Digital Humanities Infrastructures

Sigfrid Lundberg

slu@kb.dk
Digital Development and Production
The Royal Library
Post box 2149
1016 Copenhagen K
Denmark

ABSTRACT

This note discusses different initiatives as regards providing access to resources and tools aimed at researchers within the humanities. Since the general breakthrough of search services like Google, annotation and bookmarking services like del.icio.us and the like, the attitudes towards the application of computing within the humanities has changed. In addition, the concept of e-science has contributed to make scholars positive to apply computing to various problems within the humanities.

Following a discussion started around year 2000, we propose that it is possible to formalize many humanities computing tasks as pipelines between XML processing steps.

Introduction

Flanders (2005) claimed in a recent article that the situation for text analysis has changed; since the cultural position of the computer has changed. Even scholars not involved in computing are still computer users and as such consumers of huge amount of net-borne digital data just by using Google.

By taking advantage of Google's page rank in searching, the scholar has become customer of a huge enterprise employing the latest developments in hypertext, text and language technology.

Flanders describes how humanities computing has changed during the last decades; the scholars using these methods no longer perceive that their graphs, tables and statistical inferences unravels pure facts. Rather, their view of their own methodologies has been tainted by the development of arts and humanities at large. Having passed through research styles such as hermeneutics, post-modernism, structuralism and post-structuralism, the scholars involved in computing "have to not only seek and value the pattern that our tools can help us to see, but [they] have to be inquisitive about why these patterns seek us out, why we build tool to see them". (Flanders,

op. cit.)

Following the changes in computer hardware, software and cultural attitudes of scholars, we have seen a large number of initiatives that aim at establishing collections of tools as well as data for the benefit of researchers within the arts and humanities. These collections are often referred to as *infrastructure*. One such initiative established just recently is the Digital Humanities Observatory in Ireland. The very name makes you associate with a vantage point from which human heritage and creativity through the ages can be observed, but it also has connotations: An observatory is a very physical place, providing access to sophisticated scientific instruments.[1]

The idea to establish extensive digital infrastructures aimed at scholars within the humanities goes beyond earlier initiatives. With the advent of e-Science, an infrastructure implies "... methods [that] enable new research by giving researchers access to resources held on widely-dispersed computers as though they were on their own desktops. The resources can include data collections, very large-scale computing resources, scientific instruments and high performance visualisation."(Research Councils)

Digital Infrastructures

Traditionally, a digital infrastructure for the humanities (or indeed any subject area) has meant services like the UK based *Intute - the best Web resources for education and research* (Intute) and *Arts & Humanities Data service* (AHDS). The former is a general subject based information gateway (SBIG) whereas the latter makes it possible for researchers and developers within within archaeology, history, literature, language, linguistics and performing and visual arts to deposit documentation and data for the benefit of future colleagues. Obviously, such infrastructures has included services like on-line digital object archives, dictionaries and encyclopedias.

Available support structures following the advent of e-science on the World-wide Web differs. There are some very interesting research into computer aided humanities research, such as *the noraproject* (Nora), *the monk project* (Monk) and *Text analysis Portal for Research* (TAPor). The nora project produced (among other things) a web based interface for text mining and automatic classification (Plaisant et al., 2006), whereas the monk project (Metadata Offer New Knowledge) seem to concentrate on the interaction between microscopic properties of a text and metadata of the same text. The Nora & Monk projects are cross-disciplinary projects involving expertise in text mining, human computer interaction, information science and various disciplines within the humanities.TAPoR, builds upon a strong vision on the usefulness of text and language technologies in humanities research; users are allowed to create word frequency tables and KWIC con-

[1] DHO is just recently established. The organisation has recently hired quite a few people mostly from the US all having documented XML text encoding, metadata, software development and digital library expertise. This gives an even better idea of what they intend to do, than any mission statement.

cordances on the fly on texts submitted.[2] See also Rockwell (2003).

What would scholars be doing when using an infrastructure?

In order to provide digital research support of the kind envisioned by Research Councils(op. cit.), one has to have a model of what the tedious tasks are in the day-to-day work of a scholar and in particular those that may be alleviated by computing. We do not foresee computer aided innovation, inspiration & creativity and essay writing.

