Paper presented at the Conference on Printed Resources and Digital Information: The Future of Coexistence, Oklahoma University Libraries. March 2 & 3, 2006, Oklahoma City. To be published in the *Journal of Library Administration* (Forthcoming as 46, No. 2 (2007): 87-100. Also separately as a book of conference proceedings: *Print vs. Digital: The Future of Coexistence*, ed. Sul H. Lee. The Haworth Information Press, an imprint of The Haworth Press, Inc., 2007, pp. 87-100. This preprint may differ from published version.

The Digital Difference in Reference Collections.

Michael K. Buckland
School of Information,
University of California, Berkeley

Abstract

One of the very first digital library developments was the transition of bibliographies and other reference works to digital formats and the rise of online services which allowed new kinds of remote searching. But, somehow, the reference *collection* as a library service has not transferred effectively into the digital library environment. How might we re-design the functionality of the reference collection in digital environment? We approach that question through an examination of four reference genres: 1. Subject bibliographies, for which topic vocabularies and cross-references are important; 2. Gazetteers, which, when coupled with maps and bibliographies, allow new ways to search by place; 3. Chronologies, which when digitized and combined with time lines and named time periods, transform search by time; and 4. Biographical directories, which, with improved design, could link persons with their contexts in new and more effective ways. The paper presents work in a project entitled: Support for the Learner: What, Where, When, and Who.

INTRODUCTION

My topic is the reference collection. Somehow the reference collection does not seem to have made an effective transition to the digital networked environment. This is surprising because online searching of bibliographies and reference works was one of the first and most powerful drivers of libraries' transition to a digital networked environment and now there is a lot of work on the development of software to help reference librarians to provide reference help remotely. But the best kind of help is self-help and here we are concerned with the reference *collection*.

I have spent many happy and productive hours using reference collections. The needs served by a reference collection have not gone a way. What happened to the reference collection in the transition to a digital library environment? And what, if anything, should be done next?

New technology does not change the mission of a library. It simply opens up alternative means: new procedures for the same purpose. So what is the purpose of a reference collection? The reference collection is composed of a set of resources selected to serve two needs:

1. Looking up or verifying factual data, often referred to as “ready reference”; and
2. Establishing an initial outline and *context* for any topic efficiently and effectively, especially determining the what, where, when, and who aspects of whatever is of interest.

Libraries are concerned with education. The difference between memorizing and understanding is that understanding means knowing the context. This is what sir Francis Bacon meant when he said “Knowledge is power.” Knowledge is power because, if you understand context and relationships, then you know how to make things happen, and knowing how to make things happen is a form of power.

When a student, a journalist, or a researcher is curious about some topic or event, a traditional strategy is to seek the “5Ws and the H” of investigative writing: What, Where, When, Who, Why, and How. In the past, when libraries were on paper, after one had exhausted the few out-of-date reference books at home, the next step would be to go the local college or public library. And there one would find a wonderful amenity: a carefully selected collection of the library’s best and most up-to-date reference works carefully pre-arranged. There would be a biography section, with biographical dictionaries and Who’s Whos, to help with WHO questions. Also, a history section with almanacs and chronologies designed to help with WHEN questions, and a geography section with atlases and place name gazetteers to help with WHERE. For WHAT there would be general and specialized dictionaries and encyclopedias, and the subject headings in the nearby catalog designed to lead to more. WHY and HOW are less straightforward, but the basic structure was well-designed for WHAT, WHERE, WHEN and WHO, as shown in Fig. 1. In a paper-based environment the reference collection plays an important role, but that helpful structure is largely absent, or, at least, less prominent, in the digital library environment.

Reference Genre	*Vocabulary*	*Special displays*	*About*
Dictionary, Encyclopedia	Topics	Cross-references	What
Atlas, Gazetteer	Places	Maps	Where
Almanac, Chronology	Time	Timelines	When
Biographical dictionary	Persons	Interpersonal relationships	Who

Figure 1. Reference genres and their features.

Reference collections in a digital library environment

In a digital environment, one cannot see the collection. One cannot see beyond the screen, although an interface may provide some guidance. We no longer have the familiar pleasure of seeing a well-stocked collection, of being able to grasp the layout and to assess the relative size of each section, and recognize, at a glance, familiar tools that can provide answers. The valuable structured guidance of the arrangement of the reference collection and of each reference work is mostly absent. The indexes are not usually displayed and, even if browsing is supported, we usually cannot see much of the internal arrangement. Typically the cursor just winks in an empty search box.

This large gap in digital library service is ironic because one of the very first and most visible digital library developments was the transition of reference works, especially bibliographies, to digital formats, and the rise of online search services which allowed one to search reference works remotely. There is a lively literature now on “online reference," but it is concerned more with supporting the work of reference *librarians* rather than empowering self-service use of resources by library users.

