University of Groningen. In this report, I will describe the objectives of my internship, my work for the SWHi project, other experiences and the results obtained.

dr. Gosse Bouma
Supervisor Information Science

dr. Henk Ellermann
Supervisor University Library

Information Science

University of Groningen

May 2006

Contents

1 Introduction

The SWHi (Semantic Web for History) project is aimed at integrating, combining, and deducing information about early American history, based on the Early American Imprints, Series I: Evans, 1639-1800 collection, to assist general users or historians in exploring American history by using new technology offered by the Semantic Web. Although this project is initially limited to the historical domain, in the future its results can be applied to other domains and more heterogeneous resources.

This project is a cooperation between the Digital Library department of the University Library of Groningen and the Information Science department of the University of Groningen. The duration of this project is two years, from January 15, 2006 until January 15, 2008. The work is divided into three phases: identification, development, and releases. After each release, improvement cycles will be conducted.

I have worked in the identification and development phases.In this report I will explain my internship, the experiences I have gained, and a few things more.

1.1 Who?

For this internship, I have worked most closely with the following two persons.

1.1.1 Ismail Fahmi

Ismail is the main developer of the SWHi project. He has created the architecture of this project, e.g. the used tools and software. He is also a Phd student at the Alfa-Informatica department of the University of Groningen. He works for the University Library on Wednesday and Thursday, and sometimes also on Friday. I see him during these days. I have worked most closely with him on the SWHi project.

1.1.2 Henk Ellermann

Henk Ellermann is the head of the Digital Facilities department of the University Library Groningen. He has a background in Cognitive Psychology,

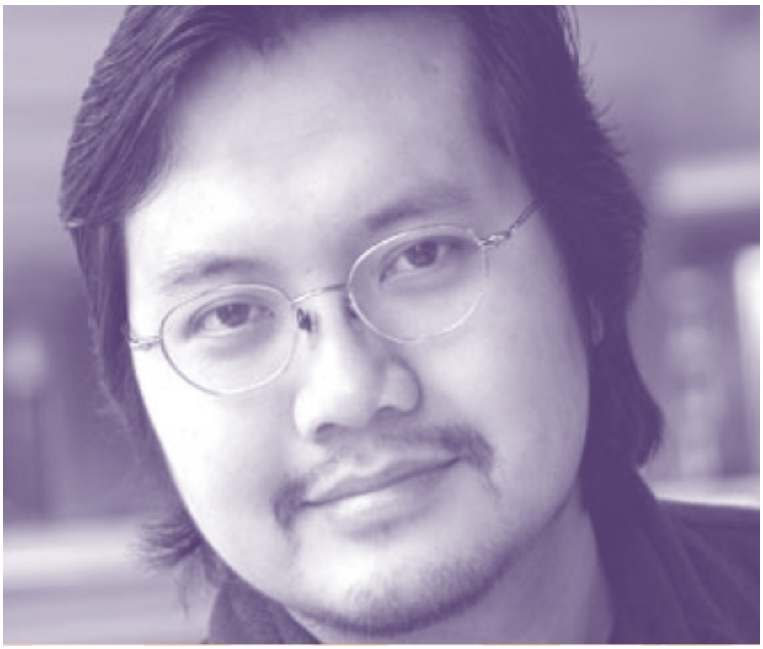

Figure 1: Ismail Fahmi

and at the same time has developed out of necessity ICT skills, such as programming.[1] As head of this department, he not only wishes to run projects, but also to be innovative, for example, that documents can be saved in a more structured format, so refined search possibilities are possible, and that relationships between documents can be presented on the Web. As an internship student, I hope to contribute partly to that vision.

