How not to work during a sabbatical

This presentation outlines the various software systems I wrote during my recent tenure as an adjunct faculty member at the University of Notre Dame.

- **How rare is rare?** - In an effort to determine the "rarity" of items in the Catholic Portal, I programmatically searched WorldCat for specific items, counted the number of times it was held by libraries in the United States, and recorded the list of the holding libraries. Through the process I learned that most of the items in the Catholic Portal are "rare", but I also learned that "rarity" can be defined as the triangulation of scarcity, demand, and value. Thus the "rare" things may not be rare at all.

- **Image processing** - By exploiting the features and functions of an open source library called OpenCV, I started exploring ways to evaluate images in the same way I have been evaluating texts. By counting & tabulating the pixels in an image it is possible to create ratios of colors, do facial recognition, or analyze geometric composition. Through these processes is may be possible to supplement the discipline of art history and criticism. For example, one might be able to ask things like, "Show me all of the paintings from Picasso's Rose Period."

- **Library Of Congress Name Authorities** - Given about 125,000 MARC authority records, I wrote an application that searched the Library Of Congress (LOC) Name Authority File, and updated the local authority records with LOC identifiers, thus making the local authority database more consistent. For items that needed disambiguation, I

created a large set of simple button-based forms allowing librarians to choose the most correct name.

- **MARC record enrichment** - Given about 500,000 MARC records describing ebooks, I wrote a program that found the richest OCLC record in WorldCat and then merged the found record with the local record. Ultimately the local records included more access points and thus proved to be more useful in a library catalog setting.

- **OAI-PMH processing** - I finally got my brain around the process of harvesting & indexing OAI-PMH content into VUFind. Whoever wrote the original OAI-PMH applications for VUFind did a very good job, but there is a definite workflow to the process. Now that I understand the workflow it is relatively easy to ingest metadata from things like ContentDM, but issues with the way Dublin Core is implement still make the process challenging.

- **EEBO/TCP** - Given the most beautiful TEI mark-up I've ever seen, I have systematically harvested the Early English Books Online (EEBO) content from the Text Encoding Initiative (TCP) and done some broad & deep but also generic text analysis against subsets of the collection. Readers are able to search the collection for items of interest, save the full text to their own space for analysis, and have a number of rudimentary reports done against the result. This process allows the reader to see the corpus from a "distance". Very similar work has been done against subsets of content from JSTOR as well as the HathiTrust.

- **VIAF Lookup** - Given about 100,000 MARC authority records, I wrote a program to search VIAF for the most appropriate identifier and associate it with the given record. Through the process I learned two things: 1) how to exploit the VIAF API, and 2) how to exploit the Levenshtein algorithm. Using the later I was able to make automated and "intelligent" choices when it came to name disambiguation. In the end, I was able to accurately associate more than 80% of the authority names with VIAF identifiers.

My tenure as an adjunct faculty member was very much akin to a one year education except for a fifty-five year old. I did many of the things college students do: go to class, attend sporting events, go on road trips, make friends, go to parties, go home for the holidays, write papers, give oral presentations, eat too much, drink too much, etc. Besides the software systems outlined above, I gah eave four or five professional presentations, attended & helped coordinate five or six professional meetings, taught an online, semester-long, graduate-level class of on the topic of XML, took many different classes (painting, sketching, dance, & language) many times, lived many months in Chicago, Philadelphia, and Rome, visited more than two dozen European cities, painted about fifty paintings, bound & filled about two dozen hand-made books, and took about three thousand photographs. The only thing I didn't do is take tests.

Eric Lease Morgan
emorgan@nd.edu

July 13, 2016