Use & understand: A DPLA beta-sprint proposal

This essay describes, illustrates, and demonstrates how the Digital Public Library of America (DPLA) can build on the good work of others who support the creation and maintenance of collections and provide value-added services against texts – a concept we call “use & understand”.

The canonical location of this proposal is http://infomotions.com/blog/2011/08/dpla/, and there the reader will find HTML, PDF, and ePub versions of this proposal for viewing in different environments.

Eric Lease Morgan <emorgan@nd.edu>
University of Notre Dame

September 1, 2011

Table of Contents

Executive summary

This Digital Public Library of America (DPLA) beta-sprint proposal "stands on the shoulders of giants" who have successfully implemented the processes of find & get – the traditional functions of libraries. We are sure the DPLA will implement the services of find & get very well. To supplement, enhance, and distinguish the DPLA from other digital libraries, we propose the implementation of "services against text" in an effort to support use & understand.

Globally networked computers combined with an abundance of full text, born-digital materials has made the search engines of Google, Yahoo, and Microsoft a reality. Advances in information retrieval have made relevancy ranking the norm as opposed to the exception. All of these things have made the problems of find & get less acute than they used to be. The problems of find & get will never be completely resolved, but they seem adequately addressed for the majority of people. Enter a few words into a search box. Click go. And select items of interest.

Use & understand is an evolutionary step in the processes and functions of a library. These processes and functions enable the reader to ask and answer questions of large and small sets of documents relatively easily. Through the use of various text mining techniques, the reader can grasp quickly the content of documents, extract some of their meaning, and evaluate them more thoroughly when compared to the traditional application of metadata. Some of these processes and functions include: word/phrase frequency lists, concordances, histograms illustrating the location of words/phrases in a text, network diagrams illustrating what author say "in the same breath" when they mention a given word, plotting publication dates on a timeline, measuring the weight of a concept in a text, evaluating texts based on parts-of-speech, supplementing texts with Wikipedia articles, and plotting place names on a world maps.

We do not advocate the use of these services as replacements for "close" reading. Instead we advocate them as tools to supplement learning, teaching, and scholarship – functions of any library.

Introduction and assumptions

Libraries are almost always a part of a larger organization, and their main functions can be divided into collection building, conservation & preservation, organization & classification, and public service. These functions are very much analogous to the elements of the DPLA articulated by John Palfrey: community, content, metadata, code, and tools & services.

This Beta-Sprint proposal is mostly about tools & services, but in order to provide the proposed tools & services, we make some assumptions about and build upon the good work of people working on community, content, metadata, and code. These assumptions follow.

First, the community the DPLA encompasses is just about everybody in the United States. It is not only about the K-12 population. It is not only about students, teachers, and scholars in academia. It is not only about life-long learners, the businessperson, or municipal employees. It is about all of these communities at once and at the same time because we believe all of these communities have more things in common than they have differences. The tools & services described in this proposal can be useful to anybody who is able to read.

Second, the content of the DPLA is not licensed, much of it is accessible in full-text, and freely available for downloading and manipulation. More specifically, this proposal assumes the collections of the DPLA include things like but not necessarily limited to: digitized versions of public domain works, the full-text of open access scholarly journals and/or trade magazines, scholarly and governmental data sets, theses & dissertations, a substantial portion of the existing United States government documents, the archives of selected mailing lists, and maybe even the archives of blog postings and Twitter feeds. Moreover, we assume the DPLA is not merely a metadata repository, but also makes immediately available plain text versions of much of its collection.

Third, this proposal does not assume very many things regarding metadata beyond the need for the most basic of bibliographic information such as unique identifiers, titles, authors, subject/keyword terms, and location codes such as URLs. It does not matter to this proposal how the bibliographic metadata is encoded (MARC, XML, linked data, etc.). On the other hand, this proposal will advocate for additional bibliographic metadata, specifically, metadata that is quantitative in nature. These additions are not necessary for the fulfillment of the proposal, but rather side benefits because of it.

Finally, this proposal assumes the code & infrastructure of the DPLA supports the traditional characteristics of a library. In other words, it is assumed the code & infrastructure of the DPLA provide the means for the creation of collections and the discovery of said items. As described later, this proposal is not centered on the processes of find & get. Instead this proposal assumes the services of find & get are already well-established. This proposal is designed to build on the good work of others who have already spent time and effort in this area. We hope to "stand on the shoulders of giants" in this regard.

Given these assumptions about community, content, metadata, and infrastructure, we will now describe how the DPLA can exploit the current technological environment to provide increasingly useful services to its clientele. Through the process we hope to demonstrate how libraries could evolve and continue to play a meaningful role in our society.

Find & get

While it comes across as trite, with the advent of ubiquitous and globally networked computers, the characteristics of data and information have fundamentally changed. More specifically, since things like books and journals – the traditional meat and potatoes of libraries – no longer need to be manifested in analog forms, their digital manifestations lend themselves to new functionality. For example, digital versions of books and journals can be duplicated exactly, and they are much less limited to distinct locations in space and time. Similarly, advances in information retrieval have made strict Boolean logic applied to against relational databases less desirable to the reader than relevancy ranking algorithms and the application of term frequency/inverse document frequency models against indexes. Combined together these things have made the search engines of Google, Yahoo, and Microsoft a reality. Compared to twenty years ago, this has made the problem of find & get much less acute.

While the problem of find & get will never completely be resolved, many readers (not necessarily librarians) feel the problem is addressed simply enough. Enter a few words into a search box, click Go, and select items of interest. We don't know about you, but we can find plenty of data & information. The problem now is what to do with it once it is identified.

We are sure any implementation of the DPLA will include superb functionality for find & get. In fact, our proposal assumes such functionality will exist. Some infrastructure will be created allowing for the identification of relevant content. At the very least this content will be described using metadata and/or the full-text will be mirrored locally. This metadata and/or full-text will be indexed and a search interface applied against it. Search results will probably be returned in any number of ordered lists: relevancy, date, author, title, etc. The interface may very well support functionality based on facets. The results of these searches will never be perfect, but in the eyes of most readers, the results will probably be good enough. This being the case, our proposal is intended to build on this good work and enable the reader to do things with content they identify. Thus we propose to build on the process of find & get to support a process we call use & understand.

Use & understand

The problem of find & get is always a means to an end, and very rarely the end itself. People want to do things with the content they find. We call these things “services against texts”, and they are denoted by action verbs including but not limited to:

> analyze • annotate • cite • compare & contrast • confirm • count & tabulate words, phrases, and ideas • delete • discuss • evaluate • find opposite • find similar • graph & visualize • learn from • plot on a map • plot on a timeline • purchase • rate • read • review • save • share • summarize • tag • trace idea • transform

We ask ourselves, “What services can be provisioned to make the sense of all the content one finds on the Internet or in a library? How can the content of a digital work be ‘read’ in such a way that key facts and concepts become readily apparent? And can this process be applied to an entire corpus and/or a reader’s personal search results?” Thus, we see the problem of find & get evolving into the problem of use & understand.

In our opinion, the answers to these questions lie in the combination of traditional library principles with the application of computer science. Because libraries are expected to know the particular information needs of their constituents, libraries are uniquely positioned to address the problem of use & understand. What do people do with the data and information they find & get from libraries, or for that matter, any other place? In high school and college settings, students are expected to read literature and evaluate it. They are expected to compare & contrast it with similar pieces of literature, extract themes, and observe how authors use language. In a more academic setting scholars and researchers are expected to absorb massive amounts of non-fiction in order to keep abreast of developments in their fields. Each disciplinary corpus is whittled down by peer-review. It is reduced through specialization. Now-a-days the corpus is reduced even further through the recommendation processes of social networking. The resulting volume of content is still considered overwhelming by many. *Use & understand is a next step in the information flow. It comes after find & get, and it is a process enabling the reader to better ask and answer questions of an entire collection, sub-collection, or individual work. By applying digital humanities computing process, specifically text mining and natural language processing, the process of use & understand can be supported by the DPLA.* The examples in the following sections demonstrate and illustrate how this can be done.

Again, libraries are almost always a part of a larger organization, and there is an expectation libraries serve their constituents. Libraries do this in any number ways, one of which is attempting to understanding the “information needs” of the broader organization to provide both just-in-time as well as just-in-case collections and services. We are living, working, and learning in an environment of information abundance, not scarcity. Our production economy has all but migrated to a service economy. One of the fuels of service economies is data and information. As non-profit organizations, libraries are unable to compete when it comes to data provision. Consequently libraries may need to refocus and evolve. By combining its knowledge of the reader with the content of collections, libraries can fill a growing need. Because libraries are expected to understand the particular needs of their particular clientele, libraries are uniquely positioned to fill this niche. Not Google. Not Yahoo. Not Microsoft.

Examples

The following sections describe specific services against texts examples.

