Harvard University Library Lab Proposal
Highbrow: A Deep Zoom Widget for Textual Annotations

by Reinhard Engels, Digital Library Software Engineer, Office for Scholarly Communication

A "Big Science" Visualization Tool for the Humanities

Bioinformatics "genome browsers" have been workhorse tools in the life sciences for the last decade. In their most basic form, these programs show zoomable stacks of annotations as (usually horizontal) bands on a linear genome sequence. At the fully zoomed out "global" level, scientists can see which regions of a genome are most densely annotated and can then zoom in to inspect individual annotations in detail. The sequence, of course, is DNA, and the annotations are mostly generated by algorithms such as BLAST or HMMER, but, conceptually at least, the sequence could just as easily be the text of Hamlet, and the annotations by Coleridge, Stephen Greenblatt -- or Harvard undergrads.

I believe that humanists and other textual scholars might find such a tool extremely useful. At the very least, it could be useful as a demonstrative teaching aid, to highlight the differences between thinkers. But, more importantly, it might also serve as an exploratory tool for identifying new openings for research (even negatively -- desperate grad students could identify "annotation deserts" of relatively uncharted territory to write their theses on).

No such textual annotation viewer currently exists. But, leveraging my experience creating similar tools for bioinformatics at the Broad Institute of Harvard and MIT, it would not be difficult for me to adapt a functional prototype that I believe would have an immediate, powerful appeal to many humanists. My working name for it is "Highbrow," short for "Highlight Browser" (and lightheartedly suggesting its Grand Literary Ambitions).

A use case

Although Highbrow could be used to display annotations on any text, its most obvious application would be on heavily annotated texts with established coordinate systems -- such as the Bible, the Koran, or the works of Plato.

Imagine that the text we are examining is the Bible. In separate bands we could map the scriptural citations of St. Augustine, Thomas Aquinas, Calvin, Nietzsche, and dozens of other thinkers. At the "global" level, Highbrow displays a density plot for each thinker's references, a kind of visual fingerprint of their relationship to the work. The regions that were of interest to Calvin but not Luther, or Calvin and Luther but not Ignatius Loyola jump out. When users zoom in sufficiently they see the citation information for individual annotations, and can click on an annotation to see a "comic book" bubble with a a snippet of the reference in context (with a hyperlink to the source, if available).

Below are some screenshots of a crude (but basically functional) prototype (on fake data) at http://osc-dev.hul.harvard.edu/highbrow/demo/kjv

Figure 1: Highbrow alpha prototype zoomed out to global view of the King James Bible

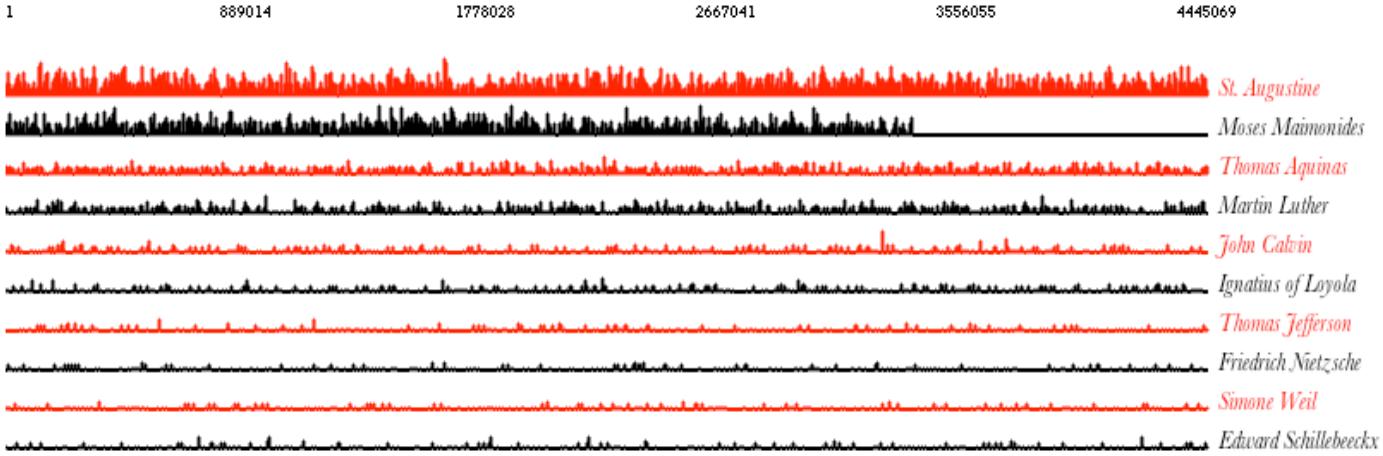
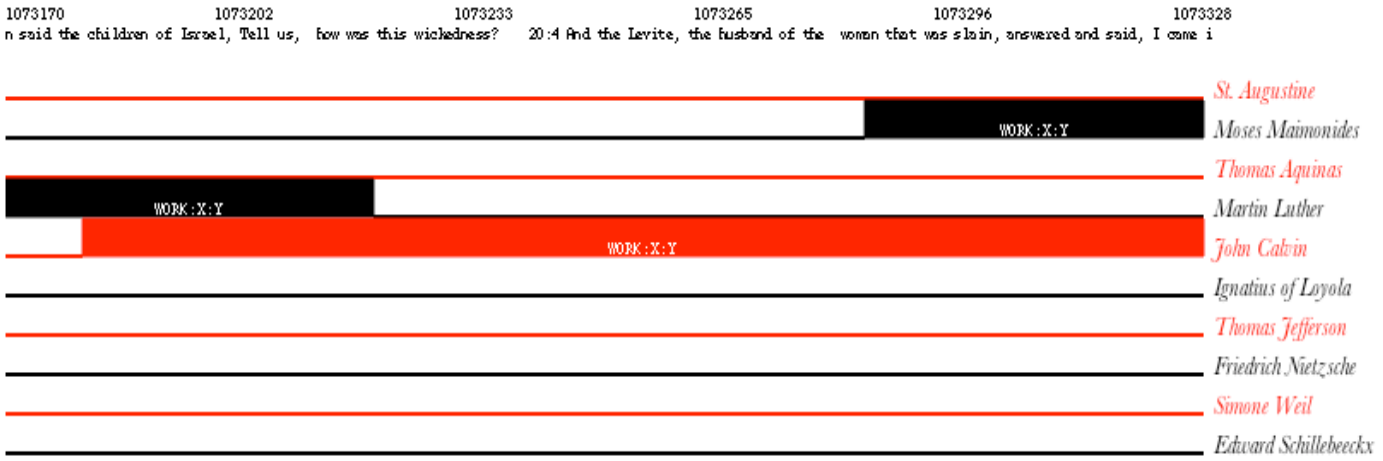