In general one may (following McCarty, 2002) divide the work of humanities computing into three fundamental branches:[3]

1 algorithmic (e.g., calculation of collocation frequencies, linguistic tagging such as parts of speech tagging or lemmatization etc)

2 metalinguistic (e.g., tagging the structure of an object that cannot be ascertained algorithmically)

3 representational (e.g., representing or transforming for viewing or into a KWIC concordance)

The three are in my view not really on equal footing: The representations are generally a product of algorithmic or metalinguistic mark-up available. But the opposite is also true analysis of collocation frequencies can obviously be made in greater detail if there is good metalinguistic markup.

The starting point for McCarty's (op. cit.) arguments is a short paper by Unsworth (2000), who investigates the idea that scholarly work can be subdivided into a finite list of primitive operations, or "self-understood"functions. According his original list, a scholar might be active

- annotating
- discovering
- comparing
- illustrating
- referring
- representing
- sampling

resources.

However, Unsworth seems think that this list can be reduced to a very brief one, "referring" and "representing". Be that as it may. The list of functions will never be complete, even if an object oriented programmer almost certainly would agree with Unsworth. The functions on the list should be recursive, in the sense that one should be able to pipe the functions together like basic Unix tools. I.e., the output of one should be possible to use as input to the next.

[2] The TaPoR is run by a consortium, and there are nodes at several universities in Canada: University of New Brunswick, University of Alberta, McMaster University, University of Toronto and University of Victoria.

[3] The division is due toMcCarty (2002)and references therein; the examples are partly mine.

Also McCarty's humanities computing research infrastructure (or rather "software"—he never mentions infrastructure) should support some list of primitive function but they should be operating within "[…] a singular world-wide entity—a heterogeneous, geographically unconstrained working environment of mutually compatible data and software to which independent, otherwise uncoordinated efforts contribute". He *does realize* that the resources within this entity should be modular and be "in more or less standard format".

An example for putting flesh on the bones. Assume that Jens is preparing a paper on Ludvig Holberg's shaping of male and female characters.[4] The idea is that Holberg, being a male early 18th century playwright, uses contemporary stereotypes as regards class and gender to shape his characters and that these stereotypes are manifested in his languages. In order to be understood these stereotypes had to be easily recognized by by a contemporary audience (who obviously should get a good laugh, or they wouldn't pay the entrance fee).

1 Jens creates an account at www.clarin.dk

2 and after configuring his workspace he starts by **discovering** a set Holberg comedies.

3 He issues a query such that the database returns a document **referencing** the names of the characters within the cast lists (*dramatis personae*) of Holberg's complete works.

4 He then characterises the persons according to gender (male and female) and class (lower, middle and upper) by **annotating** the search result.

5 Having done this, he is then **annotating** the plays such that they are referencing the complete annotated dramatis personae list

6 Next step is the **sampling** of his annotated version of Holberg's Works, such that he gets six new documents, one for each class/gender category which contain speech only.

7 Finally he is **comparing** these texts using statistical and other methods using tools made available at the www.clarin.dk web site.

My guess is that Jens would fail in this design. He will almost certainly find the character's language is more affected by whether they are speaking in the presence of people of lower or higher (or both) ranked people. This is information available in the stage notes in each scene.

Anyway, this might be regarded as positivist rantings from a natural scientist. Humanists in Denmark may, or may not, like this way of working, but this is how I understand how people work within digital humanities and in particular how I understand Unsworth's and McCarty's writings.

There are a few noteworthy things in this story: First (and foremost), each step is standard XML processing in that we have a collection of XML documents, and whatever Jens is doing is leading to a new XML. Secondly, the result of each step is piped more or less directly into the next step. That is:

[4] This is a completely fictional example

Each document, the original collection as well as the derived one must be "in more or less standard format"(McCarty, op. cit.), that is they must belong to one of a few accepted schemas.