Measure size

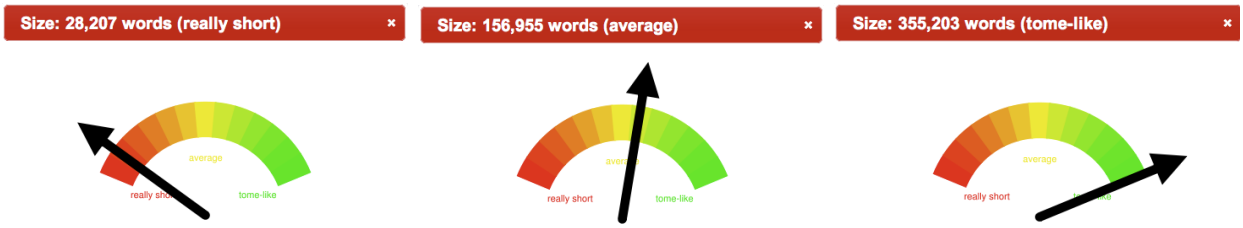
One of the simplest and most rudimentary services against texts the DPLA could provide in order to promote use & understand is to measure the size of documents in terms of word counts in addition to page counts.

Knowing the size of a document is important to the reader because it helps them determine the time necessary to consume the document's content as well as implies the document's depth of elaboration. In general, shorter books require less time to read, and longer books go into greater detail. But denoting the sizes of books in terms of page counts is too ambiguous to denote length. For any given book, a large print addition will contain more pages than the same book in paperback form, which will be different again from its first edition hard cover manifestation.

Not only can much of the ambiguity of document lengths be eliminated if they were denoted with word counts, but if bibliographic descriptions were augmented with word counts then meaningful comparisons between texts could easily be brought to light.

Suppose the DPLA has a collection of one million full-text items. Suppose the number of words in each item were counted and saved in bibliographic records. Thus, search results could then be sorted by length. Once bibliographic records were supplemented with word counts it would be possible to calculate the average length of a book in the collection. Similarly, the range of lengths could be associated with a relative scale such as: tiny books, short books, average length books, long books, and tome-like books. Bibliographic displays could then be augmented with gauge-like graphics to illustrate lengths.

Such was done against the Alex Catalogue of Electronic Texts. There are (only) 14,000 full-text documents in the collection, but after counting all the words in all the documents it was determined that the average length of a document is about 150,000 words. A search was then done against the Catalogue for Charles Dickens's *A Christmas Carol*, *Oliver Twist* and *David Copperfield*, and the lengths of the resulting documents were compared using gauge-like graphics, as illustrated below:



A Christmas Carol *Oliver Twist* *David Copperfield*

At least a couple of conclusions can be quickly drawn from this comparison. *A Christmas Carol* is much shorter than *David Copperfield*, and *Oliver Twist* is an average length document.

There will certainly be difficulties counting the number of words in documents. Things will need to be considered in order to increase accuracy, things like: whether or not the document in question has been processed with optical character recognition, whether or not things like chapter headers are included, whether or not back-of-the-book indexes are included, whether nor not introductory materials are included. All of this also assumes a parsing program can be written which accurately extracts "words" from a document. The later is, in fact, fodder for an entire computer science project.

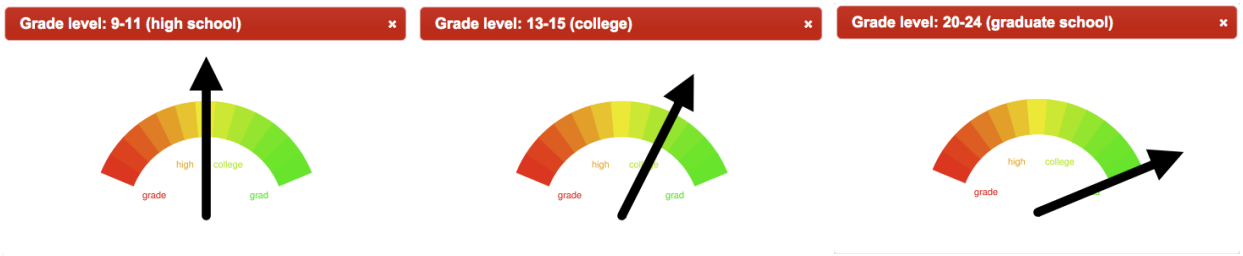
Despite these inherent difficulties, denoting the number of words in a document and placing the result in bibliographic records can help foster use & understand. We believe counting the number of words in a document will result in a greater number of benefits when compared to costs.

Measure difficulty

Measuring the inherent difficulty – readability score – of texts enables the reader to make judgements about those texts, and in turn, fosters use & understand. By including such measurements in the bibliographic records and search results, the DPLA will demonstrate ways it can "save the time of the reader".

In the last century J. Peter Kincaid, Rudolf Flesch, and Robert Gunning worked both independently as well as collaboratively to create models of readability. Based on a set of factors (such as but not limited to: lengths of documents measured in words, the number of paragraphs in documents, the number of sentences in paragraphs, the number of words in sentences, the complexity of words, etc.) numeric values were calculated to determined the reading levels of documents. Using these models things like Dr. Seuss books are consistently determined to be easy to read while things like insurance policies are difficult. Given the full-text of a document in plain text form, it is almost trivial to compute any number of readability scores. The resulting values could be saved in bibliographic records, and these values could be communicated to the reader with the use of gauge-like graphics.

In a rudimentary way, the Alex Catalogue of Electronic texts has implemented this idea. For each item in the Catalogue the Fog, Flesch, and Kincaid readability scores have been calculated and saved to the underlying MyLibrary database. Searches were done against the Catalogue for Charles Dickens's David *Copperfield*, Henry David Thoreau's *Walden*, and Immanual Kant's *Fundamental Principles Of The Metaphysics Of Morals*. The following graphics illustrate the readability scores of each. We believe the results are not surprising, but they are illustrative of this technique's utility:



David Copperfield *Walden* *Metaphysics Of Morals*

If readability scores were integrated into bibliographic search engines ("catalogs"), then it would be possible to limit search results by reading level or even sort search results by them. Imagine being able to search a library catalog for all items dealing with Neo-Platonism, asking for shorter items as opposed to longer items, and limiting things further by readability score.

Readability scores are not intended to be absolute. Instead they are intended to be used as guidelines. If the reader is a novice when it comes to particular topic, and the reader is of high school age, that does not mean they are unable to read college level material. Instead, the readability scores would be used to set the expectations of the reader and help them make judgements before they begin reading a book.

Side bar on quantitative bibliographic data

Bibliographic systems are notoriously qualitative in nature making the process of compare & contrast between bibliographic items very subjective. If there were more quantitative data associated with bibliographic records, then mathematical processes could be applied against collections as a whole, subsets of the collection, or even individual items.

Library catalogs are essentially inventory lists denoting what a library owns (or licenses). For the most part, catalogs are used to describe the physical nature of a library collection: authors, titles, publication dates, pagination and size, notes (such as "Includes index."), and subject terms. Through things like controlled vocabularies and authority lists, the nature of a collection can be posited, and some interesting questions can be answered. Examples include: what is the average age of the items in the collection, what are the collection's major subject areas, who are the predominate authors of the works in the collection. These are questions whose answers are manifested now-a-days through faceted browse interfaces, but they are questions of the collection as a whole or subsets of the collection, not individual works. They are questions librarians find interesting, not necessarily readers who want to evaluate the significance of a given work.

If the bibliographic systems were to contain quantitative data, then the bibliographic information systems would be more meaningful and more useful. Dates are a very good example. The dates (years) in a library catalog denote when the item in hand (a book) was published, not when the idea in the book was manifested. Consequently, if Plato's *Dialogs* were published today, then its library catalog record would have a value of 2011. While such a thing is certainly true, it is misleading. Plato did not write the *Dialogs* this year. They were written more than 2,500 years ago. Given our current environment, why can't a library catalog include this sort of information?

Suppose the reader wanted to read all the works of Henry David Thoreau. Suppose the library catalog had accurately denoted the all the items in its collection by this author with the authority term, "Thoreau, Henry David". Suppose the reader did an author search for "Thoreau, Henry David" and a list of twenty-five items was returned. Finally, suppose the reader wanted to begin by reading Thoreau's oldest work first and progress to his latest. Using a library catalog, such a thing would not be possible because the dates in bibliographic records denote the date of publication, not the date of first conception or manifestation.

Suppose the reader wanted to plot on a timeline when Thoreau's works were published, and the reader wanted to compare this with the complete works of Longfellow or Walt Whitman. Again, such a thing would not be possible because the dates in a library catalog denote publication dates, not when ideas were originally manifested. Why shouldn't a library catalog enable the reader to easily create timelines?

To make things even more complicated, publication dates are regularly denoted as strings, not integers. Examples include: [1701], 186?, 19--, etc. These types of values are ambiguous. Their meaning and interpretation is bound to irregularly implemented "syntactical sugar". Consequently, without all but heroic efforts, it is not easy to do any sort of compare & contrast evaluation when it comes to dates.

The DPLA has the incredible opportunity to make a fresh start when it comes to the definition of library catalogs. We know the DPLA will not want to reinvent the wheel. At the same time we believe the DPLA will want to exploit the current milieu, re-evaluate the possibilities of computer technology, and consequently refine and evolve the meaning of "catalog". Traditional library catalogs were born in an era of relative information scarcity. Today we are dealing with problems of abundance. Library catalogs need to do many things differently in order to satisfy the needs/desires of the current reader. "Next-generation library catalogs" can do so much more than provide access to local collections. Facilitating ways to evaluate collections, sub-collections, or individual items through the use of quantitative analysis is just one example.