1.2 Where?

The University of Groningen has a University Library (UB) and about twenty institutional libraries. These libraries jointly present their resources. There is one catalogue. Moreover, numerous electronic files are offered university-wide by means of network. The internship is at the Digital facilities department of University Library in Groningen. This department is responsible for unlocking and presenting all relevant digital information. I believe it is the perfect place to do my internship, and there is a huge synergy possible between Information Science and Digital Libraries. I believe I can fully use the expertise I have gained during my study and

[1] See the interview in Pictogram, February/March 2005

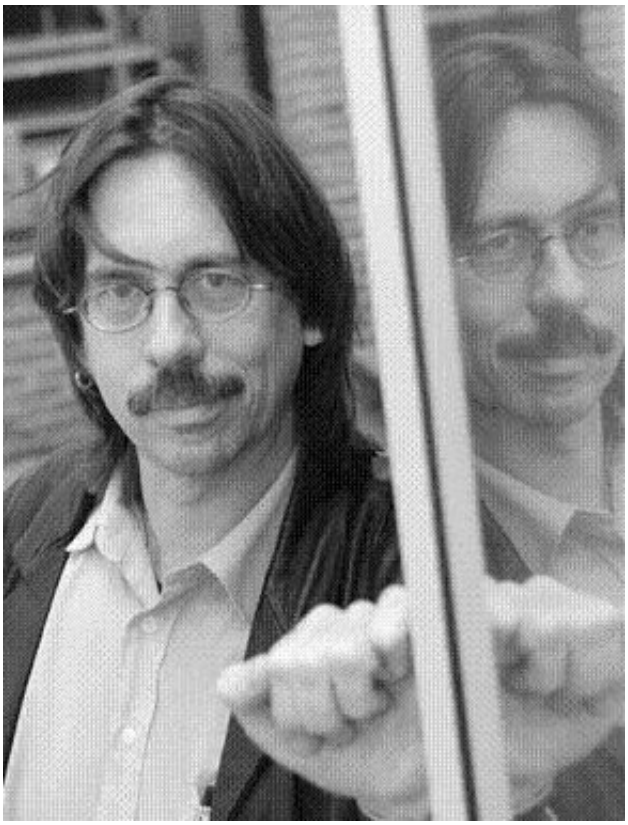

Figure 2: Henk Ellermann

deploy it for Digital Libraries.

1.3 What?

For the SWHi project, the purpose of my internship was to find out what the possibilities are of Semantic Web technologies for Digital Libraries. For this, I have developed an historical ontology and populated it with information obtained from the aforementioned Evans database. The information consists of metadata that is part of the Evans bibliography. The SWHi project is aimed at integrating, combining, and deducing information on the early American history, based on the Early American Imprints, Series I: Evans, 1639-1800 collection, to assist general users or historians in exploring American history by using new technology offered by the Semantic Web.

Well, what is the Semantic Web? Tim Berners-Lee, the ”inventor” of the

World Wide Web, always had more in mind than the current World Wide Web. Basically, the Semantic Web is making all data on the Web accessible as a huge database. The Semantic Web architecture is a system of various logic languages and relational data, connected together with local consistency, and its building blocks are binary relations and monotonic logic. It is implemented using a common symbol space (URIs). See also Berners-Lee et al. (2001)

1.4 Why?

The intership is part of the curriculum of the MA program of Information Science in Groningen. This work is relevant and interesting. Institutions which have lots of digital, computer-processable, data available, have yet to use and deploy the possibilities that Semantic Web technologies have to offer. In research communities, lots of research has been conducted. However, it is now time to really use and deploy. Within a decade, we will be using the promising features of the Semantic Web, as we are using the World Wide Web today.[2] You can be in the front, or lag behind as a snail.

What I hope that I have achieved is:

- Obtain experience in Research & Development (R&D)
- Learn more about my research interests, e.g. Semantic Web
- Use the knowledge and results to write my MA thesis

2 Working Environment and Time

What to say about my working environment? My work place was situated at the Cataloging Department of the University Library. I had a desk with a computer. The workers do not talk much, generally speaking. The desk was big, so there was enough room to put my documents. As my work was progressing, the computer I was using was not powerful enough. Everything was lagging on the screen. At my request, the memory of the PC got expanded to 2 Gigabyte, which was a relief.

