Semantic Squirrels

Hugh Glaser

Electronics & Computer Science
University of Southampton
Southampton SO17 1BJ
+44 (0)23 8059 3670
hg@ecs.soton.ac.uk

Abstract. We argue that data should be acquired now. Every day that goes by data is lost. We propose Semantic Squirrels, a community-enabled low technology solution to data acquisition to achieve this data acquisition, while other more difficult problems wait to be resolved.

Introduction

The most exciting thing about the Web is the data. The most frustrating thing about the Web is the lack of data.

In browsing the Web, it can often seem like the world began around 1990, although much of the most exciting content, such as census data and government records comes from earlier times. Where the data is not available in electronic form or at all, it can be created from paper records or from collective memory if there are sufficient cooperative individuals. Friends Reunited is the clear example of the latter. Unfortunately recreating this data is an expensive, patchy and error-prone activity.

It is reasonable to suggest that most of the people who recorded the data we now search, aggregate and retrieve on the Web had little idea of the use to which we are now able to put it

Turning to the Semantic Web, it is also reasonable to suggest that the same thing will happen again. We struggle as best we can to imagine the developments of the next few years, and what sort of systems will emerge. However, we do not know.

What I do know is that when the brave new world of what the Web and Semantic Web becomes finally arrives, I will regret my lack of foresight in gathering data for it to feed on. I do not intend to have those regrets.

So what are the major barriers to gathering this data? There are a number of questions:

- Where can it be found?
- What format should it be stored in?
- What should be kept, and how should it be structured (what is the ontology)?
- Where should it be kept?
- Who might have access to it?

It is our contention that the actual gathering of data only has one real barrier – it needs to be found.

Collecting Data

As we go about our daily lives, it is evident that we leave electronic footprints, either deliberately or incidentally. Much of this can easily be recorded on the machines we use in our offices, homes, and that we carry around with us: our desktop machines, laptops, PDAs and the device formerly known as the mobile phone.

Some of the data is quite obvious: many of us keep our photographs online, maintain an email archive, and add extra data such as GPS tracks. In fact applications such as Jet Photo Studio (http://www.jetphotosoft.com/web/) that combine photographs with GPS data and maps to present the photographs on a map against a calendar begin to give a sense of what might be achieved if more data can be acquired.

For example I can currently use my laptop to gather:

Photographs; GPS data (I carry a GPS receiver at all times); All email, both in and out, in both Mail and Entourage formats; The MAC address of my current wifi access point; My IP address (after decoding any NATing); All files changed today; Current weather conditions; Address book; Diary; iTunes library listing; Safari (web browser) Bookmarks and History; Everything my laptop hears (but not always); what the camera sees once a minute.

Almost all of this happens completely incidentally, and has been for many months. There are many other sources that could be added to this.

This further data is even harder to gather and therein lies the problem. It is often "owned" by the application or operating system. For a person with the right expertise it is not hard to get, but such people may not have skills in the Semantic Web technologies. Similarly, people with Semantic Web skills may not have the skills to grasp the data.

So a major objective of the Semantic Squirrel activity is to separate out the functions of different experts. The Squirrels themselves are scripts and code which extract the data from wherever it is secreted, and put it away (in a Larder) for future processing.

We have set up a Semantic Squirrel web site that aims to provide the focus for the activity (http://semantic-squirrel.org/). It is organized as a wiki, so that all interested parties can contribute, and it is expected that the structure of the site and the nature of Squirrels will evolve as time goes by.

Users are invited to contribute Squirrels, as well as the scripts that take the data that has been squirreled away and turn it into more directly useful forms such as RDF against specific ontologies. Utilities to help to manage squirrels are also needed.

In conclusion, we believe that a collaborative activity of this sort has great potential, and I have a personal and selfish interest; I want to get people to write Squirrels for me, especially of the "difficult" sort described above. So I hope that the reader feels inspired to get involved; visit http://semantic-squirrel.org/ to do so.

I thank my colleagues, the EPSRC AKT Project (GR/N15764/01) and the EU ReSIST Project (IST-4-025764-NOE) for their support.