Measure concept

By turning a relevancy ranking algorithm on its head, it is be possible to measure the existence of concepts of a given work. If this were done for many works, then new comparisons between works would be possible, and again, making it possible for the reader to easily compare & contrast items in a corpus or search results. Of all the services against texts examples in this proposal, we know this one is the most avant-garde.

Term frequency/inverse document frequency (TFIDF) is a model at the heart of many relevancy ranking algorithms. Mathematically stated, TFIDF equals:

$$(c / t) * \log(d / f)$$

where:

c = number of times the query terms appear in a document
t = total number of words in a document
d = total number of documents in a corpus
f = total number of documents containing the query terms

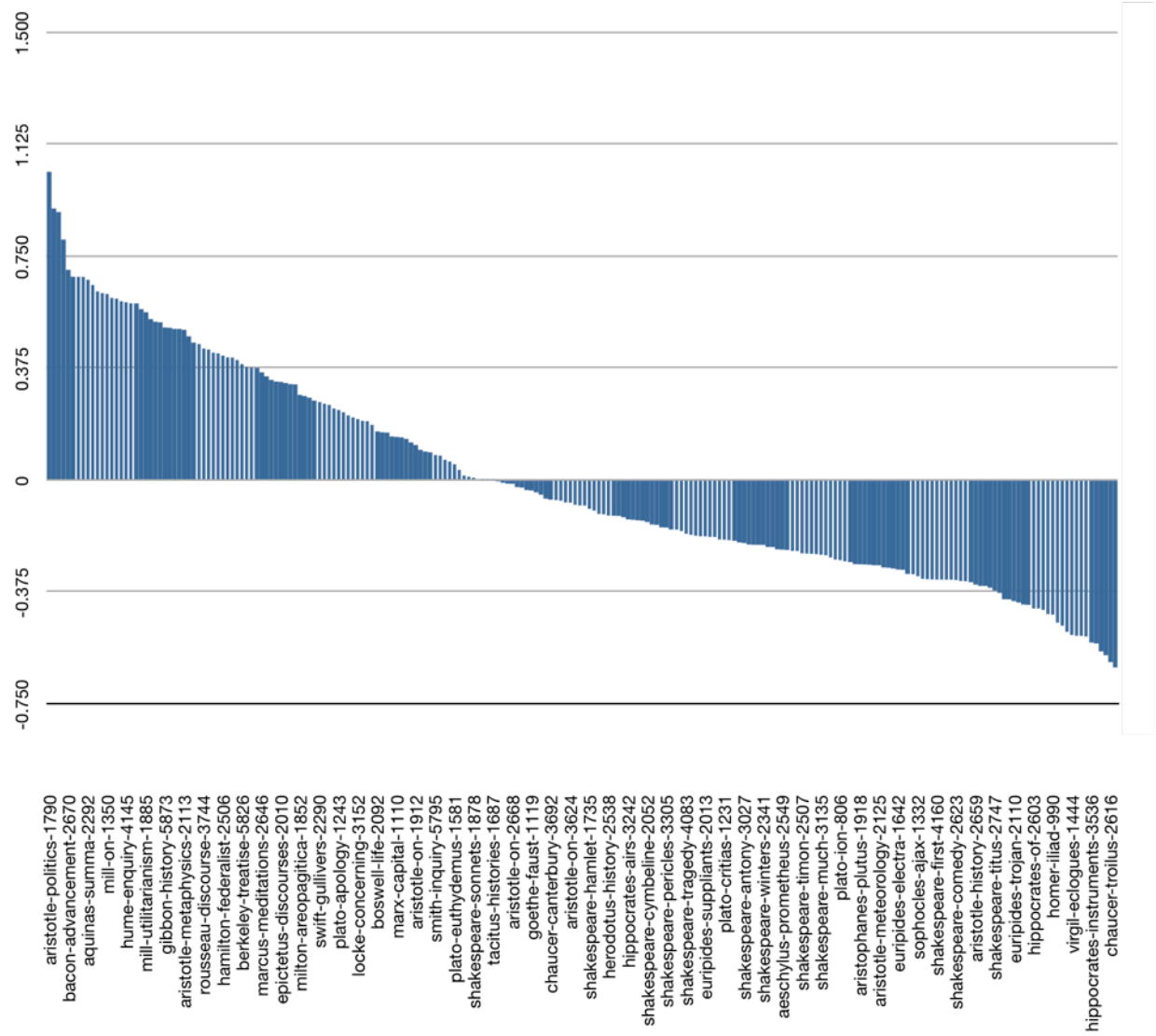
In other words, TFIDF calculates relevancy ("aboutness") by multiplying the ratio of query words and document sizes to the ratio of number of documents in a corpus and total frequency of query terms. Thus, if there are three documents each containing the word "music" three times, but one of them is 100 words long and the other two are 200 words long, then the first document is considered more relevant than the other two.

Written language – which is at the very heart of library content – is ambiguous, nuanced, and dynamic. Few, if any, concepts can be completely denoted by a single word or phrase. Instead, a single concept may be better described using a set of words or phrases. For example, music might be denoted thusly:

art • Bach • Baroque • beat • beauty • blues • composition • concert • dance • expression • guitar • harmony • instrumentation • key • keyboard • melody • Mozart • music • opera • percussion • performance • pitch • recording • rhythm • scale • score • song • sound • time • violin

If any document used some or all of these words with any degree of frequency, then it would probably be safe to say the document was about music. This "aboutness" could then be calculated by summing the TFIDF scores of all the music terms in a given document – a thing called the "document overlap measure". Thus, one document might have a total music "aboutness" measure of 105 whereas another document might have a measure of 55.

We used a process very similar to the one outlined above in an effort to measure the "greatness" of the set of books called *The Great Books Of The Western World*. Each book in the set was evaluated in terms of it use of the 102 "great ideas" enumerated in the set's introduction. We summed the computed TFIDF values of each great idea in each book, a value we call the Great Ideas Coefficient. Through this process we determined the "greatest" book in the set was Aristotle's *Politics* because it alluded to the totality of "great ideas" more than the others. Furthermore, we determined that Shakespeare wrote seven of the top ten books when it comes to the idea of love. The following figure illustrates the result of these comparisons. The bars above the line represent books greater than the hypothetical average great book, and the bars below the line are less great than the others.



Measuring the “greatness” of the *Great Books of the Western World*

The DPLA could implement very similar services against texts in one and/or two ways. First, it could denote any number of themes (like music or “great ideas”) and calculate coefficients denoting the aboutness of those themes for every book in the collection. Readers could then limit their searches by these coefficients or sort their search results accordingly. Find all books with subjects equal to philosophy. Sort the result by the philosophy coefficient.

Second, and possibly better, the DPLA could enable readers to denote their own more specialized and personalized themes. These themes and their aboutness coefficients could then be applied, on-the-fly, to search results. For example, find all books with subject terms equal to gardening, and sort the result by the reader’s personal definition of biology.

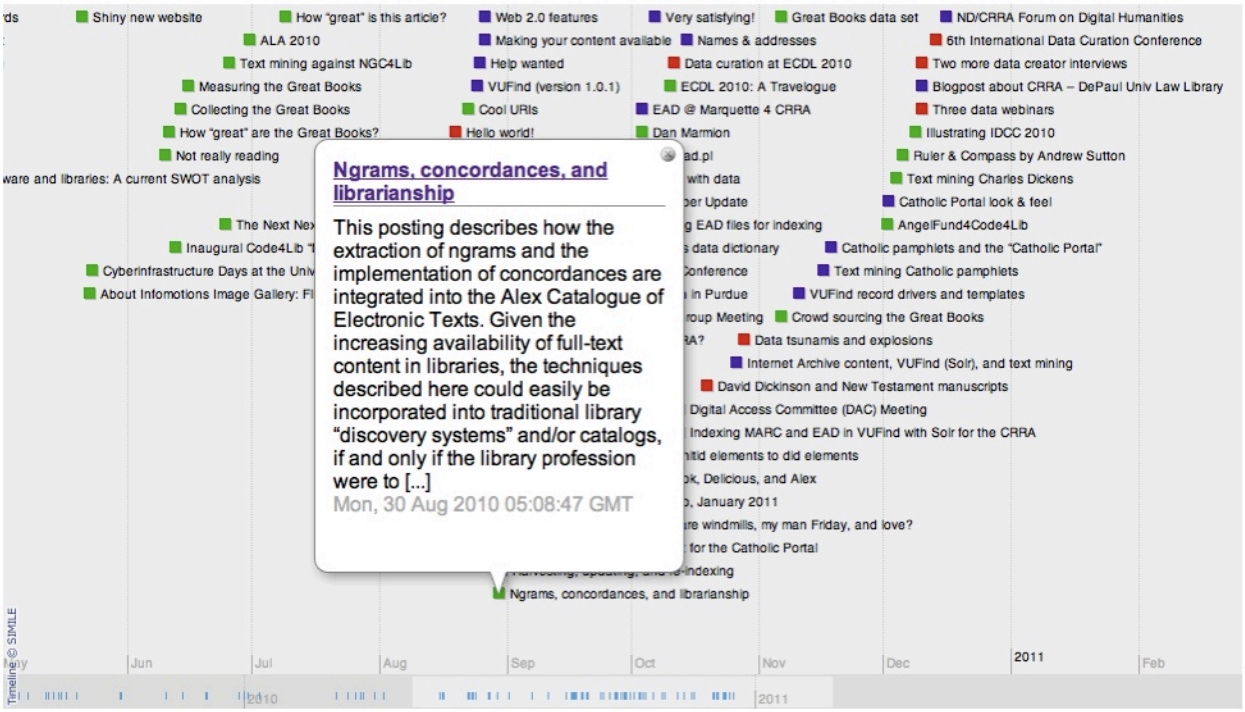
As stated earlier, written language is ambiguous and nuanced, but at the same time it is, to some degree, predicable. If it were not predicable, then no one would be able to understand another. Because of this predicability, language, to some degree, can be quantified. Once quantified, it can be measured. Once measured it can be sorted and graphed, and thus new meanings can be expressed and evaluated. The coefficients described in this section, like the measurements of length and readability, are to be taken with a grain of salt, but they can help the reader use & understand library collections, sub-collections, and individual items.

