Think "Different"

by Karen Coyle

Keynote, Dublin Core, 2012 and Emtacl12

"Think Different" was an Apple company advertising campaign, which may be familiar to you. It caused a bit of a scandal in the U.S. because "think different" is grammatically incorrect – it should be "think differently," with *differently* being the proper adverbial form. Around the time of the death of Steve Jobs, a man who really did think different, I read that 1) he knew his grammar and 2) his intention was to use different as a noun. So the message was: what ever you would usually think, you would normally think, instead think of something different – a different way of looking at things, a different style, a different assumption about how things work.

This is essentially the same concept as the saying "think outside the box," except that latter has become an overused cliché, and I must say that it is often not clear to me where the box is that is being thought outside of.

I'm going to explain what I mean with some examples, because the concept is hard to define in words.

Case 1: The Open Library and Alphabetical Order

The Internet Archive has a bibliographic catalog called the Open Library. It is a large database with about 25 million bibliographic records, and it is also the access to the over one million digitized books that the Archive holds and provides access to. When they were beginning to create the catalog they realized that they needed some sources of bibliographic data. They took in some

data from Amazon, but they also obtained library bibliographic data, from the Library of Congress and from other libraries.

None of the Archive staff members had any prior experience with library data and at first they thought: no problem. Then they started looking at the data and decided: "ok, problem." So I worked with them as a kind of translator between their goals and the data in library catalog records.

There were some significant differences in the goals between the Open Library and most library catalogs, a big one being that the Open Library would be wiki-like and anyone, really anyone, could edit the data, just like they can in Wikipedia. This meant that some of the particular details of library data had to be smoothed out so that the input and editing could use simple "fill in the box" forms. Some aspects of library data just wouldn't fit into this model.

This seemed to be working just fine, and then we ran into the "alphabetical order" problem.

The catcher in the rye

Many titles have extra words at the beginning that get in the way of an expected alphabetical order. Most users would look for this work under "C" not "T," and the number of entries under "The" in a large library catalog would be enormous. So library data has various was to indicate what part of the title is to be ignored for the purposes of putting titles in alphabetical order.

There are some systems that use marks to indicate the non-filing part:

/The /catcher in the rye

Others, like MARC, make the cataloger indicate the number of bytes that must be ignored when sorting:

[4] The catcher in the rye

The rules for what entails an initial article for filing purposes differs from language to language, so explaining to users how to input this data was going to be very complex, and in an open database with potentially any internet user being an editor, this just was not going to work.

So the Open Library folks did something absolutely different – they asked: why do we need alphabetical order?

When I told this story to a large auditorium of U.S. librarians, they howled with laughter. What do you mean "Why do we need alphabetical order?" Of course you need alphabetical order. Alphabetical order is the very basis of library cataloging; it permeates our cataloging rules, even the new rules that presumably are being designed, at least in part, for the Semantic Web.

Well, Google, the most popular place to search for information on the entire planet, doesn't present results in alphabetical order. Nor does Amazon. Even OCLC's Worldcat, the world's largest database of library bibliographic data, does not present in alphabetical order as
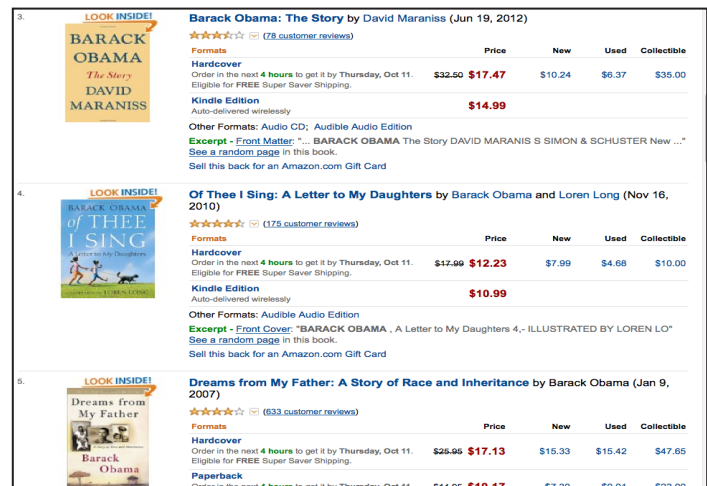

WorldCat results for "Barack Obama"

Google results for "public library"

Amazon results for "Barack Obama"

its default. And even that absolutely classic alphabetical list, the telephone book, is now accessed as a search, and in some cases does not return its results in alphabetical order.

