



A Non-Technical
Guide to Muninn
for Historians

Nick Gunz

2009-05-30, v1.2

www.muninn-project.org

A Non-Technical Guide to Muninn for Historians

Nick Gunz, 2009-05-30, v1.2

The Muninn project is made up of academics from the UK, the US and Canada, in disciplines ranging from biostatistics to literature. We have a project, for which we are currently trying to get funding, to take millions of documents from the First World War and turn them into a huge research database.

This database will be very useful for you, the historians of the First World War. Those of you who work in very data driven areas, such as social or demographic history, will be given a vast new repository of raw and semi-processed information. Those of you who work in more qualitative specialities will be given a tool which allow you to set your records in their proper context with incredible rapidity and scope.

But we need your help to make this work, for two reasons:

1. Our project relies on the idea that we will be producing information that is useful to the wider research community. We need to be able to consult with you while we're still creating that big database so that we can be sure to collect the right information and format it in the right way.
2. Nobody is going to give us money to do this research unless we can prove to them that people like you want this database to exist. We need to collect names of people who are prepared to support our funding applications by either endorsing our project as a Good Thing or, even better, by offering to take part in our research as an advisor, a full participant, or even a leader.

What We Hope to be Able to Do

Think of it like Facebook for the First World War. I'm not sure you're familiar with 'social networking' sites like Facebook (online at facebook.com), but they're a bit like the phone book only much, much smarter. In a phone book, you look up a person's name and you get their phone number. In a site like Facebook, you look up that person and you find photographs of them, lists of people that they're friends with, what kind of movies they like, when they were born, and all sorts of other random information that they choose to list there.

If that doesn't sound very impressive, imagine you had access to the whole Facebook database for, say, market research. In seconds you could find the average age of people who were fans of the movie *Paths of Glory*. Let's say you want to sell video games. You wouldn't have to advertise to people who like video games, you could get lists of *friends of people* who like video games and send them heavy hints in the week or so before their gamer buddy has a birthday. These social networking sites are revolutionising market research and advertising, and people are making millions.

So imagine if we could revolutionise history research in the same way. Imagine if you had a document, say the personnel file of a private soldier. With a few keystrokes you could get a list of all the other personnel files which relate to the one in your hand: who his close comrades were, who his immediate superiors were. You could get a map of where he went, when. If he was wounded you could find out how common that wound was. If he survived the war, and left a pension record, you could find out whether any of his neighbours were also veterans.

But why stop there? What was the average age of sergeants at the Battle of the Somme? Give me a density map of the locations of people from the Lincolnshire Wolds on 4 March 1917. Give me an organisation chart of the chaplaincy, not as it existed in theory but as it was actually constituted, on the day of Dogger Bank.

Now, in the name of responsibility, it is important to remember that this database will also be subject to important limitations. While I believe that the above questions would be reasonable uses of our data products, the answers could only be based on a *sample* of the full data which existed at the end of the First World War. There are two reasons for this. The first is that we will only have a selection of the documents that were originally created. Many of the paper documents, sitting in archives, have yet to be digitised. We can't electronically read images that we can't get. Other documents have been destroyed since the end of the war, particularly in the British archives. We will have millions of documents, but we won't have *all* the documents that were ever created.

The second reason is that even with the great processing capacity of Sharcnet, computer reading is never going to be anywhere as good as human-entered data. We estimate that we will only be able to get a maximum of 80% accuracy in our scans. This might make certain kinds of history, like linking documents to names in census records, problematic. Nevertheless, 80% of millions is a lot. For many kinds of research, having 80% of the whole documentary corpus is better than having 100% of 5% of the documentary corpus.

To some extent, this second problem can be mitigated by linking the Muninn data together with other archives which are smaller but higher in quality. Jay Winter suggested a very clever version of this idea to me when we spoke a few weeks ago. Apparently there's been a long-running controversy about whether front-line service made people more prone to suffer health problems later in life. Veterans organisations said yes, and said the government should compensate them. Governments said no, and that they didn't want to pay the money. Well, our database couldn't solve this problem directly, but we could tell you who served in the front line and who didn't. Link that through high quality name lists, like the ones developed by ancestry.co.uk, to census data and you would more or less have your answer.

Ultimately, we have only thought of a fraction of the questions our database could answer. That's why we want to talk to you, get your input and your ideas, and give you the data to play with for your own work.

How it Would Work

So how could all this be achieved? Well, this is where I'm afraid we have to get a little bit technical, but I'll try to be as clear as I can.

All over the world, archives are in the process of digitising their paper records, photographing and scanning them to make digital images which can be stored on a computer. This is good for a whole number of reasons, but it doesn't really add up to this 'Facebook for WWI' idea because these images are still just pictures of paper documents. Computers aren't very good at reading pictures yet, and to make our database we need to know what the actual text on the documents says. Reading all these millions of documents would take a computer of immense power.

Fortunately, we have one. For the next couple of years or so, Muninn is going to have access to one of the most powerful computers in the world. It's called Sharcnet. The main node, in Waterloo Ontario, is a room the size of a football field filled with thousands and thousands of computer chips all working together to crunch numbers at a phenomenal rate. It cost a hundred million dollars to build, it has its own electrical substation, it probably costs a fortune just to keep it powered on, and they're letting us use it for free.