Plot on a timeline

Plotting things on a timeline is an excellent way to put events into perspective, and when written works are described with dates, then they are amenable to visualizations.

The DPLA could put this idea into practice by applying it against search results. The reader could do a search in the "catalog", and the resulting screen could have a link labeled something like "Plot on a timeline". By clicking the link the dates of search results could be extracted from the underlying metadata, plotted on a timeline, and displayed. At the very least such a function would enable the reader to visualize when things were published and answer rudimentary questions such as: are there clusters of publications, do the publications span a large swath of time, did one particular author publishing things on regular basis?

The dates in traditional bibliographic metadata denote the publication of an item, as mentioned previously. Consequently the mapping of monographs may not be useful as desired. On the other hand, the dates associated with things of a serial nature (blog postings, twitter feeds, journal articles, etc.) are more akin to dates of conception. We imagine the DPLA systematically harvesting, preserving, and indexing freely available and open access serial literature. This content is much more amenable to plotting on a timeline as illustrated below:



Timeline illustrating when serial literature was published

The timeline was created by aggregating selected RSS feeds, parsing out the dates, and plotting them accordingly. Different colored items represent different feeds. Each item in the timeline is hot providing the means to read the items' abstracts and optionally viewing the items' full text.

Plotting things on a timeline is another way the DPLA can build on the good work of find & get and help the reader use & understand.

Count word and phrase frequencies

Akin to traditional back-of-the-book indexes, word and phrase frequency tabulations are one of the simplest and most expedient ways of providing access to and overviews of a text. Like tables of contents and indexes, word and phrase frequencies increase a text's utility and make texts easier to understand.

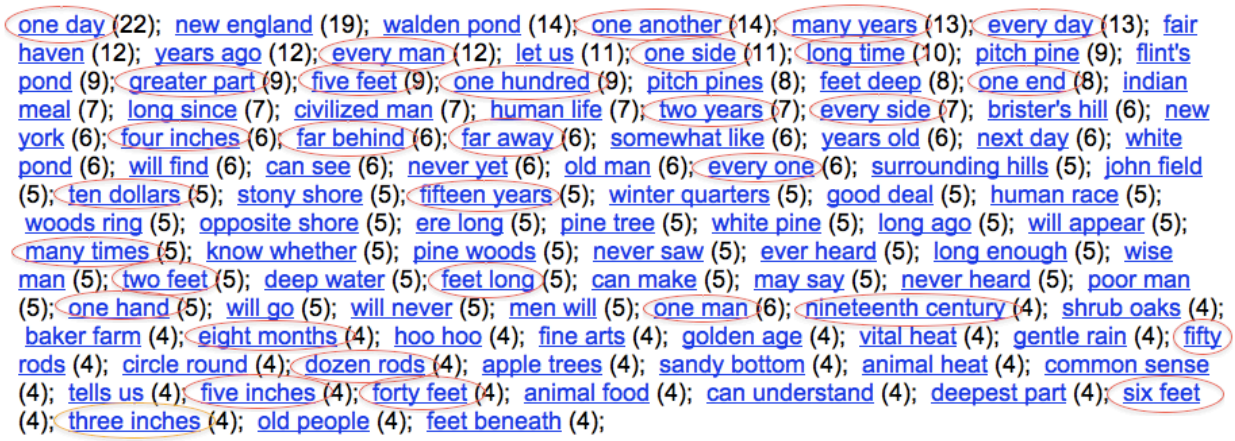
Back-of-the-book indexes are expensive to create and the product of an individual's perspective. Moreover, back-of-the-book indexes are not created for fiction. Why not? Given the full-text of a work any number of back-of-the-book index-like displays could be created to enhance the reader's experience. For example, by simply tabulating the occurrences of every word in a text (sans, maybe, stop words), and then displaying the resulting list alphabetically, the reader can have a more complete back-of-the-book index generated for them without the help of a subjective indexer. The same tabulation could be done again but instead of displaying the content alphabetically, the results could be ordered by frequency as in a word cloud. In either case each

entry in the “index” could be associated with an integer denoting the number of times the word (or phrase) occurs in the text. The word (or phrase) could then be linked to a concordance (see below) in order to display how the word (or phrase) was used in context.

Take for example, Henry David Thoreau's *Walden*. This is a piece of non-fiction about a man who lives alone in the woods by a pond for just about two years. In the book’s introduction Ralph Waldo Emerson describes Thoreau as a man with a keen sense of physical space and an uncanny ability for measurement. The book itself describes one person’s vision of what it means to be human. Upon the creation and display of the 100 most frequently used two-word phrases (bigrams), these statements about the book are born out. Notice the high frequency of quantitative references as well as reference to men.

Compare *Walden* to James Joyce’s *Ulysses*, a fictional work describing a day in the life of Leopold Bloom as he walks through Dublin. Notice how almost every single bigram is associated with the name of a person.

Interesting? Some people may react to these illustrations and say, “So what? I already knew that.” To which we reply, “Yes, but what about those people who haven’t read these texts?” Imagine being able to tabulate the word frequencies against any given set of texts – a novel, a journal article, a piece of non-fiction, all of the works by a given author or in a given genre. The results are able to tell the reader things about the works. For example, it might alert the reader to the central importance of a person named Bloom. When Bloom is mentioned in the text, then maybe the reader ought to be extra attention to what is being said. Frequency tabulations and word cloud can also alert the reader to what is not said in a text. Apparently religion is not a overarching theme in either of the above examples.



The 100 most frequent two-word phrases in *Walden*

The 100 most frequent two-word phrases in *Ulysses*

It is possible to tabulate word frequencies across texts. Again, using *A Christmas Carol*, *Oliver Twist*, and *David Copperfield* as examples, we discover the 6-word phrase "taken with a violent fit of" appears in both *David Copperfield* and *A Christmas Carol*. Moreover, the bigram "violent fit" appears on all three works. Specifically, characters in these three Dickens stories have violent fits of laughter, crying, trembling, and coughing. By concatenating the stories together and applying concordancing methods to them (described below) we see there are quite a number of violent things in the three stories:

```
n such breathless haste and violent agitation, as seemed to betoken so
ood-night, good-night!' The violent agitation of the girl, and the app
sberne) entered the room in violent agitation. 'The man will be taken,
o understand that, from the violent and sanguinary onset of Oliver Twi
one and all, to entertain a violent and deeply-rooted antipathy to goi
eep a little register of my violent attachments, with the date, durati
cal laugh, which threatened violent consequences. 'But, my dear,' said
in general, into a state of violent consternation. I came into the roo
artly to keep pace with the violent current of her own thoughts: soon
ts and wiles have brought a violent death upon the head of one worth m
 There were twenty score of violent deaths in one long minute of that
id the woman, making a more violent effort than before; 'the mother, w
 as it were, by making some violent effort to save himself from fallin
behind. This was rather too violent exercise to last long. When they w
 getting my chin by dint of violent exertion above the rusty nails on
en who seem to have taken a violent fancy to him, whether he will or n
peared, he was taken with a violent fit of trembling. Five minutes, te
, when she was taken with a violent fit of laughter; and after two or
he immediate precursor of a violent fit of crying. Under this impressi
and immediately fell into a violent fit of coughing: which delighted T
```

```
of such repose, fell into a violent flurry, tossing their wild arms ab
 and accompanying them with violent gesticulation, the boy actually th
ght I really must have laid violent hands upon myself, when Miss Mills
 arm tied up, these men lay violent hands upon him – by doing which,
 every aggravation that her violent hate – I love her for it now – c
 work himself into the most violent heats, and deliver the most wither
terics were usually of that violent kind which the patient fights and
 me against the donkey in a violent manner, as if there were any affin
 to keep down by force some violent outbreak. ‘Let me go, will you,--t
hands with me - which was a violent proceeding for him, his usual cour
en.’ ‘Well, sir, there were violent quarrels at first, I assure you,’
revent the escape of such a violent roar, that the abused Mr. Chitling
t gradually resolved into a violent run. After completely exhausting h
, on which he ever showed a violent temper or swore an oath, was this
ullen, rebellious spirit; a violent temper; and an untoward, intractab
fe of Oliver Twist had this violent termination or no. CHAPTER III REL
in, and seemed to presage a violent thunder-storm, when Mr. and Mrs. B
f the theatre, are blind to violent transitions and abrupt impulses of
ming into my house, in this violent way? Do you want to rob me, or to
```

These observations simply beg other questions. Is violence a common theme in Dickens' works? What other adjectives are used to a greater or lesser degree in Dickens' works? How do the use of these adjectives differ from other authors of the same time period or within the canon of English literature?