99 Results for Fitzgerald in Chicago, IL

Patricia Fitzgerald   **S Bell Ave**
(Age 60-64)   **Chicago, IL**

Robert W Fitzgerald   **S Bell Ave**
(Age 60-64)   **Chicago, IL**

Pearlie Fitzgerald   **W Jackson Blvd**
(Age 65+)   **Chicago, IL**

Danyel C Fitzgerald   **S Loomis St**
(Age 30-34)   **Chicago, IL**

A "white pages" search online

And yet, when librarians create bibliographic data they are following rules that not only assume but actually dictate that the key decisions that are made about the metadata must be done in support of alphabetical order.

It's true that alphabetical order was once the only discovery mechanism in the library catalog.

It was the discovery mechanism because no other technology of discovery was available in the analog card catalog. For the last 50 years we have had database management systems that can reach into any part of the library data record and find any text string or combination of strings within the record. We no longer use, and should of course no longer design our data for, the linear retrieval of the past.

What is the nature of alphabetical order in terms of knowledge organization? Here is a set of things. What is the meaning of this set of objects?

They don't have any obvious conceptual or semantic relationship between them. They are, however, in alphabetical order.

Apple   Book   Candle   Donkey

A   B   C   D

Alphabetical order is not about knowledge, it's about words. It's not conceptual, it's not semantic. It is an accident of language that this first object, an apple, was given the word "apple" that precedes the word of the second object, "book." There is no meaning that would place apples before books.

Alphabetical order is an accident of language, and different languages have different accidents. This makes alphabetical order a very poor discovery method in a multilingual environment.

However, I can show you various groupings of things that, even if you haven't seen them before, probably make sense to you.

This is not a quirk; cognitive scientists can explain, at least to some extent, why our brains work well with concepts. And yet, in spite of this evidence, and in spite of what we know about "modern technology" (that is, technology that has been around for 50 years) the most recently developed library cataloging rules still have within them in key areas the assumption of the predominance language terms in alphabetical order. This is a failure to think clearly about a problem that needs solving because you have a default solution that cannot be questioned, a solution with a long history that has been successful in the past. This is an inability to "think different."

Case 2: The Book

The book is such an icon in our culture: it represents learning, it represents libraries, it represents religious beliefs. We all know what the book is: it has pages, usually paper with printing on it, it has a title page, chapters with headings, a binding that holds it all together, and page numbers.
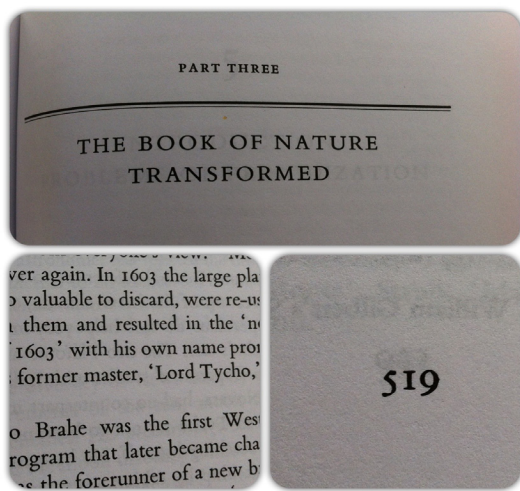
For over 500 years we've had this fairly standardized object that we know so very well, and then, suddenly, it changes: the e-book is invented.

It goes from being a mass-produced object with all instances being the same to a display of continuous, fluid text with no fixity of the meaning of "page." If the change the font, the amount of text that is display on the virtual page changes. If your device is a different size or shape from someone else's, what is on your page is different from what is on theirs. Page numbers are meaningless in this environment. Some educational institutions did experiments with ebooks for their classes, but they ended up rejecting them because it was too hard, in the classroom, to get everyone on the "same page." Where once a professor could instruct his class to "turn to page 87," there is no equivalent in the ebook. Using percentages doesn't work because the point measured as "17%" of a large book could cover a great deal of text.

