Notes for Talk

Welcome

- Thanks for inviting me - my research doesn't have an explicit focus on crowdsourcing or community participation *per se*, and I tend to have more of an interest in the mechanics of digital humanities and in particular databases.
    - This talk is somewhat at the geeky end of the discipline, but I want to be clear that I'm not just going to talk at you for twenty minutes about databases!
    - Instead, I'm going to discuss some of the history and importance of linking up data in the Digital Humanities
    - and introduce what may be a new type of datastore you've not come across - graph databases - and their potential for use in future digital humanities work.

The *Digital* in Digital Humanities

- In the abstract, I said this talk will focus on the "digital" in digital humanities.
    - Let's take a moment to reflect and see exactly what I mean by that: how the "digital" and the "humanities" come together.
        - So, at their most fundamental level, computers are very simple machines that use the

movement of electrons in a processor to perform logic. They combine the results of these movements, and store them, in order to process input and produce output.

- By contrast, at the most sophisticated level, as users, we see computers do breathtaking things every day.
    - Routing
    - Social Media
    - Video Calling
    - Amazon Alexa / Google Home
- As you might have guessed: quite a lot happens in this middle bit here, and it’s quite important.
- And, at the risk of simplifying massively, from a digital humanities perspective, this middle bit here is **what we do**.
- If we look at a digital humanities project, for example the classic study of the Medici Family as a social network by Padgett and Ansell (1993) we can see how this works out.
    - At the very “humanities” end of the spectrum, we might have a question like “how was power wielded in 15th century Florence?”
    - And, digital end of the spectrum, we begin to model records of individuals and places in a network or graph, that

we can process on a computer to model the properties of these entities, and reveal patterns of interactions that wouldn’t be visible otherwise.

Brief Historical Perspective: 3 Ages of the Digital in Digital Humanities

The entire history of the discipline of digital humanities involves using new computer technologies to help us explore human experiences. The technologies we use to do this have altered and developed considerably over the years since the beginnings of the discipline of digital humanities.

As a brief bit of history of the discipline, I divide this into three different “phases” or styles that have progressed since the mid-twentieth century.

Before DH, linking together information was seen as one of the main goals of developing information technologies and computing systems.

- The Memex: Vannevar Bush
    - Just after WWII, proposes a device to allow the linking together of information into “flows”, which can be stored and shared between individuals.
- Very influential piece of historical writing, great article to read.

Claimed as inspiration by Doug Engelbart, who would go on to invent the computer mouse, and was an influential figure in the development of the early internet and computing technologies.

Once these technologies came into being, they began to be used by scholars of the humanities.

- In the early 1950s and even earlier we see the humanities computing movement, which used mechanical computers to explore correlations and linkages between data.
    - Fr Roberto Busa - often considered the founder of Humanities Computing, began his *Index Thomisticus* using Hollerith Punch-card computers to explore word freuquencies in the works of Thomas Aquinas.
        - 1954: William Aydelotte: voting records of historical parliamentarians
        - 1949: Harriet and Frank Owsley - *Plain Folk of the Old South*: linking together census records.
        - 1974: Fogel and Engerman - *Time on the Cross*: linking different quantitative historical records, taken to the extreme - very controversial at the time.
- In the 1990s, this changes with the invention of the WWW.
    - The WWW explicitly built on the idea of connecting information in a non-hierarchical way. From Tim Berners-Lee's original proposal:

        - *“the method of storage must not place its own restraints on the information.”*
        - *“This is why a ‘web’ of notes with links (like references) between them is far more useful than a fixed hierarchical system.”*
        - *“When describing a complex system, many people resort to diagrams with circles and arrows. Circles and arrows leave one free to describe the interrelationships between things in a way that tables, for example, do not. The system we need is like a diagram of circles and arrows, where circles and arrows can stand for anything.”*
    - The web was embraced by the digital humanities as a way of connecting information, and of displaying and sharing the connections between these data with others.
        - 1993 - Valley of the Shadow of Death Project
        - modern day - Old Bailey Online
- Nowadays we’re entering a new phase in digital humanities, what you might term the era of “Big Data” - whatever that means. In all sorts of disciplines, we see data being accumulated into databases, and then the use of statistical and computational techniques to perform what is often termed “knowledge discovery” inside of that database in order to extract information about these systems. My favourite example:

        - Machine Learning: Authorial London project at Stanford University - extract location information from the works and biographies of well-known authors, created a digital interactive map of London.

So I argue that DH as a discipline is all about using new computing technologies to link data about human activity together, in order to reveal more about the world we live in and human experiences. So far, we can see three distinct phases in the use of these technologies: early quantitative studies, the use of the WWW, and now knowledge discovery in databases.

Conceptually, this new style of DH is quite interesting. Often, it allows us to play with data before reaching conclusions, to explore and interact with our data without having to be necessarily constrained by it to begin with. This is good for exploratory work, getting a “feel” for data, which is essential in the humanities.

It also marries up very interestingly with a 1962 conceptual proposal by the aforementioned Doug Engelbart, entitled *Augmenting Human Intellect: A Conceptual Framework*. In this paper, Engelbart proposed a new type of human thought that could emerge from computer use.

The development of abstract thoughts through continual, iterative close collaboration with computers through a visual interface which he called: automated external symbol manipulation

- This might lead, he argued, to *"concepts that we have never yet*

imagined" (Engelbart, 1962, p.25)

And it’s exactly this kind of novel, interactive exploration of data in databases that the new style of DH is undertaking. Hopefully, I’ll convince you of that with a demonstration in the second half of this talk.