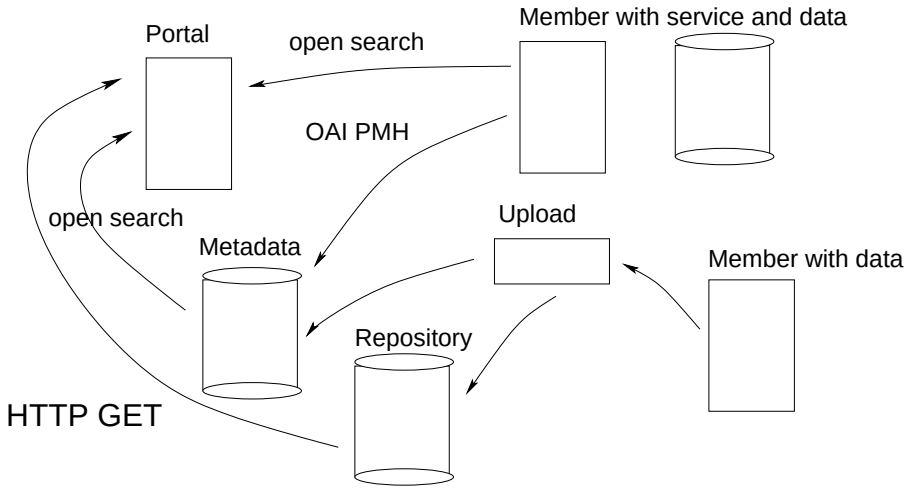

Figure 1. Possible overall architecture of the clarin.dk site, with the back-end aggregator systems shown. For the portal architecture, please refer to Figure 2.

Building a service

Merging the ideas quoted and presented in the previous sections, we realize that the humanities computing infrastructure is more or less the same as the World-wide Web. In many respects much of what is envisioned in the literature is more or less what Tim O'Reilly is evangelising in his set of papers and talks within the linked from his web site *peer-to-Peer Networking, Web Services, and the Emergent Internet Operating System* (O'Reilly). He does not mention humanities resources, but the his visions about the tools are similar. Since many of these exists already, we will not start from scratch.

The Danish Clarin network will consist by member organisations, whose staff is contributors to as well as users of the collections maintained within the network. Some point within the network must—for practical—reasons be assigned the role of central hub (that is, users need somewhere to find information about the services provided). Also, a central index helps when it comes to provide fast and efficient resource discovery. The network must have facilities for for dissemination of metadata. This site will also be the home of user contributed data and possibly also mirrors of the other repositories (Figure 1).

There will be a need for a set of protocols for metadata harvesting as well as for distributed searching. At this time, the most promising candidates for these functions are *Protocol for Metadata Harvesting* (Open Archives Initiative), *Object Reuse and Exchange* (Open Archives Initiative) and *OpenSearch/1.1/Draft 3* (OpenSearch), respectively. Mirroring (if at all) will be made by HTTP based on the data revealed via OAI-PMH by member sites. Members without database driven access to data will have to provide

data through other means, for example manual file upload.

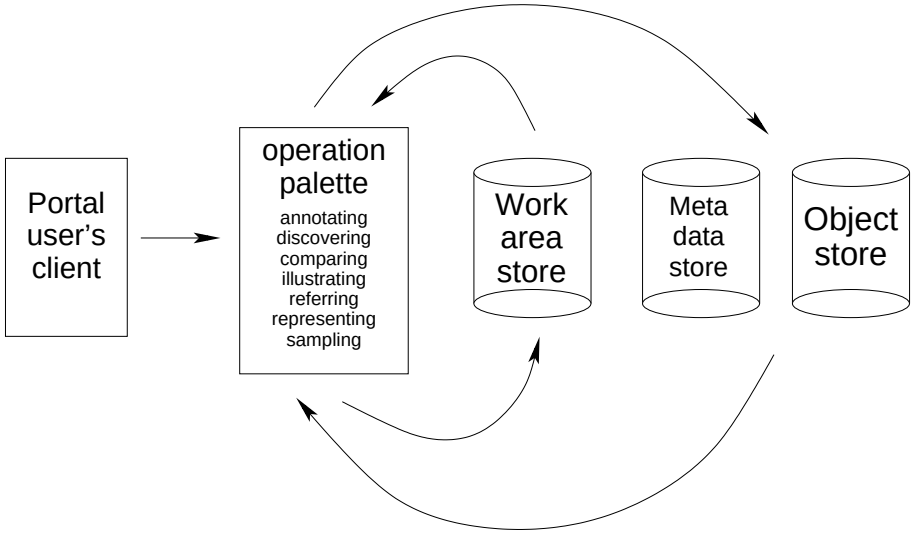

Figure 2. Possible architecture of the clarin.dk portal.