Unsolvable problem? Hardly. In fact, this problem was solved before the advent of printing.
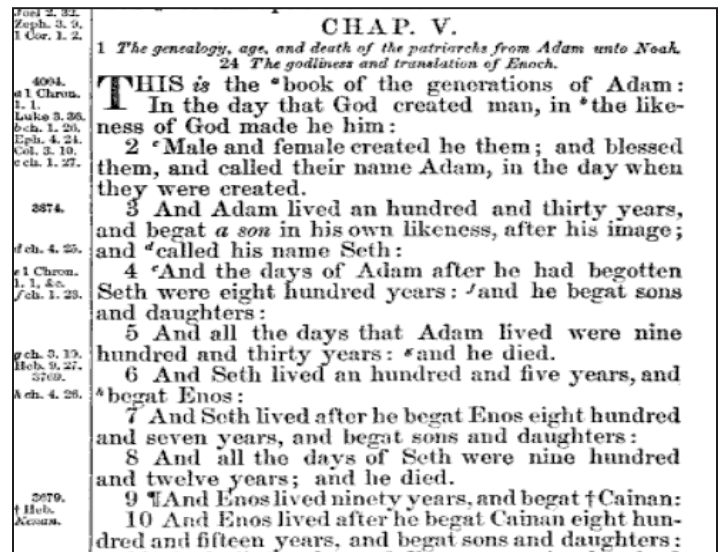
When books were in manuscript form there was no concept of page numbers because each copy was unique. Page numbers came into use only with the mass production of books through printing. As a book that decidedly preceded printing, the Bible has developed a highly effective way of numbering chapters and verses such than anyone with a different copy can refer to exactly the same portion of text.

To give a more modern example, Ludwig Wittgenstein, admittedly not your average author, wrote his philosophical works with numbers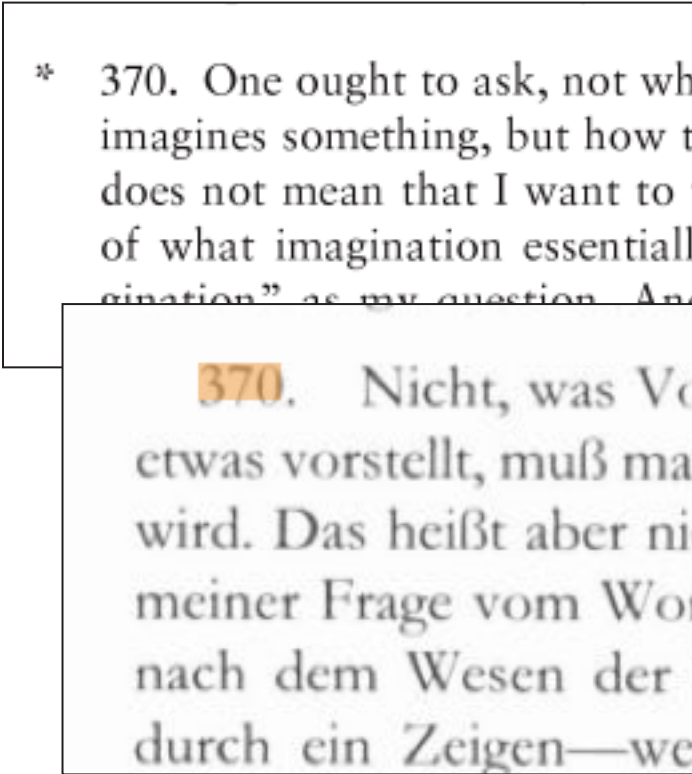 on each paragraph. This means that persons reading his work, even in different translations, can reference the same precise point in his books.

Wittgenstein's paragraph numbering

In a printed book these numbers have to be visible and they are admittedly unattractive and distracting. In ebooks these place markers can be hidden until needed, as appears to be the Kindle approach. The ebook software could make it easy to cite individual points in the book, and perhaps even ranges.

I'll give one more example, which is a bit odd but definitely a case of thinking different.