The working hours are very flexbible. I have not counted the numbers of hours I have worked, but it is certainly fulltime. You can create your

[2]For example, see this BBC News report: http://news.bbc.co.uk/1/hi/technology/5013146.stm

own schedule. I usually started a few hours later, and also worked a few hours longer. There are three breaks in a day, as Ismail told me. It was at 10:00 AM, at noon and at 3:00 PM. In the cafetaria, there was a special section for personell. You could get free coffee, tea or chocolate milk. At my workplace, I was told by an elderly lady, that I am not allowed to drink. She almost scolded me. I thought that was strange and I was very disappointed that she was speaking so disrespectful to me. I did not know her, and I believe, neither did she know me. However, generally speaking, I was treated with respect, especially by Henk and Ismail.

You need to self-disciplined and able to work on your own. There is not much supervision, which I regarded as an advantage, because you are free to develop your own ideas. I weekly saw Ismail on Wednesday and Thursday, sometimes also on Friday. We had a weekly meeting on Thursday. Sometimes, Gosse, my Information Science supervisor, was also present at these meetings. We have discussed our weekly results and talked about how to proceed next. This happened almost every week, well, at least for the first few months.

3 My work

In this section, I will explain what my I have done and experienced during my internship period from February to June.

3.1 Preparation

The first week of the internship, it was mainly getting used to the working environment. I have setup the computer to run Linux, as I cannot work in Windows. I have studied some literature to get an ideas on how to proceed with the project. The task was to develop an ontology which has American history as its domain. Anything else was not really clear yet, so that void has to be filled up first. I kept notes of my progress on a blog, which is a journal where you can post messages ad hoc. During these months, I have posted about 50 posts. My idea was to at least post a weekly update, but of course I have posted much more.[3]

The literature I have studied in the first week was mainly about (*i*) the data I was going to use (*ii*) how ontologies are build (*iii*) how Semantic

[3]See `http://semweb.weblog.ub.rug.nl/home`

Web technologies are now used for History (*iv*) general theories about the Semantic Web.

We have to collaborate on software development, and that is why Ismail had setup CVS. This is a way how research groups collaborate on a joint project. You store the latest files on a central computer. You update the work you have did by committing it to that server and you can get the latest update by checking it out there. This was new to me. I have used CVS to daily update my work and collaborate with Ismail.

3.2 Tasks

For the real technical work and more details, I'd like to refer to my thesis. In this section, I will explain in general terms what I have done, which can hopefully be understood by laymen.

I have studied the metadata I was given. It was MARC21 metadata in XML format, which is highly structured data using a common bibliographic schema called MARC21. That file was about 150 MB big. My plan was to automatically extract information from that data by using scripts that I have written. However, to do that, I first need to know what the data looks like and what information is there. This way, the ontology can also be modeled. That data are huge, but there is a way to cope with that.

3.2.1 Know the data

I have written a script that counts what is in metadata using a script. Since this data is highly structured, I can match patterns and use regular expressions. This way I have counted the number of records, data- and subfields, combination between these two, and their meaning. For example, in MARC21 you have the code *100 a*. The following table shows that it is a personal name, occurred 17577 times in the metadata or 1.07% out of the total number of codes.

MARC21	Number	Percent	Description
100 a	17577	1.07	MAIN ENTRY. Personal name

3.2.2 Setup the historical ontology

My task was to put all that information into an ontology. However, how should that ontology look like? Since the point of the Semantic Web is to reuse existing resources as much as possible, we have decided to do the same. We have used an existing out-of-the-box historical ontology. This saved us time and effort to develop a similar ontology. That existing ontology was not complete enough for us, so I have also mapped an existing Topic Hierarchy, which is compiled by librarians, and added it. Furthermore, the ontology was expanded by automatically generating the types of documents out of the metadata and including these as concepts. The ontology was setup and designed with the tool Protege, which is a popular Ontology Editor and Knowledge Acquisition System.