The portal's architecture is depicted in Figure 2. The main function in the stack of operations discussed above is**discovering**. For authenticated users, we foresee that all transactions between the system will be in a standardised XML format rather than HTML. In this vision the rendition of documents is performed in the user's web client. The user will thus be able to process any documents delivered with his or her favourite native XML tools. We propose that the portal is built as an engine based on *TEI lite* (Burnard & Sperberg-McQueen, 2006). It is possible to produce very sophisticated hypertext documents in this format for direct rendition client side (the current document is an example of this).

Given a standard, modular and extensible method to implement and document XML processing pipelines, we are confident that the pipeline stack of Figure 2 is implementable. There are systems available suitable for the purpose, the most promising one is *XProc, An XML Pipeline Language* (Walsh et al., 2008). A conforming XProc processor accepts one or more XML documents as input and produces the same as output. It should support XSLT 1.0 and 2.0 for transformation, XPath 2.0 and XQuery 1.0 for querying. XQuery may be used for transformation as well as querying. With this arsenal we envisage that the whole set Unsworth's functions can be implemented without programming in the meaning that staff will have to write an extensive amount of (say) Java code to implement the individual pipes.

References

Arts and Humanities Data Service `<URL: http://ahds.ac.uk/>`

Burnard, Lou, and C. M. Sperberg-McQueen, 2006. *TEI Lite: Encoding for Interchange: an introduction to the TEI Revised for TEI P5 release* `<URL: http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html>`

Digital Humanities Observatory <URL: http://www.dho.ie/>

Flanders, Julia, 2005. Detailism, Digital Texts, and the Problem of Pedantry. *TEXT Technology*. Vol. 14(2), pp. 41-70. <URL: http://texttechnology.mcmaster.ca/pdf/vol14_2/flanders14-2.pdf>

Intute - the best Web resources for education and research <URL: http://www.intute.ac.uk/>

McCarty, Willard, 2002. Humanities Computing: Essential Problems, Experimental Practise. *Literary and Linguistic Computing*. Vol. 17(1), pp. 103-125. <URL: http://llc.oxfordjournals.org/cgi/content/abstract/17/1/103>

Monk project <URL: http://www.monkproject.org/>

Nora project <URL: http://www.noraproject.org/>

O'Reilly, Timothy *Peer-to-Peer Networking, Web Services, and the Emergent Internet Operating System* <URL: http://tim.oreilly.com/p2p/index.csp>

Open Archives Initiative *Protocol for Metadata Harvesting* <URL: http://www.openarchives.org/ore/>

Open Archives Initiative *Object Reuse and Exchange* <URL: http://www.openarchives.org/ore/>

Open Search *OpenSearch/1.1/Draft 3* <URL: http://www.opensearch.org/Specifications/OpenSearch/1.1>

Plaisant, Catherine, James Rose, Bei Yu, Loretta Auvil, Matthew G. Kirshenbaum, Martha Nell Smith, Tanya Clement, and Greg Lord, 2006. Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL'06)*. <URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.3281>

Research Councils [the UK] *What is e-Science?* <URL: http://www.rcuk.ac.uk/escience/default.htm#phMain>

Rockwell, Geoffrey, 2003. What is Text Analysis, Really? *Literary and Linguistic Computing*. Vol. 18(2), pp. 209-219. <URL: http://llc.oxfordjournals.org/cgi/content/abstract/18/2/209>

TAPoR - Text Analysis Portal for Research <URL: http://portal.tapor.ca/>

Unsworth, John, 2000. Scholarly Primitives: what methods do humanities researches have in common, and how might our tools reflect this? *Part of a symposium on "Humanities Computing: formal methods, experimental practise" sponsored by King's College, London, May 13, 2000.* <URL: http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html>

Walsh, Norman, Alex Milowski, and Henry S. Thompson, 2008. *XProc: An XML Pipeline Language* <URL: http://www.w3.org/TR/xproc/>