

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.

Categories and Subject Descriptors

H.3.5 [Information Systems]: Online Information Systems – *Web-based services*. D.2.12 [Software]: Software Engineering – *Interoperability*.

General Terms

Design, Human Factors, Standardization.

Keywords

Semantic Web; Knowledge Acquisition.

1. 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.

2. GATHERING DATA

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.

Of course these issues are active research topics in a wide range of organizations, and there is currently a lot of interest.

A good starting point to explore the questions above is the Memories for Life project (<http://www.memoriesforlife.org/>), although this tends to emphasize the photographic issues. Research laboratories of major companies in computer manufacture or mobile technology are heavily involved in this area; the MyLifeBits project from Microsoft Research is a prime example

(<http://research.microsoft.com/barc/mediapresence/MyLifeBits.aspx>).

Much of the work on the Semantic Desktop also addresses these issues, since it is concerned with finding and processing data on personal computers. A recent workshop at the International Semantic Web Conference 2005 in Galway provides an overview of this activity (The Semantic Desktop – Next Generation Information Management & Collaboration Infrastructure).

In the next section we look at the data we would like to collect, and motivate Semantic Squirrels. Following that we look at some typical Squirrels, and then provide some concluding remarks.

3. 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 mobile phones.

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)

Almost all of this happens completely incidentally, and has been for many months. There are many other sources that could be added to this. It would be easy to record everything my laptop sees but there are more difficult sources. I use a computer as a Personal Video Recorder (PVR), so it would be possible to record which programmes I watch; my mobile phone records the numbers of outgoing and incoming calls, as well as text messages; I could log the application with which I am currently interacting. In the near future, much of what I buy will come with RFID tags, which will provide a rich source.

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 proposed Semantic Squirrel activity is to separate out the functions of different experts.

We must tap into the expertise of the people who know how to get the data, without imposing an overhead on them to learn Semantic Web or related technologies. They should be able to simply build Squirrels that grasp these Nuts from the applications and operating systems and squirrel them away in a Larder (usually a directory of folder on the local machine).

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.

4. EXAMPLE SQUIRRELS

Squirrels can of course be as complex as the author likes, but the objective is that they should be very quick for a knowledgeable person to write, so that the barrier to recording the data is as low as possible.

Looking at some Squirrels will clarify things, while exposing some of the many issues. The examples are taken from MacOS, but are similar to that which would run on any Unix box or DOS shell.

First we look at web browser data:

```
cp /Users/hg/Library/Safari/History.plist /Larder/Today/
plutil -convert xml1 -- /Larder/Today/*.plist
```

This simple command picks up the browser history and puts it in the Larder (where all the Squirrels put their Nuts). It already breaks the rules slightly, as it post-processes the data, however we

consider it desirable to convert data to text, in this case plist XML, rather than leave it in the alternative binary plist format.

Here is how we squirrel mail messages:

```
mv /Users/hg/Library/Mail/Mailboxes/Today.mbox/Messages/*
/Larder/Today/Mail
```

This moves all the messages from a mail folder called Today in the Mail application. This introduces the idea that there may be components in applications that work in tandem with external scripts.

Finally a script to gather changed files:

```
cd /Users/hg/My\ Documents/
find . -mtime -1 -exec cp {} /Larder/Today/Files \;
```

In addition to the Squirrels themselves, there is some infrastructure that is required to make them function. This can conveniently be done using a shell script that is fired by crontab jobs. Primitive versions of this infrastructure and some other Squirrels can be found in the Developer section of the Semantic Squirrel web site. It will be necessary to have more sophisticated tools for managing and installing Squirrels. We envisage a system comparable with rpm or apt-get, or the graphical fink-commander, which would allow users to choose the Squirrels they want, and then have them installed in the appropriate format for their application and operating system.

5. CONCLUSIONS

We have put forward an argument for separating the gathering of data from the processing. We have suggested that there is much data that could be gathered, and that this should be put aside during this Semantic Summer, so that we will be able to store it through the Semantic Winter and feed on it when the Semantic Spring technologies really start to blossom. However it is interesting to speculate what might be done with some of this data.

The most obvious is to restore context using many of the cues we find so valuable. For example before I go to a social or work meeting, it should be possible to give me a sense of the last related event, perhaps playing back the closing words, showing me emails and websites I had looked at around that time, play the music I listened to on the way, and remind me there was a thunderstorm.

It will also be interesting to see how the symbiosis grows between the user and the data stored, as retrieval gets better. I am already less careful in filing email messages knowing that I can easily retrieve them from the Larder (I use Spotlight for this). I no longer record my business mileage at the time, as I simply look it up in the GPS data in my Larder.

It will be clear to some readers that much of the data being gathered is similar to that needed by systems in Ubiquitous and Pervasive computing. We expect that Nuts will enable systems to track behaviour over extended periods of time to understand changes, and support activities. It is also the case that we are already using Nuts as research data to explore the possibilities of such systems.

In the space of this document we have had to leave many issues without discussion. The biggest of these is to do with issues of how the nuts might be processed. It is clear that for the present it is possible to process them into resources such as knowledge bases using scripts. However, might it not be possible that

knowledge acquisition tools in the future could just process the text and extract the knowledge with little programmatic input?

Another serious question is that of privacy. Many people would consider much of the data gathered to be very private. We have therefore assumed that no data is communicated to the outside world, although it may be that some data is acquired from outside, such as news feeds or weather. It is interesting to speculate what Semantic Web technologies can do if the user refuses to allow personal data to be shared; it seems that the ability to mediate between the many resources that a single individual has is still a powerful tool on a single machine.

Storage space is another consideration. With the exception of multimedia, it is quite hard to generate excessive data in comparison with modern disk size. The data listed above, without the multimedia, averages less than 10 MB a day. This is less than

3 GB a year. Even with pictures it is still only a few GB. For extensive sound and quality video, however, some special care needs to be taken.

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.

6. ACKNOWLEDGMENTS

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