Databases Introduction: Relational vs. Graph

Databases are thus becoming very important, and understanding them essential for work in the digital humanities.

- When we think of databases generally, you will almost immediately be told about **relational** databases, like MySQL, PostgreSQL or SQL Server.
    - They have tables with rows and columns, and records share IDs through a primary key
    - You’d think that because they’re called *relational* databases, they’re really good at relationships, right?
    - Wrong. They are rubbish at relationships, and I’ll show you why. The only reason they’re called relational is because they use *relational algebra* to process their records.
- Like TBL said:

    - *“When describing a complex system, many people resort to diagrams with circles and arrows.”*
    - So, I thought I’d describe this current situation:
        - a complex international collaboration between two major universities
        - in different cities
        - this talk is part of a bigger event (symposium)
        - which is part of another event (the summer school)
        - I’m a student at this university
        - but have the role of speaker
        - you’re students at the other university
        - you’re in the audience
        - the talk has a topic
- I could go on. So, I got out my trusty pens and put this down on a whiteboard.

- But, what if I wanted to put that in a relational database?

    - Well, I’m going to need a bigger whiteboard, but here goes…
    - What if I wanted to add more data or change it around?
        - Well, asking a question is sort of easy
        - But, what about changing a relationship?
        - And why can’t I name relationships?
- It’s clear relational databases are going to be clunky here. The kind of database which excels with this kind of data, and which

I argue will be important for the future of the digital humanities, however, are graph databases.

    - Graph databases are explicitly meant to be “whiteboard friendly”, meaning that you can chop and change data and model and re-model it on the fly very easily.
    - And so I was able to take that big messy diagram and turn it into a database query very easily.

- Graphs also allow the scanning of large datasets at scale to find patterns: this wouldn’t be possible in relational databases, because to combine multiple tables in a single query can b every slow and memory intensive.

- This is using freely available software: Neo4J. It comes as a community edition, or there’s a paid enterprise edition. It offers a browser to query data, or a REST API, and uses a language called cypher to manipulate data like I’ve shown you.

    - Graph databases, and particularly neo4j - which is very user friendly - have enormous potential for the humanities in linking together different datasets to explore their properties.
        - Proven very useful as a tool in investigative journalism
            - ICIJ and the Panama Papers
            - Donald Trump Business Deals

Geographical Information in Digital Humanities

Now, I thought I'd finish off this talk by using a geographical example, looking at **Volunteered Geographical Information**, and seeing how we can represent place in our databases, and combine this with non-place data.

- In the humanities, we study human actions, and everything has to happen somewhere. So, before we can model this locational data in a database, we have to understand how we model location non-digitally.

- It's often revealing to look at just how many ways you can answer a simple question.

    - **Where are we?**
        - London
        - Anatomy Lecture Theatre, 6th Floor Strand Building, Strand Campus, King's College London, Strand, London WC2R 2LS
        - 51°30'41.2"N 0°06'57.5"W
        - "Five minutes from Waterloo Bridge"
        - Westminster
        - The UK
        - About 40m above sea-level
    - All of these are valid answers, and could be found in the

type of data we deal with in the humanities. However, linking these data together remains a big challenge.

        - Each introduces its own biases etc.
            - Prime meridian
            - Ordinance survey
            - US land parcels

- But, it’s once that we can still tackle using lines and circles.
    - And in Neo4J, we can do this very easily, and so it becomes very easy to combine information about location with any other relevant information we choose, without having to worry ourselves too much with the rigid requirements of a traditional GIS.

London Pubs Example

This is all very good in theory, but when I was putting together this talk, I decided I needed an example to demonstrate the possibilities for using these kinds of database.

- So, I thought, summer students in London - what might be a nice example close to home that would interest you, and me, and hopefully all of us. What might we have experience of?

- Then it dawned on me: the pub.

- So, to demonstrate the potential offered by graph databases, I decided to do a little experiment and combine two sources of

volunteered geographical information in order to demonstrate the potential offered by these sorts of database.

    - I found two sources of information about pubs very easily.
        - One, of course, was the open street map, and the other was a website called Beer in the Evening, which contains lots of user generated reviews of pubs, and lots of useful volunteered information, like:
            - Does it have a pool table?
            - Is there free wifi?
            - Does it serve food?
    - The only problem is, these two datasets aren't linked, and there are lots of little inconsistencies between the datasets:
        - Is "The Lord John Russell" the same as "Lord John Russell"?
            - If they're at the same location, yes
        - Also, pubs shut down, but their listings may stay up. Likewise, they might change name or ownsership.

- How do we model this in a graph, and what can we do with it once we have it?

    - Like I mentioned above, the aim here isn't necessarily to address an immediate research question, but more to

browse and facilitate knowledge discovery in the database, and to see what we can find out about the data once it’s in our DB.
    - So, I scraped the data and, with a little bit of simple preprocessing put it in the graph. Then, with just two queries to combine the two datasets, I managed to combine 60% of all listings - which is reasonably good going considering pub closures, and the fact that the datasets don’t align perfectly.
- With this in mind, let’s see what we can explore base on the db.
    -

Conclusions

- My aim in the above wasn’t to educate you about London pubs *per se*, but rather to demonstrate what’s possible - with relative ease - when using new genrations of data store to link together highly complex data. Here, we can visually explore a network about a relevant topic related to human beings with ease.
- And, as this new phase of DH technology grows, I hope that you can see how we are getting closer and closer to the kind of **enhanced symbol manipulation** that Engelbart talked about way back in 1962.