There is a group of people who have taken upon themselves to translate some classical works that are in the public domain to series of QR code barcodes. They call their project "Books to Barcodes." Each QR code represents some amount of text, usually a few sentences or maybe a short paragraph. This is obviously not an ideal way to read a book like *Pride and Prejudice*, but it does demonstrate that snippets of text can be turned into a machine-readable format that can be decoded by a variety of modern devices. This doesn't in itself give you location between texts on different devices or in different formats, but this odd project may provide a clue to how we can solve this problem in a cross-platform way.

Books as QR codes

Thinking "Different" About Libraries

The classical library is a thing of great beauty. The wooden shelves, leather-bound books and the cathedral-like atmosphere of contemplation, the distant ceiling that gives one a sense of reaching upward, transcending – who of us who love books and learning wouldn't want to study here?

This model is still at the heart of most modern libraries. It may be the case that the library as conceived even many thousands of years ago is so right that it is still relevant. However, the technological changes over the last century, and in particular those that have led to a vastly larger variety of information media, must certainly mean that the "books on a shelf" model is no longer the most relevant one.

Today's media have an emphasis on interactivity that was inconceivable in the past. Books have been called a "slow conversation" between authors and readers, and some readers go on to extend that conversation in their own publications. In the traditional book world it is not easy to see how the books interact. There are some overt connections, such as footnotes, but that is only part of the story. Technology today can help us see other connections, such as when documents share readers.

We can also learn about interaction between documents through their proximity in bibliographies, syllabi, and perhaps even the shelves they share. An aspect of the conversation is knowing what books an author or reader had available in her time, which gives us a context for that person's part of the conversation. And we know that books are active when they are read, and that the reader is as important as the book itself.

But libraries insist on treating books as objects; objects to be organized, not as knowledge.

In fact, libraries are very "thing" oriented, and the library view of its collection is that of things that are owned and must be managed and controlled. Library cataloging is all about the physical format, and placing those things in a linear order. Retrieval is primarily language-based, requiring that the user be able to name what she is looking for.

This concept of a controlled information world is out of date, as is the concept that information comes in a tangible form, pre-packaged. While print dominated for a significant amount of time (from about 1450 to the late 1800's), our information has been in less tangible formats since the invention of the telegraph in 1844. Telephone, radio, television, and now the Internet have taken over the book's previous monopoly on information. In fact, something that we couldn't even imagine just a few years ago is how the telephone has combined with the Internet to become a multi-media information and conversation machine.

Yet in libraries we are still putting considerable effort into the linear arrangement of books on a shelf. It is a myth that the library's shelves are meaningful to the user. We tell users that when they go to the shelf looking for one book they will find others on the same topic. That may happen sometimes, but the limitation of having only one shelf location means that some books are also kept far from each other. And what is the user to make of those numbers on the spines of the books? At no point does the library show users what those numbers mean. And they are meaningful – in fact they have a whole knowledge structure in them, which is seen by the one librarian who assigns the number, and no one else. It's a secret code. How can that help people find things?

We have to move libraries from organizing things to knowledge discovery. This means that the library can be only one part of the information picture since a great deal of information is outside of the library. It means emphasizing the relationships between things, not their place in a linear order. It also means allowing interoperability and that means treating library users as contributors to the knowledge universe, not just as consumers. It means, yes, giving up control, and that is going to be the hardest thing for librarians to do.

Libraries have been slow to augment their catalogs with more information beyond the bare catalog record. WorldCat is now providing more links to resources outside the catalog itself. Amazon still provides much more information, and more interaction, than the library catalog which still mostly does not let the user contribute at all.

card

Worldcat

Amazon

However, the length of the page is not a good measure of how useful the catalog is. One problem that I see is that all of these still focus on a single item; they are still very "thing" based. For Amazon that is a function of its purpose, which is to sell things. The library could take a different view. Some library catalogs are adding topic maps and various facets, but the focus is still on the individual things, not relationships between them. And in fact it is easier to organize things than it is to manage a complex of relationships, but that's not our mission. In fact, I have a new mission statement for libraries:

> The mission of the library is not to gather physical things into an inventory, but to **organize human knowledge that has been very inconveniently packaged.**

To do this, we must un-package our data so that any information can combine with any other information, and let the user of the data determine what is the focus, and what relationships are meaningful.