While works of fiction are the basis of most of the examples, there is no reason why similar processes couldn't be applied to non-fiction as well. We also understand that the general reader will not be interested in these sorts of services against texts. Instead we see these sorts of services more applicable to students in high school and college. We also see these sorts of services being applicable to the scholar or researcher who needs to "read" large numbers of journal article. Finally, we do not advocate the use of these sorts of tools as a replacement for traditional "close" reading. These tools are supplements and additions to the reading process just as tables of contents and back-of-the-book indexes are today.

Display in context

Concordances – one of the oldest literary tools in existence – have got to be some of the more useful services against texts a library could provide because they systematically display words and concepts within the context of the larger written work making it very easy to compare & contrast usage. Originally implemented by Catholic priests as early as 1250 to study religious texts, concordances (sometimes called "key word in context" or KWIC indexes) trivialize the process of seeing how a concept is expressed in a work.

As an example of how concordances can be used to analyze texts, we asked ourselves, "How do Plato, Aristotle, and Shakespeare differ in their definition of man?" To answer this question we amassed all the works of the authors, searched each for the phrase "man is", and displayed the results in a concordance-like fashion. From the results the reader can see how the definitions of Plato and Aristotle are very similar but much different from Shakespeare's

Plato's definitions

```
stice, he is met by the fact that man is a social being, and he tries to harmoni
ption of Not-being to difference. Man is a rational animal, and is not – as man
ss them. Or, as others have said: Man is man because he has the gift of speech;
wise man who happens to be a good man is more than human (daimonion) both in lif
ied with the Protagorean saying, 'Man is the measure of all things;' and of this
```

Aristotle's definitions

```
ronounced by the judgement 'every man is unjust', the same must needs hold good
ts are formed from a residue that man is the most naked in body of all animals a
ated piece at draughts. Now, that man is more of a political animal than bees or
hese vices later. The magnificent man is like an artist; for he can see what is
lement in the essential nature of man is knowledge; the apprehension of animal a
```

Shakespeare's definitions

```
 what I have said against it; for man is a giddy thing, and this is my conclusio
 of man to say what dream it was: man is but an ass, if he go about to expound t
e a raven for a dove? The will of man is by his reason sway'd; And reason says y
n you: let me ask you a question. Man is enemy to virginity; how may we barricad
er, let us dine and never fret: A man is master of his liberty: Time is their ma
```

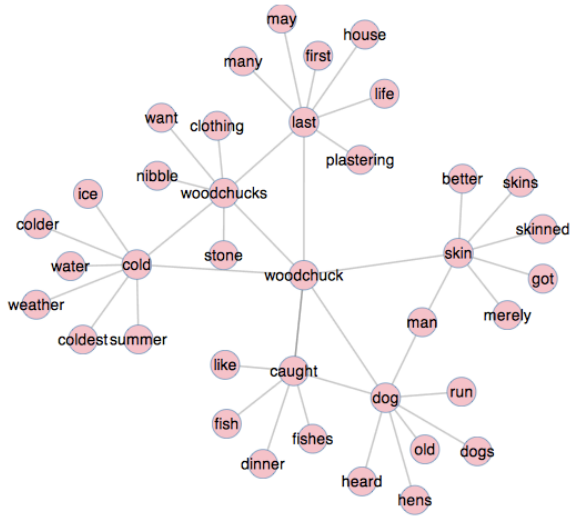
We do not advocate the use of concordances as the be-all and end-all of literary analysis but rather a pointer to bigger questions. Think how much time and energy would have been required if the digitized texts of each of these authors was not available, and if computers could not be applied against them. Concordances, as well as the other services against texts outlined in this proposal, make it easier to ask questions of collections, sub-collections, and individual works. This ease-of-use empowers the reader to absorb, observe, and learn from texts in ways that was not possible previously. We do not advocate these sort of services against texts as replacements for traditional reading processes, but rather we advocate them as alternative and supplemental tools for understanding the human condition or physical environment as manifested in written works.

Herein lies one of the main points of our proposal. *By creatively exploiting the current environment where full-text abounds and computing horsepower is literally at everybody's fingertips, libraries can assist the reader to "read" texts in new and different ways – ways that make it easier to absorb larger amounts of information and ways to understand it from new and additional perspectives.* Concordances are just one example.

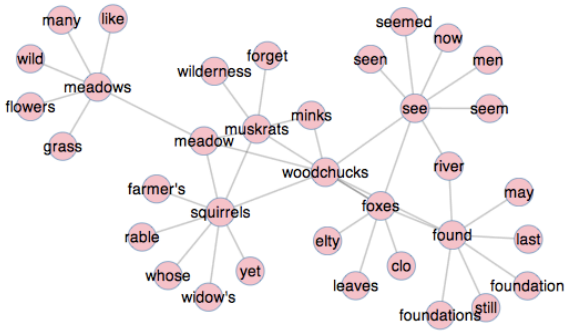
Display the proximity of a given word to other words

Visualizing the words frequently occurring near a given word is often descriptive and revealing. With the availability of full-text content, creating such visualization is almost trivial and have the potential for greatly enhancing the reader's experience. This enhanced reading process is all but impossible when the written word is solely accessible in analog forms, but in a digital form the process is almost easy.

For example, first take the word woodchuck as found in Henry David Thoreau's *Walden*. Upon reading the book the reader learns of his literal distaste for the woodchuck. They eat is beans, and he wants to skin them. Compare the same author's allusions to woodchucks in his work *Two Weeks On The Concord And Merrimack Rivers*. In this work, when woodchucks are mentioned he also alludes to other small animals such as foxes, minks, muskrats, and squirrels. In other words, the connotations surrounding woodchucks and between the two books are different as illustrated by the following network diagrams:



"woodchuck" in *Walden*



"woodchuck" in *Rivers*

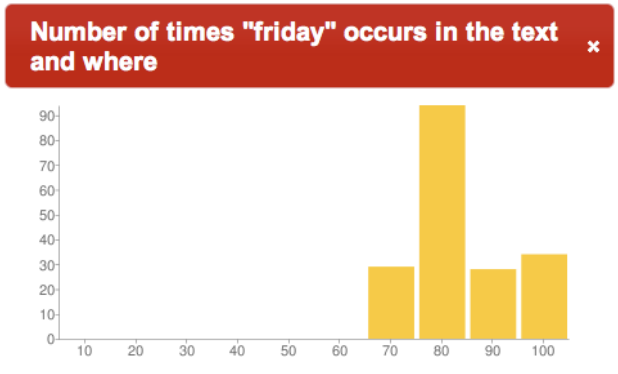
The given word – woodchuck – is in the center. Each of the words connected to the given word are the words appearing most frequently near the given word. This same process is then applied to the connected words. Put another way, these network diagrams literally illustrate what an author says, "in the same breath" when they use a given word. Such visualizations are simply not possible through the process of traditional reading without spending a whole lot of effort. The DPLA could implement the sort of functionality described in this section and make the reader's experience richer. It demonstrates how libraries can go beyond access (a problem that is increasingly not a problem) and move towards use & understand.

We do not advocate the use of this technology to replace traditional analysis, but rather to improve upon it. This technology, like all of the examples in the proposal, makes it easier to find interesting patterns for further investigation.

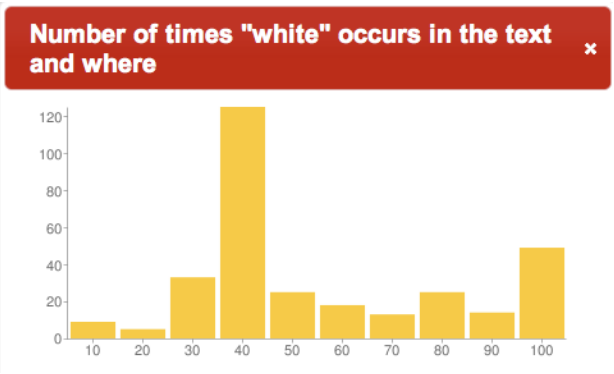
Display location of word in a text

Sometimes displaying where in a text, percentage-wise, a word or phrase exists can raise interesting questions, and by providing tools to do such visualizations the DPLA will foster the ability to more easily ask interesting questions.