Figure 2: Highbrow alpha prototype Zoomed in to individual annotations:



Scalability

Extremely large data sets could be supported using established deep zoom techniques of creating “pyramids” of tiled data at multiple resolutions, as do Google Maps for geographic data and Microsoft Seadragon for high resolution images. In the case of textual annotations, the bottom of the pyramid would consist of bins of raw data, chunks of individual annotations that only need to be retrieved at very close zoom. Bins of increasingly dense summary scores would be stored at higher levels of the pyramid. This would allow a global summary view, even of massive amounts of data, to load quickly, and provide seamless zoom down to the individual annotation level without requiring that the entire data set be downloaded over the network.

How big are literary annotation data sets? According to James W. Wiles’s 1995 compilation, there are 31,817 biblical citations in the complete works of St. Augustine. For a human being, this is an impressive output -- but it is small by bioinformatics standards. There is no question

that the visualization paradigms which have already proven themselves in the field of genomics will be able to handle multiple data sets of this size.

Implementation

In order to reach the widest possible audience, I intend to write Highbrow using javascript and the HTML 5 canvas element for browser native graphics. Canvas is an open standard supported in Firefox, Safari, Chrome, and (soon) IE 9. Unlike heavyweight solutions like Java Applets or Adobe Flash, it works in both IOS (iPhone/iPad) and Android based mobile devices. The high-level processing.js library provides a robust, readable, rapid-development environment for creating canvas based graphics.

The code will be entirely client side javascript, except possibly for a supplemental data conversion script (to generate the deep zoom pyramid for large datasets from a simpler input file). This means it can be easily installed and configured by the humanists themselves (and the more technically inclined can just “view source” to see what is going on under the hood).

The Big Challenge: Getting the Data

Getting existing sets of annotations mapped to the correct sequence coordinates will be far more work than the merely technical challenge of coding the viewer. Still, once domain experts see this tool in action on a sample data set, and if the annotation input format is simple, documented, and easy to understand, I believe they will be highly motivated (and far more qualified than I) to put in the work to massage their data so that it can be viewed by this tool. It is also a well-defined problem that could be effectively crowdsourced (at least in part).

For the initial implementation, I will write crude import and analysis scripts to get real, sizable, (but incomplete) data sets in place -- an “80% solution” sufficient to give a good (and perhaps even useful) demonstration of the potential of Highbrow, and a data core for more qualified experts to build on.

Resouces & Deliverables

Coding a usable version of the viewer suitable for public “beta” release should take no more than three weeks of my time, with another week for data generation, documentation, and testing. I could spread this out over the course of more than one month so as not to neglect other tasks: perhaps half of my time for two months.

