

Data Mining with Criminal Intent
Interim Report
July 30, 2010

Daniel Cohen
Center for History and New Media
George Mason University

It has been a busy first six months for the Data Mining with Criminal Intent project. As an international collaboration, unsurprisingly the first month was spent setting up modes of interaction and strategic planning. The team established monthly phone calls to discuss progress on various fronts and to iron out any issues related to the functional specifications of the proposed work. We also set up an online project management system for the exchange of documents, timelines, and messages.

Next we designed and built a project website at <http://criminalintent.org>, which is hosted at the Center for History and New Media. That website not only laid out our plans for the coming year but also began to open up our thinking about how we were actually going to exchange and analyze data. To date, all project partners have blogged about their thoughts, and we have posted updates to the site. We plan to continue this open process so others can learn from our Digging into Data experience.

We also had two in-person meetings of a majority of the project partners during the first six months, one in Canada and one in the UK. The first explored in greater technical depth some of the possibilities for data mining the Old Bailey collection. Methods from commercial “data mart” techniques, to normalized compression distance, to more free-form “conjecture” prospecting were assessed both from a code perspective and, more importantly, from a historical perspective.

Out of that first meeting and the regular online and phone conversations we settled on two major modes for the interaction of Old Bailey, Zotero, and Voyeur. The first, which we nicknamed the “slice,” was to move a large saved search—effectively a slice of the Old Bailey database that met certain requirements—over to Voyeur without pulling down the full text into Zotero. We are currently working through the best way to send that large text corpus directly, based on a specific URI encapsulated in Zotero. The second mode, nicknamed “hunt and peck,” is intended to satisfy the method we heard from historians where they handpick certain documents for further analysis. In this case, individual documents from Old Bailey would indeed be pulled down the wire into Zotero before being packaged and sent to Voyeur.

The second meeting occurred in London just before DH2010. The project principles reviewed the existing work, including modifications and plugins to Old Bailey, Zotero, and Voyeur, to plan for getting them to production stage for use by historians. The Old Bailey API is now essentially up and running, structured around tagged data that can be grabbed by Zotero. We settled on normalized compression distance as our first major

effort of analysis by the Voyeur team, although initially (given the large corpus transfer issue) this might have to be done on a pre-processed rather than transfer-and-analyze-on-the-fly basis (i.e., the entire Old Bailey database would be pre-processed by Voyeur, something that we hope to avoid in the final version because it is not replicable by other collections Zotero users might want to mine). We also went over the mockups of the interfaces.

We also have been discussing the best way to save collections as generated from the Old Bailey API in Zotero. We are particularly working on a way to ensure that a functioning version of the URI that generated a specific collection could be saved in Zotero in association with the study collection. For the average historian to understand what is going on in the process, we also realized that we need a plain language statement interpretation of that URI (e.g., “This URI points to all murders by teenagers between 1830 and 1850”) that would be understandable as a representation of how they arrived at the search results that made up a study collection. This search URI would then need to be capable of being re-imported into the Old Bailey API for further refinement and use (e.g., if the historian wanted to change the end date from 1850 to 1870).

We also have thought through a range of research narratives, and how they might work with our proposed interface: starting, for instance, from a single distinct trial in the Old Bailey and searching for similar texts. The process is complicated by the multi-variant nature of the tagged data (i.e. a ‘trial’ can contain multiple crimes, defendants and verdicts), but various strategies have been proposed to overcome this.

In short, we have made substantial intellectual and technical progress on the project in the first six weeks. The collaboration is going very smoothly, and we have structures in place to complete the project on track over the next year.