For example, what comes to mind when you think of Daniel Defoe's *Robinson Curose*? Do you think of a man shipwrecked on an island and the cannibal named Friday? Ask yourself, when in the story is the man shipwrecked and when does he meet Friday? Early in the story? In the middle? Towards the end? If you guessed early in the story, then you would be wrong because most of the story takes place on a boat, and only three-quarters of the way through the book does Friday appear, as illustrated by the following histogram:



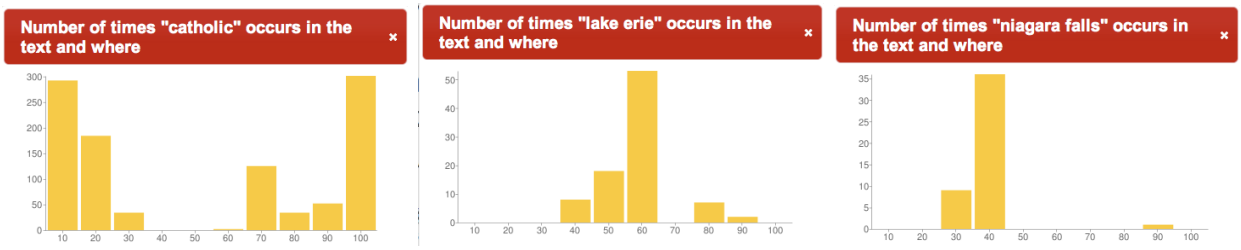
"Friday" in *Curose*



"white" in *Moby Dick*

We all know that Herman Melville's book *Moby Dick* is about a sailor hunting a great white whale. Looking at a histogram of where the word "white" appears in the story, we see a preponderance of its occurrence forty percent the way through the book. Why? Upon looking at the book more closely we see that one of the chapters is entitled "The Whiteness of the Whale", and it is almost entirely about the word "white". This chapter appears about forty percent through the text. Who ever heard of an entire book chapter whose theme was a color?

In a Catholic pamphlet entitled *Letters of an Irish Catholic Layman* the word "catholic" is one of the more common and appears frequently in the text towards the beginning as well as the end:



"catholic" in *Layman* "lake erie" in Layman "niagara falls" in *Layman*

After listing the most common two-word phrases in the book we see that there are many references to places in upper New York state:

de la (271); new york (125); castle bishop (110); upper canada (98); united states (93); lake erie (88); le p (83); le pere (82); separate schools (76); irish people (75); years ago (70); let us (69); sir john (65); roman catholic (58); les peres (57); dans la (53); et de (56); jesus christ (47); via erie (47); precious blood (46); niagara falls (46); et les (48); prize essay (45); et le (47); lake ontario (44); la compagnie (43); lower canada (44); irish catholic (44); great britain (41); par le (41); per cent (40); catholic church (41); au canada (40); dans le (39); de quebec (39); de jesus (38); fran ais (37); annual report (37); goat island (36); et la (37); irish church (35); des jesuites (34); long point (34); one hundred (34); talbot settlement (33); grand river (33); samuel ryerse (32); ninth annual (32); de saint (33); sir charles (32); roman catholics (32); catholic minority (32); du canada (32); catholic review (31); court house (31); les iroquois (31); les jesuites (31); que le (31); home rule (30); remedial bill (30); erie railway (30); quarter sessions (29); state reservation (29); de leur (30); compagnie de (29); colonel talbot (28); la nouvelle (28); s d (28); inclined railway (27); square miles (27); dans les (27); privy council (26); one day (28); la salle (26); took place (26); three years (26); que les (26); kettle creek (25); erie canal (25); la colonie (25); lake huron (25); manitoba act (25); le pays (25); via n (25); par les (25); de ce (25); remedial order (24); sur les (24); de ces (24); many years (24); de montreal (24); pour le (24); dominion government (23); de ans (23); la france (23); catholic schools (23); thomas welch (22); due form (22); land league (22); twenty years (22);

The 100 most frequently used two-word phrases in *Letters of an Irish Catholic Layman*

Looking more closely at the locations of "Lake Erie" and "Niagara Falls" in the text, we see that these things are referenced in the places where the word "catholic" is not mentioned.

Does the author go off on a tangent? Are there no catholics in these areas? The answers to the questions, and the question of why are left up to the reader, but the important point is the ability to quickly "read" the texts in ways that were not feasible when the books were solely in analog form. Displaying where in a text words or phrases occur literally illustrates new ways to view the content of libraries. These are examples of how the DPLA can build on find & get and increase use & understand.

Elaborate upon and visualize parts-of-speech analysis

Written works can be characterized through parts-of-speech analysis. This analysis can be applied to the whole of a library collection, subsets of the collection, or individual works. The DPLA has the opportunity to increase the functionality of a library by enabling the reader to elaborate upon and visualize parts-of-speech analysis. Such a process will facilitate greater use of the collection and improve understanding of it.

Because the English language follows sets of loosely defined rules, it is possible to systematically classify the words and phrases of written works into parts-of-speech. These include but are not

limited to: nouns, pronouns, verbs, adjectives, adverbs, prepositions, punctuation, etc. Once classified, these parts-of-speech can be tabulated and quantitative analysis can begin.

Our own foray's into parts-of-speech analysis, where the relative percentage use of parts-of-speech were compared, proved fruitless. But the investigation inspired other questions whose answers may be more broadly applied. More specifically, students and scholars are often times more interested in what an author says as opposed to how they say it. Such investigations can gleaned not so much from gross parts-of-speech measurements but rather the words used to denote each parts-of-speech. For example, the following table lists the 10 most frequently used pronouns and the number of times they occur in four works. Notice the differences:

Walden	**Rivers**	**Northanger**	**Sense**
I (1,809)	it (1,314)	her (1,554)	her (2,500)
it (1,507)	we (1,101)	I (1,240)	I (1,917)
my (725)	his (834)	she (1,089)	it (1,711)
he (698)	I (756)	it (1,081)	she (1,553)
his (666)	our (677)	you (906)	you (1,158)
they (614)	he (649)	he (539)	he (1,068)
their (452)	their (632)	his (524)	his (1,007)
we (447)	they (632)	they (379)	him (628)
its (351)	its (487)	my (342)	my (598)
who (340)	who (352)	him (278)	they (509)

While the lists are similar, they are characteristic of work from which they came. The first – *Walden* – is about an individual who lives on a lake. Notice the prominence of the word "I" and "my". The second – *Rivers* – is written by the same author as the first but is about brothers who canoe down a river. Notice the higher occurrence of the word "we" and "our". The later two works, both written by Jane Austin, are works with females as central characters. Notice how the words "her" and "she" appear in these lists but not in the former two. It looks as if there are patterns or trends to be measured here.

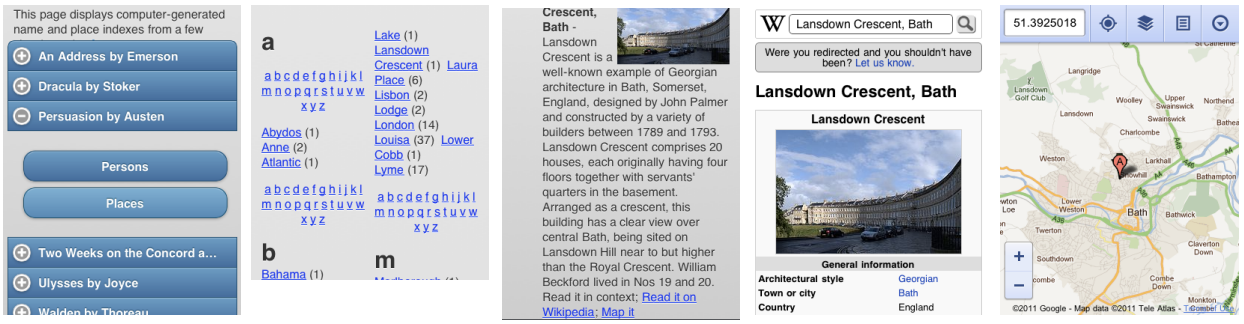
If the implementation of the DPLA were to enable the reader to do this sort of parts-of-speech analysis against search results, then the search results may prove to be more useful.

Nouns and pronouns play a special role in libraries because they are the foundation of controlled vocabularies, authority lists, and many other reference tools. Imagine being able to extract and tabulate all the nouns (things, names, and places) from a text. A word cloud like display would convey a lot of meaning about the text. On the other hand, a simple alphabetical list of the result

could very much function like a back-of-the-book index. Each noun or noun phrase could be associated with any number of functions as such but not limited to:

- look-up in a controlled vocabulary list in order to find more
- look-up in an authority list in order to find more
- show in context of the given work (concordance)
- elaborate upon using a dictionary, thesaurus, encyclopedia, etc.
- plot on a map

We demonstrated the beginnings of the look-up functions in a *Code4Lib Journal* article called "Querying OCLC Web Services for Name, Subject, and ISBN". The concordance functionality is described above. The elaboration service is common place in today's ebook readers. Through an interface designed for mobile devices, we implemented a combination of the elaborate and plot on a map services as a prototype. In this implementation the reader is presented with a tiny collection of classic works. The reader is then given the opportunity to browse the names or places index. After the reader selects a specific name or place the application displays a descriptive paragraph of the selection, an image of the selection, and finally, hypertext links to a Wikipedia article or a Google Maps display.