It should be possible to ask a question of the library catalog, and to get an answer, not just a list of items.

> "What were the most popular book subjects from 1830-1840?"
>
> (WorldCat answer, using kw:history because a search on date alone isn't allowed: **141,076**)

Presenting results as a list of items no longer works because library holdings have gone beyond human scale. A search on the term "history" and limited to books published from 1830-1840 retrieves over 140,000 records. Even with a few hundred it would be hard for a human to look at these records and determine what the most popular subjects were in that time period. We need to apply computation to perform tasks that humans cannot do on their own. That's what computers are good at.

An example of computation on library data is WorldCat identities, which shows timelines that give the user a quick snapshot of the author.
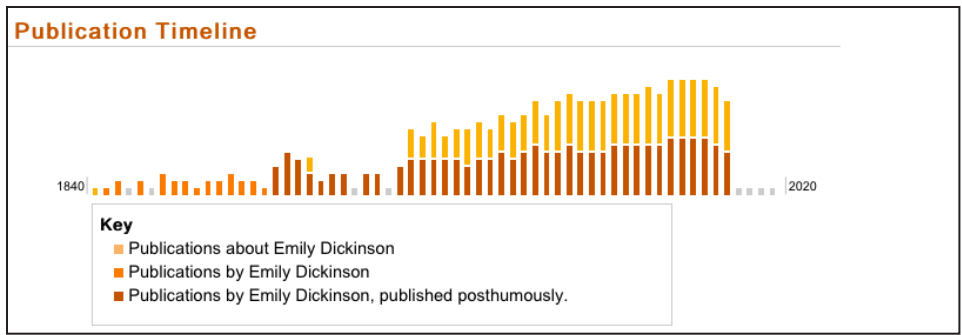
Another example is on the subject pages in OpenLibrary, where you can see that the term "human evolution" appears first in 1859, with Darwin, but gains ground only after about 1960. The topic "love" however has been written about at least since the beginning of printing, and undoubtedly even before. This is useful information in itself, and this is just using the library metadata.
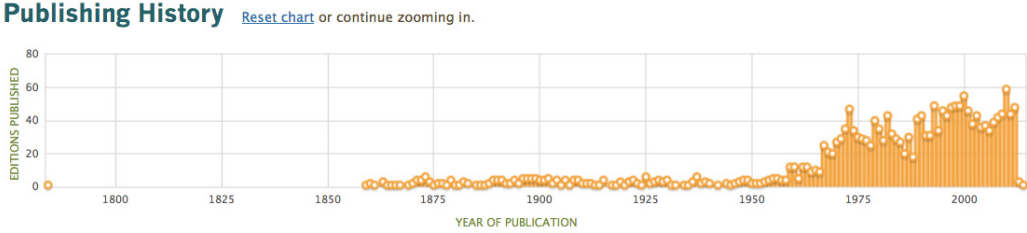
Timeline for subject "human evolution"
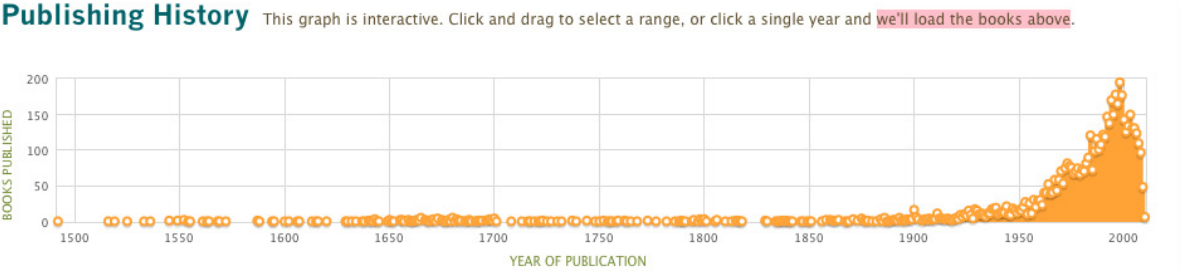
Timeline for subject "love"

