The Next Next-Generation Library Catalog

With the advent of the Internet and wide-scale availability of full-text content, people are overwhelmed with the amount of accessible data and information. Library catalogs can only go so far when it comes to delimiting what is relevant and what is not. Even when the most exact searches return 100′s of hits what is a person to do? Services against texts — digital humanities computing techniques — represent a possible answer. Whether the content is represented by novels, works of literature, or scholarly journal articles the methods of the digital humanities can provide ways to compare & contrast, analyze, and make more useful any type of content. This essay elaborates on these ideas and describes how they can be integrated into the "next, next-generation library catalog".

Find is not the problem

Find is not the problem to be solved. At most, find is a means to an end and not the end itself. Instead, the problem to solve surrounds use. The profession needs to implement automated ways to make it easier users do things against content.

The library profession spends an inordinate amount of time and effort creating catalogs — essentially inventory lists of things a library owns (or licenses). The profession then puts a layer on top of this inventory list — complete with authority lists, controlled vocabularies, and ever-cryptic administrative data — to facilitate discovery. When poorly implemented, this discovery layer is seen by the library user as an impediment to their real goal. Read a book or article. Verify a fact. Learn a procedure. Compare & contrast one idea with another idea. Etc.

Instead of focusing on find, the profession needs to focus on the next steps in the process. After a person does a search and gets back a list of results, what do they want to do? First, they will want to peruse the items in the list. After identifying items of interest, they will want to acquire them. Once the selected items are in hand users may want to print, but at the very least they will want to read. During the course of this reading the user may be doing any number of things. Ranking. Reviewing. Annotating. Summarizing. Evaluating. Looking for a specific fact. Extracting the essence of the author's message. Comparing & contrasting the text to other texts. Looking for sets of themes. Tracing ideas both inside and outside the texts. In other words, find and acquire are just a means to greater ends. Find and acquire are library goals, not the goals of users.

It starts with counting

The availability of full text content in the form of plain text files combined with the power of computing empowers one to do statistical analysis against corpora. Put another way, computers are great at counting words, and once sets of words are counted there are many things one can do with the results, such as but not limited to:

- measuring length
- measuring readability or any other index
- measuring frequency of n-grams
- charting & graphing analysis
- analyzing measurements
- drawing conclusions

For example, suppose you did the perfect search and identified all of the works of Plato, Aristotle, and Shakespeare. Then, if you had the full text, you could compute a simple table such as Table 1.

The table lists who wrote how many works. It lists the number of words in each set of works and the average number of words per work. Finally, based on things like sentence length, it estimates grade and reading levels for the works. Given such information, a library "catalog" could help the patron could answer questions such as:

Author	Works	Words	Average	Grade	Flesch
Plato	25	1,162,461	46,499	12-15	54
Aristotle	19	950,078	50,004	13-17	50
Shakespeare	36	856,594	23,794	7-10	72

- Which author has the most works?
- Which author has the shortest works?
- Which author is the most verbose?

When a person counts the number of times individual words are used by each of the authors,

the words “one”, “good”, and “man” are in the top twenty-five. Is there a pattern here? If one word contains some meaning, then do two words contain twice as much meaning? Counting the significant two-word phrases (bigrams) in each corpus a new pattern appears. Shakespeare uses many names, but Plato and Aristotle use many concepts. For example, some of the names in Shakespeare include King Henry, Mark Antony, and Duke Vincentio. While Plato and Aristotle mention human nature, human mind, false opinion, essential nature, practical wisdom, and scientific knowledge.

A concordance is particular useful for finding and displaying words and phrases in context. What does each author have to say about a “good man”?

Plato

```
ngth or mere cleverness. To the good man, education is of all things the most pr
 Nothing evil can happen to the good man either in life or death, and his own de
but one reply: 'The rule of one good man is better than the rule of all the rest
 SOCRATES: A just and pious and good man is the friend of the gods; is he not? P
ry wise man who happens to be a good man is more than human (daimonion) both in
```

Aristotle

```
ons that shame is felt, and the good man will never voluntarily do bad actions.
reatest of goods. Therefore the good man should be a lover of self (for he will
hat is best for itself, and the good man obeys his reason. It is true of the goo
theme If, as I said before, the good man has a right to rule because he is bette
d prove that in some states the good man and the good citizen are the same, and
```

Shakespeare

```
r to that. SHYLOCK Antonio is a good man. BASSANIO Have you heard any imputation
p out, the rest I'll whistle. A good man's fortune may grow out at heels: Give y
t it, Thou canst not hit it, my good man. BOYET An I cannot, cannot, cannot, An
hy, look where he comes; and my good man too: he's as far from jealousy as I am
 mean, that married her, alack, good man! And therefore banish'd -- is a creatur
```

Digital humanities, and “catalogs”

The previous section is just about the most gentle introduction to digital humanities computing possible, but can also be an introduction to a new breed of library science and library catalogs.

It began by assuming the existence of full text content in plain text form — an increasingly reasonable assumption. After denoting a subset of content, it compared & contrasted the sizes and reading levels of the content. By counting individual words and phrases, patterns were discovered in the texts and a particular idea was loosely followed — specifically, the definition of a good man.

While the software used to do this analysis were really toys, the potential they represent are not. It would not be too difficult to integrate their functionality into a library “catalog”. Assume the existence of significant amount of full text content in a library collection. Do a search against the collection. Create a subset of content. Click a few buttons to implement statistical analysis against the result. Enable the user to “browse” the content and follow a line of thought.

Again, find is not the problem to be solved. People can find more information than they require. Instead, people need to use and analyze the content they find. This content can be anything from novels to textbooks, scholarly journal articles to blog postings, data sets to collections of images, etc. The process outlined above is an example of services against texts, a way to “Save the time of the reader” and empower them to make better and more informed decisions.

The next “next generation library catalog” is not about find, instead it is about *use*. Integrating digital humanities computing techniques into library collections and services is just one example of how this can be done.

Eric Lease Morgan
University of Notre Dame

June 24, 2010