Screen shots of services against texts on a mobile device

Given the amount of full text content that is expected to be in or linked from the DPLA's collection, there is so much more potential functionality for the reader. The idea of a library being a storehouse of books and journals is rapidly become antiquated. Because content is so readily available on the 'Net, there is a need for libraries to evolve beyond its stereotypical function. By combining a knowledge of what readers do with information with the possibilities for full text analysis, the DPLA will empower the reader to more easily ask and answer questions of texts. And in turn, make it easier for the reader to use & understand what they are reading.

Disclaimer

People may believe the techniques described herein run contrary to the traditional processes of "close" reading. From our point of view, nothing could be further from the truth. We sincerely believe the techniques described in this proposal supplement and enhance the reading process.

We are living in an age where we feel like we are drowning in data and information. But according to Ann Blair this is not a new problem. In her book, *Too Much to Know*, Blair chronicles in great detail the ways scholars since the 3rd Century have dealt with information overload. While they seem obvious in today's world, they were innovations in their time. They included but were not limited to: copying texts (St. Jerome in the 3rd Century), creating concordances (Hugh St. Cher in the 13th Century), and filing wooden "cards" in a "catalog" (Anthanasius Kircher 17th Century).

St. Jerome

Hugh St. Cher

Anthanasius Kercher

Think of all the apparatus associated with a printed book. Books have covers, and sometimes there are dust jackets complete with a description of the book and maybe the author. On the book's spine is the title and publisher. Inside the book there are cover pages, title pages, tables of contents, prefaces & introductions, tables of figures, the chapters themselves complete with chapter headings at the top of every page, footnotes & references & endnotes, epilogues, and an index or two. These extras – tables of contents, chapter headings, indexes, etc. – did not appear in books with the invention of the codex. Instead their existence was established and evolved over time.

In scholarly detail, Blair documents how these extras – as well as standard reference works like dictionaries, encyclopedias, and catalogs – came into being. She asserts the creation of these things became necessary as the number and lengths of books grew. These tools made the process of understanding the content of books easier. They reenforced ideas, and made the process of returning to previously read information faster. Accordingly to Blair, not everybody thought these tools – especially reference works – were a good idea. To paraphrase, "People only need a few good books, and people should read them over and over again. Things like encyclopedias only make the mind weaker since people area not exercising their memories." Despite these claims, reference tools and the apparatus of printed books continue to exist and our venerable "sphere of knowledge" continues to grow.

Nobody can claim understanding of a book if they read only the table of contents, flip through the pages, and glance at the index. Yes, they will have some understanding, but it will only be tertiary. We see the tools described in this proposal akin to tables of contents and back-of-the-book indexes. They are tools to find, get, use, and understand the data, information, and knowledge a book contains. They are a natural evolution considering the existence of books in digital forms. The services against texts described in this proposal enhance and supplement the reading process. They make it easier to compare & contrast the content of single books or an entire corpus. They make it faster and easier to extract pertinent information. Like a back-of-the-book index, they make it easier to ask questions of a text and get answers quickly. The tools described in this proposal are not intended to be end-all and be-all of textual analysis. Instead, they are intended to be pointers to interesting ideas, and it is left up to the reader to flesh out and confirm the ideas after closer reading.

Digital humanities investigations and specifically text mining computing techniques like the ones in this proposal can be viewed as modern-day processes for dealing with and taking advantage of information overload. Digital humanists use computers to evaluate all aspects of human expression. Writing. Music. Theater. Dance. Etc. Text mining is a particular slant on the digital humanities applying this evaluation process against sets of words. We are simply advocating these processes become integrated with library collections and services.

Software

This section lists the software used to create our beta-sprint proposal examples. All of the software is open source or freely accessible. None of the software is one-of-a-kind because each piece could be replaced by something else providing similar functionality.

- **Alex Catalogue of Electronic Texts** (http://infomotions.com/alex/) - This is a collection and full-text index of approximately 14,000 public domain documents from the areas of American and English literature as well as Western philosophy. This "digital library", created and maintained by the author since 1994, is a personal "sandbox" and laboratory for the implementation of new ideas in librarianship.
- **Google Charts** (http://code.google.com/apis/chart/) - Implemented through a Javascript API (application programmer interface), Google Charts enabled us to create the histograms in the "display location of word in a text service". It also provided the gauge-like graphics for the "measure size" and "measure difficulty" services.
- **Google Maps** (http://code.google.com/apis/maps/) - Another Javascript API, Google Maps was a part of the "plot on a map" service.
- **Lingua::Concordance** (http://search.cpan.org/dist/Lingua-Concordance/) - A Perl module, Lingua::Concordance was used to implement the "display in context" service. This module was written by the author.
- **Lingua::EN::Ngram** (http://search.cpan.org/dist/Lingua-EN-Ngram/) - Another Perl module written by the author, Lingua::EN::Ngram was used to count and tabulate the words and n-length phrases in a given text. It plays a crucial role "count word and phrase frequencies" service.
- **Lingua::Fathom** (http://search.cpan.org/dist/Lingua-EN-Fathom/) - This Perl module formed the basis of the "measure size" and "measure difficulty" services since its primary purpose is to calculate Fog, Flesch, and Kincaid readability scores.

- **Lingua::Stem::Snowball** (http://search.cpan.org/dist/Lingua-Stem-Snowball/) - This Perl module plays a role in the "measure concept" service. Given words as input, it outputs the words' roots (or "stems"). These roots were then searched against the index of Alex Catalogue to determine the number of documents (f) containing the root. This value was then used to calculate TFIDF.
- **Lingua::TreeTagger** (http://search.cpan.org/dist/Lingua-TreeTagger/) - This a Perl interface to set of cross-platform binary applications whose purpose is to classify parts-of-speech. Lingua::TreeTagger was used to compare & contrast the ways pronouns were used in four classic works of literature.
- **MyLibrary** (http://search.cpan.org/dist/MyLibrary/) - This is a digital library framework written in Perl. At its core are modules to manage library resources, librarians, and patron descriptions. Inter-relationships between resources, librarians, and patrons can be controlled through the creation and maintenance of facet/term combinations. MyLibrary was co-written by the author and implemented the concept of facets before faceted browse became popular. MyLibrary, in combination with Solr, forms the functional basis of the Alex Catalogue.
- **Protovis** (http://mbostock.github.com/protovis/) - This is the Javascript library used to visualize the "display the proximity of a given word to other words" service.
- **SIMILE Widgets Timeline** (http://www.simile-widgets.org/timeline/) - This is a Javascript library used to display timelines. It was used in the "plot on a timeline" service.
- **Solr** (http://lucene.apache.org/solr/) - Solr is probably the most popular open source indexer in use by the library community, if not else where. It is used to index the full-text of the Alex Catalogue. It was also used to determine the value of f in the "measure concept" service.
- **Stanford Named Entity Recognizer** (http://nlp.stanford.edu/software/CRF-NER.shtml) - This is the set of Java programs used to extract the names and places from a document. These names and places were then linked to Wikipedia or plotted on a map – the "elaborate upon and visualize parts-of-speech" service.

This short list of software can be used to create a myriad of enhanced library services and tools, but the specific pieces of software listed above are not so important in and of themselves. Instead, they represent types of software which already exist and are freely available for use by anybody.

Services against texts facilitating use & understand can be implemented with a wide variety of software applications. The services against texts outlined in this proposal are not limited to the software listed in this section.

Implementation how-to's

Putting into practice the services against text described in this proposal would not be a trivial task, but process is entirely feasible. This section outlines a number of implementation how-to's.

Measurement services

The measurement services (size, readability, and concept) would ideally be done against texts as they were added to the collection. The actual calculation of the size and readability scores are not difficult. All that is needed is the full text of the documents and software to do the counting. (Measuring concepts necessitates additional work since TFIDF requires a knowledge of the collection as a whole; measuring concepts can only be done once the bulk of the collection has been built. Measuring concepts is also a computationally intensive process.)

Instead, the challenge includes denoting locations to store the metadata, deciding whether or not to index the metadata, and figuring out how to display the metadata to the reader. The measurements themselves will be integers or decimal numbers. If MARC were the container for the bibliographic data, then any one of a number of local notes could be used for storage. If a relational database were used, then additional fields could be used. If the DPLA wanted to enable the reader to limit or sort search results by any of the measurements, then the values will need to be indexed. We would be willing to guess the underlying indexer for the DPLA will be Solr, since it seems to be the current favorite. Indexing the measurements in Solr will be as easy as creating the necessary fields to a Solr configuration file, and adding the measurements to the fields as the balance of the bibliographic data is indexed. We would not suggest creating any visualizations of the measurements ahead of time, but rather on-the-fly and only as they were needed; the visualizations could probably be implemented using Javascript and embedded into the DPLA's "catalog".

Timeline services

Like the measurements, plotting the publication dates or dates of conception on a timeline can be implemented using Javascript and embedded into the DPLA's "catalog". For serial literature (blogs, open access journal articles, Twitter feeds, etc.) the addition of meaningful dates will have already been done. For more more traditional library catalog materials (books), the addition of dates of conception will be labor intensive. Therefore such a thing might not be feasible. On the other hand, this might be a great opportunity to practice a bit of crowd sourcing. Consider making a game out

of the process, and try to get people outside the DPLA to denote when Plato, Thoreau, Longfellow, and Whitman wrote their great works.

