Notes of the user group meeting on Feb 27, 2014

Participants: Douglas Duhaime, Matthew Wilkens, Miao Chen, Michelle A. Paolillo, Sayan Bhattacharyya, Loretta Auvil

Douglas Duhaime's shared his research story with HTRC, and the participants also discussed some issues related to HTRC today.

Doug is a graduate student from Notre Dame. He is interested in natural philosophy of science in 18th century, tracing relationship between natural philosophy and science. He uses "philosophical transactions", one of the greatest scientific publication in England back then.

He just started using HTRC resources, collections resources of the philosophical transactions and other literatures citing philosophical transactions.

Sayan :are there certain words that keep chaining over time, that you're looking for?
Doug: orthography keeps changing from 16th to 18th century, words keep changing. But he is not particularly looking at words, he's looking for common sequence of words, to look for common substrings between two publications. He has used HTRC resources to look for such substrings.

Sayan: this is an opportunity for HTRC to find how scholars can use our resources. What would be some tools that might need for literature comparison?

Doug: The 1st is to it'd be nice to be able to make large connections on HTRC, the larger the better. It improves my odds for finding common strings. It would be good to have all philosophical transactions in his collection.

How large?
about 1000 maybe.

Matt Wilken: maybe thousands, hundred thousands, but not a million.
WCSA is finalizing this week, or in next couple of weeks.

Will WCSA lead to support large collection?
Sayan: it will support heterogenous collections, e.g. connecting some online resources not part of HathTrust to use's work set.

Doug: something tricky about API level stuff is it doesn't acquire full text in order.
His methods uses text window to find substrings, you can't use only bag-of-words for that. You need to know order of text.
Loretta: if you query Solr directly, you can find one word near another word directly.
It doesn't tell where and how many time it happens in text, it just tells you the volume.

Matt: similar bigrams "King will", and "Kinge will", it would be nice to set an edit distance threshold, and mark those similar ones as match.
Loretta: you can use some * match scheme.

Matt: the quality of underlying text, how good is the OCR? how about the quality?
Loretta: poor.

Matt: has there been discussion on running some clean routine?
Loretta: there is some clean routine in the portal, depending on what you are doing (which algorithm).

Matt: the thing really nice would be able to create some collections, submit script on these collections and get output through a general portal.
Loretta: there is an aspect for generating virtual machine for people to use, not ready; two is have portal for people.

Doug wondered the data structure of text in the current form. One problem he encountered is, he put together some collection, 1660-1701 all the work, run the Marc Downloader, he found one text that was include but actually fell outside the time frame. Loretta will look into it.

Mentioned he used relative frequency for words in text, when he attempted to load, the progress is spinning. He will try again later.