Then it was my task to map the extracted information to the ontology. The extracted information is used to generate instances, and an instance consists of properties. So the information that is extracted out of the metadata was used to compose the values of properties and to classify them to the proper concept. This is all done automatically using scripts that I have written. I have found out that the existing out-of-the-box ontology was not good enough, so I have included more schemas, in particular Dublin Core and FOAF.

3.2.3 Populate the ontology

The ontology was then automatically populated, which means feeding instances to it. For example, all codes *100 a* are names of persons, which is mapped to the property *foaf:name* of the instances of the concept *Person*. This happened automatically with literally thousands of values, and with many more properties. For more technical information, I have to refer to my thesis.

MARC21	**Description**	**Schema**
100 a	MAIN ENTRY. Personal name	foaf:name

Following these steps the ontology became ready for use. But is it good enough, e.g. what is its power? In the future, this ontology can and will be enriched with information from more sources, such as Wikipedia and other digital encyclopedia. Facts and relations between these facts can be recovered, but to what extend? This question can be answered by querying the ontology, which is my way to evaluate the ontology.

3.2.4 Evaluate with Questions

Event	TimeEvent
"Anglo-French War, 1755-1763"	"1755-1763"
"Anglo-French War, 1793-1802"	"1793-1802"
"Anglo-Spanish War, 1739-1748"	"1739-1748"
"Augustus II 1697-1733"	"1697-1733"
"Austrian Succession, War of, 1740-1748"	"1740-1748"
"Canadian Invasion, 1775-1776"	"1775-1776"
"Catherine II 1762-1796"	"1762-1796"
"Clark's Expedition to the Illinois, 1778-1779"	"1778-1779"
"Colonial period ca 1600-1775"	"1600-1775"
"Confederation 1783-1789"	"1783-1789"
"Constitutional period 1789-1809"	"1789-1809"
"Consulate and First Empire 1799-1815"	"1799-1815"
"English West Indian Expedition, 1739-1742"	"1739-1742"
"English colony 1763-1784"	"1763-1784"
"French and Indian War 1755-1763"	"1755-1763"
"French occupation 1798-1801"	"1798-1801"
"Fries Rebellion, 1798-1799"	"1798-1799"
"George II 1727-1760"	"1727-1760"
"George III 1760-1820"	"1760-1820"
"Helvetic Republic 1798-1803"	"1798-1803"
"Jacobite Rebellion, 1745-1746"	"1745-1746"
"King Georges War 1744-1748"	"1744-1748"

Figure 3: Results of some events that happened in the life of George Washington

This is done with Sesame, which is an RDF framework to store repositories, such as this ontology. Using the query language SeRQL, you can query that RDF repository.

So I have learned:

- how to setup Sesame

- how to use query language SeRQL

The results look promising. See figure 3. To see the types of questions, you can check out this site: `http://evans.ub.rug.nl/~junte/query.html`

3.2.5 Deploy it

Figure 4: SWHi web interface

Eventually, these questions can really be deployed. This is done using a PHP application, which was engineered by Ismail. I have helped deploying the queries in this application in PHP. See figure 4 for the web interface of that application. For example, it will show when someone was born or died, who he knew, what events occurred in his lifetime, his publications, etc. This shows that the ontology can already be regarded as a nice knowledge base about American history between 1600 and 1800. Some relevant questions can be answered. More uses are possible, for instance by using the ontology for query expansion, or by offering suggestions to users using web agents, or allow Semantic Browsing by automatically creating semantic links within documents. In my thesis, this will be elaborated.

4 Conclusion

I have met the three objectives I have set in the beginning, which are:

- Obtain experience in Research & Development (R&D)
- Learn more about my research interests, e.g. Semantic Web
- Use the knowledge and results to write my MA thesis

I have now some experience in R&D work, and have learned more things about the Semantic Web, especially in practical terms, and I have gathered enough information and expertise to write my MA thesis. There was enough flexibility and freedom to do my job. I have had a good time, and it was rich experience.

References

Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):34–44, 2001.