Frequency, concordance, proximity, and locations in a text services

Implementing the frequency, concordance, proximity, and locations in a text services require no preprocessing. Instead these services can all be implemented on-the-fly by a program linked from the DPLA's "catalog". These services will require a single argument (a unique identifier) and some optional input parameters. Given a unique identifier, the program can look up basic bibliographic information from the catalog including the URL where the full-text resides, retrieve the full-text, and do the necessary processing. This URL could point to the local file system, or, if the network was deemed fast and reliable, the URL could point to the full-text in remote repositories such as the Internet Archive or the HathiTrust. These specific services against texts have been implemented in the Catholic Research Resources Alliance "Catholic Portal" application using "Analyze using text mining techniques" as the linked text. This is illustrated below:

Screen shot of the "Catholic Portal"

By the middle of September 2011 we expect the Hesburgh Libraries at the University of Notre Dame will have included very similar links in their catalog and "discovery system". These links will provide access to frequency, concordance, and locations in a text services for sets of digitized Catholic pamphlets.

Parts-of-speech services

Based on our experience, the parts-of-speech services will require pre-processing. This is because the process of classifying words into categories of parts-of-speech is a time- and computing-intensive process. It does not seem feasible to extract the parts-of-speech from a document in real time.

To overcome this limitation, we classified our small sample of texts and saved the result in easily parsable text files. Our various scripts were then applied against these surrogates as opposed to the original documents. It should be noted that these surrogates, while not only computationally expensive, were also expensive in terms of disk space consuming more than double the space of the original.

We suggest one or two alternative strategies for the DPLA. First, determine what particular items from the DPLA's collection may be the more popular. Once determined, have those items pre-processed outputting the surrogate files. These pre-processed items can then be used for demonstration purposes and generate interest in the parts-of-speech services. Second, when readers want to use these services against items that have not been pre-processed, then have the readers select their items, supply an email address, process the content, and notify the readers when the surrogates have been created. This second approach is akin to the just-in-time approach to collection development as opposed to the just-in-case philosophy.

Priorities

Obviously, we think all of the services against texts outlined above are useful, but practically speaking, it is not feasible to implement all of them once. Instead we advocate the following phased approach:

1. **Word/phrase frequency, concordance, proximity, and locations in a text services** - We suggest these services be implemented first, mostly because they can be written outside any "discovery system" hosted by the DPLA. Second, these services are the root of many of the other services, so it will be easier to build the others once these have been made available.
2. **Measurements of size and readability** - Calculating the values of size and readability on-the-fly is possible but is limiting in functionality. Pre-processing these values is relatively easy, and incorporating the result into the "discovery system" has many benefits. This is why we see these two services as the second highest priority.
3. **Plot dates of publication on a timeline** - Plotting dates will be easy enough if the content in question is of a serial nature and the dates represent "dates of conception". But we are not sure content of a serial nature (blog postings, open

access journal literature, Twitter feeds, etc.) will be included in the DPLA's collection. Consequently, we suggest this service be implemented third.

4. **Parts-of-speech analysis** - Implementing services based on parts-of-speech will almost certainly require pre-processing as increase local storage requirements. While these costs are within the DPLA's control, they are expenses that may inhibit implementation feasibility. That is why they are listed fourth in the priority order.

5. **After crowd sourcing the content, plot dates of conception on a timeline** - We think this is one of the easier and more interesting services, especially if the dates in question are "dates of conception" for books, but alas, this data is not readily available. After figuring out how to acquire dates of conception for traditional catalog-like material – through something like crowd sourcing – implementing this service my be very enlightening.

6. **Measure ideas** - This is probably the most avant-garde service described in the proposal. Its implementation can only be done after the bulk of the DPLA's collection has been created. Furthermore, calculating TFIDF for a set of related keyword is computationally expensive. This can be a truly useful and innovative service, especially if the reader were able to create a personal concept for comparison. But because of the time and expense, we advocate this service be implemented last.

Quick links

This section lists most of the services outlined in the proposal as well as links to blog postings and example implementations.

Word frequencies, concordances

These URLs point to services generating word frequencies, concordances, histograms illustrating word locations, and network diagrams illustrating word proximities for *Walden* and *Ulysses*.

- http://infomotions.com/sandbox/concordance/?id=thoreau-walden-186
- http://infomotions.com/sandbox/concordance/?id=etext4300

Word/phrase locations

Using the text mining techniques built into the "Catholic Portal" the reader can see where the words/phrases "catholic", "lake erie", and "niagara falls" are used in the text.

- http://www.catholicresearch.net/concordances/?id=tormarc_lettersofirishca00iris

Proximity displays

Using network diagrams, the reader can see what words Thoreau uses "in the same breath" when he mentions the word "woodchuck". These proximity displays are also incorporated into just about every item in the Alex Catalogue

- http://infomotions.com/blog/2011/01/visualizing-co-occurrences-with-protovis/

Plato, Aristotle, and Shakespeare

This blog posting first tabulates the most frequently used words by the authors, as well as their definitions of "man" and a "good man".

- http://infomotions.com/blog/2010/06/the-next-next-generation-library-catalog/

Catholic Portal

The "Portal" is collection of rare, uncommon, and infrequently held materials brought together to facilitate Catholic studies. It includes some full text materials, and they are linked to text mining services.

- http://www.catholicresearch.net/Record/tormarc_lettersofirishca00iris

Measuring size

In this blog posting a few works by Charles Dickens are compared & contrasted. The comparisons include size and word/phrase usage.

- http://infomotions.com/blog/2010/12/text-mining-charles-dickens/

Plot on a timeline

This blog posting describes how a timeline was created by plotting the publication dates of RSS feeds.

- http://infomotions.com/blog/2010/12/mits-simile-timeline-widget/

Lookup in Wikipedia and plot on a map

After extracting the names and places from a text, this service grabs Linked Data from DBedia, displays a descriptive paragraph, and allows the reader to look the name or place up in Wikipedia and/or plot it on a world map. This service is specifically designed for mobile devices.

- http://dh.crc.nd.edu/sandbox/ner/mobile.html

Parts-of-speech analysis

This blog posting elaborates on how various parts of speech were used in a number of selected classic works.

- http://infomotions.com/blog/2011/02/forays-into-parts-of-speech/

Measuring ideas

The "greatness" of the Great Books was evaluated in a number of blog postings, and the two listed here give a good overview of the methodology.

- http://infomotions.com/blog/2011/03/how-great/
- http://infomotions.com/blog/2010/06/measuring-the-great-books/

Summary

In our mind, the combination of digital humanities computing techniques – like all the services against texts outlined above – and the practices of librarianship would be a marriage made in heaven. By supplementing the DPLA's collections with full text materials and then enhancing its systems to facilitate text mining and natural language processing, the DPLA can not only make it easier for readers to find data and information, but it can also make that data and information easier to use & understand.

We know the ideas outlined in this proposal are not typical library functions. But we also apprehend the need to take into account the changing nature of the information landscape. Digital content lends itself to a myriad of new possibilities. We are not saying analog forms of books and journals are antiquated nor useless. No, far from it. Instead, we believe the library profession has figured out pretty well how to exploit and take advantage of that medium and its metadata. On the other hand, the possibilities for full text digital content are still mostly unexplored and represent a vast untapped potential. Building on and expanding on the education mission of libraries, services against texts may be a niche the profession – and the DPLA – can help fill. The services & tools described in this proposal are really only examples. Any number of additional services against texts could be implemented. We are only limited by our ability to think of action words denoting the things people want to do with texts once they find & get them. By augmenting a library's traditional functions surrounding collection and services with the sorts of things described above, the role of libraries can expand and evolve to include use & understand.

About the author

Eric Lease Morgan considers himself to be a librarian first and a computer user second. His professional goal is to discover new ways to use computers to provide better library service. He has a BA in Philosophy from Bethany College in West Virginia (1982), and an MIS from Drexel University in Philadelphia (1987).

While he has been a practicing librarian for more than twenty years he has been writing software for more than thirty. He wrote his first library catalog in 1989, and it won him an award from Computers in Libraries Magazine. In a reaction to the "serials pricing crisis" he implemented the Mr. Serials Process to collect, organize, archive, index, and disseminate electronic journals. For these efforts he was awarded the Bowker/Ulrich's Serials Librarianship Award in 2002. An advocate of open source software and open access publishing since before the phrases were coined, just about all of his software and publications are freely available online. One of his first pieces of open source software was a database-driven application called MyLibrary, a term which has become a part of the library vernacular.

As a member of the LITA/ALA Top Technology Trends panel for more than ten years, as well as the owner/moderator of a number of library-related mailing lists (Code4Lib, NGC4Lib, and Usability4Lib), Eric has his fingers on the pulse of the library profession. He coined the phrase "'next-generation' library catalog". More recently, Eric has been applying text mining and other digital humanities computing techniques to his Alex Catalogue of Electronic Texts which he has been maintaining since 1994. Eric relishes all aspects of librarianship. He even makes and binds his own books. In his spare time, Eric plays blues guitar and Baroque recorder. He also enjoys folding origami, photography, growing roses, and fishing.