For this first iteration, only the most basic functionality would be supported: (deep) zoom, pan, and inspection of individual annotations. The goal would be to deliver something simple, but solid, fast, and scalable for users to start experimenting with.

Logical next steps might be: text and annotation search, “deep” permalinks to particular zoom states, interactive editing of annotations, and batch selection and processing. But users might have different priorities and these should direct future development. We can evaluate whether further effort is warranted depending on the response to the initial release.

Beyond a few megabytes of static web hosting space, no server side resources will be required.

Measuring Success

The best way to measure success will be if scholars, librarians, and interested amateurs download and install the widget to publicly display their data.

Another valuable metric will be end user traffic to library hosted pages that have embedded the widget.

Because a first iteration will be relatively easy to produce, we can assess from there whether further development is warranted. It is a low-risk project with high potential upside.

Internal and External Collaborators

Although it would be useful to anyone who wanted to display large numbers of textual annotations from multiple sources, I believe Highbrow would be of particular interest to literary scholars, theologians, classicists, philosophers, and perhaps even linguists. I am eager to find Harvard faculty members to partner with, but I believe this will be easier once a functional prototype exists for them to play with.

From having attended the MIT Hyperstudio Information Visualization in the Humanities conference and the Harvard Digital Scholarship Summit 2010, I got the sense that many humanists consider themselves under-served by digital technologies relative to their peers in the sciences. Highbrow would be a (small) step towards rectifying that imbalance.

I spoke briefly with Phil Desenne of ATG about the possibility of plugging a visualization tool like Highbrow into the Common Collaborative Media Annotation Framework his group is working on. He seemed very receptive to the idea and I am eager to follow up with him. I do not think this will impact development of the first release, but it might be a fruitful long-term collaboration.

There is also a w3c semantic web annotation standard called annotea with an associated firefox plugin called annozilla, which do not seem to have gathered much traction, but are worth looking into further. I would also like to explore the possibility of plugging into the Amazon kindle highlight sharing architecture.

When I was at the Open Repositiories 2010 Conference, I discussed the idea of collaborating on a literary annotation browser with Youssef Mikhail Bassily, Head of Software and System Development at the Bibliotheca Alexandrina. I would love to approach him again about this if the project is approved. A joint venture of Harvard with the modern incarnation of the greatest library of the ancient world and preeminent digital library of the Arabic-speaking world might have symbolic, as well as technical benefits.

Highbrow “Comparables”

Since plotting annotations and search terms against literary works seems obviously desirable and technically not very difficult, it is a little surprising that it has not been done before. But after much googling and conversations with experts here at Harvard (Phil Desenne at ATG, Mark Shiefsky in Classics, Alexander Parker of the Digital Humanities initiative), I am convinced that it really has not been done yet.

Below are the closest comparable tools that I was able to find. Some of them are very interesting in their own right, but none does quite what Highbrow will do. Furthermore, far from being “competition,” several of them present opportunities for collaboration.

Arboreal / Archimedes

Arboreal is a tool for viewing and annotating XML texts. It was developed as part of project Archimedes here at Harvard. While very powerful as a page-by-page editing tool, it does not do much in the way of visualization. I am working with Mark Shiefsky, the principal investigator of project Archimedes here at Harvard, to explore how Highbrow could complement the tools he already has in place.

Common Collaborative Media Annotation Framework

Harvard’s Academic Technology group is working on a general-purpose framework to support annotation of resources in multiple data formats including text, images, audio, and video. When I showed Highbrow to the project leader, Phil Desenne, he was very excited about plugging it into their framework to provide a high-level overview, which they are currently lacking.

Genome Browsers

Genome Browsers were my initial point of inspiration for Highbrow, and there are quite a number of them in active use. While Highbrow borrows many of their visualization techniques and metaphors, none of them is well-suited for viewing non-genomic sequences as is.

Google Ngram Viewer

This is a fantastic tool launched just this month (December 2010). It allows users to plot the frequency of search term appearance in the google books corpus by date. It supports tremendous scale (“4% of all books ever written”), but the user-interface is very bare-bones, and it is not useful for investigation of individual works. It also has no support for annotations. The authors of this tool are also obviously inspired by genomics, using the term “culturomics.” Since many of the principal investigators are at Harvard, it may be worth exploring some form of collaboration.

Understanding Shakespeare

Stephen Thiel at the University of Potsdam recently produced a series of static images depicting various trends in Shakespeare’s works. Although interesting, these visualizations are not-

interactive, and difficult to reproduce for other texts. I also believe that at least some of the underlying data he presents could be more effectively communicated using Highbrow's zoomable graphs (perhaps I will contact him about this).