However, library metadata itself has some serious problems as data, as exhibited by a record for "The origin of species" by Darwin, that has the publication date 2009. Nothing in the data tells us that the text is that of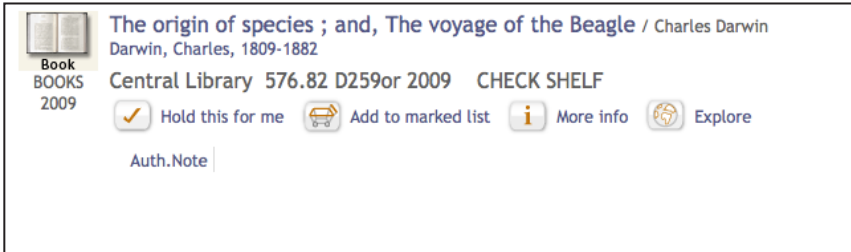 1859, because the emphasis is on the package, not on the content. And this means that the library user misses some key context, which is how the slow conversation of books has taken place. Darwin makes little sense if you think his ideas are from 2009.

The origin of species ; and, The voyage of the Beagle / Charles Darwin
Darwin, Charles, 1809-1882
Central Library 576.82 D259or 2009   CHECK SHELF
✓ Hold this for me   Add to marked list   i More info   Explore
Auth.Note

In fact, this view greatly interrupts the conversation about science and evolution.

New Dimensions

We need to add some new dimensions to the library. The library today is 2-D, relying heavily on linear order -- linear order of shelves, linear order of headings in the catalog, and linear order of catalog records that are retrieved with searches.

The first dimension that we need to add is linking; linking between items in the library based on any aspect of the concepts they contain; linking from the library to information outside of the library; and linking from the main information resource in the users' environment, the Web, to the library. With links, the library can become 3-D.

The fourth dimension is time. It needs to be possible to follow thoughts and ideas in the library as the develop over time. This includes knowing what works influenced an author or inventor and what new discoveries that person may have contributed to. Readers should be able to reconstruct the context of information that they encounter, so that they can better understand the world that knowledge addressed.

The fifth dimension is people. What is in the library was created by people, and people will use the services and materials of the library in unpredictable ways. People will understand and create new knowledge using the library. That knowledge may be totally new to the world or just new for that user. It will combine thoughts from the person's life and information previously encountered.

Addendum: Is Linked Data the Answer?

I have beena proponent of exploration into the use of linked data for libraries for a number of years. Therefore you might expect that I would say: Yes, linked data is the answer!

Instead, I want to insert here a word of caution. No, I'm not going to declare that we should abandon ideas of making use of linked data, but I do want to caution against having "an answer" to the issues and problems that we face. When you have "an answer" you tend to stop looking for new ideas and new solutions. You also tend to only consider problems that the "answer" can address.

It is unreasonable, and even dangeous, to think that any one technology will be the solution to every process and service that we want to provide. So although there is much to be gained by using linked data to create a web that connects libraries to other information resources, we have more to do than simply linking.

One strong assumption in the library field is that what we have to contribute to the Web and the greater information world is the contents of our bibliographic databases. Yet there would be little to be gained by flooding the Web with hundreds of millions of records, most of which already exist. In fact, the Web is awash in bibliographic data, from booksellers like Amazon and Barnes and Noble, to Google Books, which is a mix of actual books and bibliographic data, to book fan sites like LibraryThing and GoodReads. Although some of the data in library catalogs may be rare or unique, most of what we would contribute would be duplication of data that already exists.

What libraries do have, however, that no one else does is that we know where the user can borrow or use materials in her nearby community. It is library holdings that is key to providing service and furthering knowledge creation. It is also key to providing visibility for libraries as users explore resources on the Web.

Connecting users to library resources from the Web is an extention of what we already provide using the OpenURL: offering the user access to copies of materials from within non-library contexts. There may be more than one way to provide this service. The search engines are exploring the use of microformats, in par-

ticular one called "schema.org" to enhance the information that can be presented to users when they search the Web. Google refers to this as "rich snippets" and shows examples that lead users directly to resources within or related to a Web page:

In a similar way, some searches might return a direct link to resources in the user's library community:

Some of this can be facilitated with linked data, but linked data in itself will not make it possible to effectively and efficiently provide these local holdings. To provide actual library services through general Web software we will need to "think different" — even different from linked data.