The catch is that they're letting us use it, in large part, because they want to use our project to stress test

the system. In a couple of years time, they won't need us for this any more, and we loose our computer. There's another catch, too, which is that by and large historians, like you and I, don't have the computer programming skills necessary to make the computers read the documents and build the database. We need computer scientists to help us do this, and the computer scientists are only interested in helping us at the moment because it will help them understand certain mathematical problems of their own faster. In three or four years time, somebody will have solved those problems even without our 'help', and the computer scientists will have moved on.

In a sense, we are phenomenally lucky that the right group of people, and the right piece of equipment, has been served up on a plate for us at exactly the right time. On the other hand, if we don't play our cards right, and if we don't get funding for the project, the computer people will move on and our opportunity will just evaporate, perhaps not coming around again for many decades.

How Can You Help?

I want to be very clear that I am not asking you to overthrow your current research programme and join some kind of unit. Historians tend to value their independence and to plan their work years in advance. We understand that, we think it's fine, and we want to find a way of working with you that respects and accommodates your existing interests and schedule.

Having said that, we *do* need your help, and we need it for the two purposes which I described at the top of this article: we want you to help make Muninn better, and we want you to help make Muninn possible in the first place. Let's deal with these one at a time.

As we go through the process of extracting data from documents and rolling it up into a database, we will need advice. There are a lot of problems we can solve by being mathematically clever. But at the end of the day, history texts are history texts, with all the subtlety and slipperiness and complexity that that implies. At some point there is just no substitute for being able to ask a real expert.

Not only that, but we are in the business of producing data 'products' which we hope that you will be able to use. As much as we might want them to be, the design of those data products cannot be 'neutral'. They are going to be biased toward one kind of research or another, and we want them to be biased towards the kind of research that is actually going to be done with them. That means that we want to be able to consult with you, while we're still doing the data extraction and organisation work.

The flip side of this coin is that we want to convince potential funding bodies that our work will give a good return on their academic funding dollar. We need to convince them that lots of people want this research to happen, will use the datasets that we produce, and are willing to help out to make sure our project is a success.

What we propose to do is set up a quasi-formal 'advisory team' composed of WWI experts in various different sub-specialities. There would be a website, possibly an email list, and we would submit to you samples of the freshest data and focussed questions relating to your area of expertise. In return for this service, you would get academic recognition, possible co-authorship on technical publications, and the chance to shape the Muninn WWI database such that it will be of maximum value for your future research.

Timeline

Muninn is a long term project, but at time of writing we are going through a critical phase with regard

to funding. Our first major grant application, to the 'Digging into Data' competition (diggingintodata.org), is due on 15 July 2009. In practice, this means that we need to get all the components of our application together by 1 July. This funding would not kick in until academic year 2010-11, so we are also looking for seed money to start next year. We would need about \$40K in order to start right away. The computer scientists seem to think that this is a small amount of money, but speaking as a historian it seems like a lot. We may or may not, therefore, be able to start next year.

Things That You Can Do

If you want to participate in Muninn, you can do so on any one of a number of levels, here listed in rising order of commitment:

- a) Lend us your name. Let us put your name on a list of people who want Muninn to happen and support our project getting funded. We can submit this list with our funding applications.
- b) Write a letter: Even better would be a formal letter of support, preferably opining that this research is interesting and useful, and that you would likely be an end-user of our data.
- c) Agree, in principal, to be on the advisory team: If you think that you would like to be on the advisory body which shaped the data produced by Muninn WWI, please let us know. You would not only be helping this research take shape in the future, but your reputation and experience would lend gravitas to our grant, making it more likely for the project to proceed.
- d) Join are team as a core member: If you want to join our team as a core member, we would be more than happy to have you. You'd put your name on the website, you'd put your email address on our mailing list, you'd put your signature on the grant application(s) and (potentially) your co-authorship on some of the papers that we produce. You could then have as much or as little control of the project as you want, from being an occasional recipient of Muninn project spam emails to doing actual, serious research work as a major part of the project.
- e) Join as a Principal Investigator: We are still looking for somebody, somewhere in the UK, to agree to fill out the paperwork and be the principal investigator on the UK part of the grant. This would not, in practice, necessarily mean that this person would have to organise a big chunk of the project. If they chose to do this, they could be something akin to a 'lab head' in a science project, taking a parental interest and controlling the flow of money. Alternately, they could be a really serious leading figure in the project. The UK part of the money would be enough to pay for a postdoc.
- f) ... sky's the limit. If you can give us support in any other way, be it moral, intellectual or financial, we would be very happy to receive it.

Conclusion

I hope that you are as excited about the possibilities for this project as I am. If you have any comments, questions, suggestions or complaints, please don't hesitate to email me at ndg25@cam.ac.uk. If you wish to talk to me in person, I can phone you up at any time of your convenience. Just email me and we can set up a time to talk. I hope to hear from you soon.