The Internet Public Library has an attractive reference collection. If we look at it, we can see that it is a replica of the technology of the codex. There is a dominant hierarchical structure: Go to the collection; find the section; select a reference work; look in the index; find the detail desired. It is convenient in the sense that you can use it from anywhere, twenty-four hours a day, but it is weaker than the paper version in that you sometimes cannot browse the index. You may reach an empty query box and have to guess what term to use and how it is spelled. Then you climb back out and drill down again, one resource at a time, repeatedly, until you are satisfied or give up.

But digital reference does not have to be that way. If you are not using books you do not have to follow the constraints of the technology of the paper codex. Digital technology allows links to be direct and horizontal if two conditions are met: There needs to be *procedural interoperability,* of which the Z39:50 search and retrieve protocol is an example; and there needs to be *vocabulary interoperability*, of which Dewey's Relativ Index to his Decimal Classification is an example.

Technology Change

New technology brings more than technological change and ordinarily comes in two stages. At first the new technology is used to perform existing work in a new way: *to do the same things differently and better*. The earliest printers initially designed type that resembled manuscript writing. The more interesting change comes later in the second stage. With greater familiarity with the characteristics of the new technology, it becomes a matter of *doing better different things* (Buckland 1992). The reference collection presented in the Internet Public Library can be regarded as being in the first stage, a digital replica of the book-based environment: Doing the same thing differently and better. We can now consider the challenge of moving the reference collection into the second stage, going with the flow of digital technology to do different better things. To explore how that might be done we consider the second, more complex purpose of the reference collection, supporting the user's need to establish context by learning about What, Where, When, and Who, and relationships between them.

A series of studies at Berkeley indicates, I believe, how we might reconstruct the attractive functionality of the traditional reference collection in a digital environment. These studies were a collaboration between researchers in the School of Information (formerly the School of Library and Information Studies) and the Electronic Cultural Atlas Initiative (ECAI), an informal collaboration among scholars worldwide to advance education and research in the humanities and social sciences through increased attention to place and time (Buckland & Lancaster 2004). To understand human activities you have to know about the cultural context. What else had been happening in that community at that time. Further, time and place provide a unifying framework across all disciplines and provides an organizing principle for bringing together scholarly resources of many different kinds. The mission of ECAI is not to construct a single cultural atlas, but -- a quite different agenda -- to operate at multiple levels to advance academic best practices: Advocacy for attention to place and time; encouraging the development of infrastructure through collaboration, standards, and technology; and gaining practical experience and providing proof-of-concept. To the extent to which these goals are achieved and resources become network-accessible and interoperable, scholars will become able to compose temporally dynamic cultural maps for themselves drawing on each others' resources.

In what follows, we go back to basics and start with the assumption that the purpose of a reference collection is to provide answers to the four basic questions What, Where, When, and Who. These four facets are different in kind. Distinct reference genres exist for each. Each has special display requirements (as shown in Figure 1) and, in practice, they are very closely entangled with each other.

WHAT – Topic lists – Cross-references within and between vocabularies

Search and selection depends on categorization, which we use as a general term to include indexing, classification, and every other form of ordered arrangement. The forms vary -- indexes, lists of subject headings, thesauri, category codes, classifications, and so on – but they are all descriptive vocabularies and traditionally were called, collectively, "documentary languages." As with natural languages, the meanings of terms evolve and vary between groups. The optimal choice of search term in any given resource may be unclear and a search term that works within one resource may not be the best term to use in another resource even for the same topic. We cope with these anomalies in a paper-based environment because we tend to use one resource at a time, because the number of resources available is usually small and stable, because a printed display of an index enables us to survey the options and overall arrangement, and because it is easier to recognize a topical name than to guess at it. For example if we were interested in martial arts movies, one could recognize as relevant the Library of Congress Subject Heading "Hand-to-hand finding, oriental, in motion pictures," but how many people would have imagined that as a heading to look under? We learn to navigate topic terminology with more or less success. A central feature of indexing and classification is the effort invested in internal consistency and much of the training is in how to establish cross–references within a thesaurus so as to achieve internal coherence.

There always have been two kinds of mapping: Documents are assigned (mapped) to categories; and our queries also have to be mapped to the categories. Library science has heavily emphasized the first and underestimated the importance of the second, even though Melvil Dewey thought that his Relativ Index was at least as important as the Decimal Classification. Dewey considered his index to the classification to be at least as important as the classification itself because to would lead users from whatever words they were familiar with to the correct point in the unfamiliar "vocabulary" of classification numbers. One does not search effectively or economically in unfamiliar resources because learning to use resources effectively takes time and experience and is an important ingredient in effectiveness for scholar and reference librarian alike.

It is not good enough to assume that the use of verbal keywords resolves problems because language is dynamic, unstable, ambiguous, and multiple. Language, especially vocabulary evolves in communities. That is why and how one can often tell where someone is from and what their occupation is from the words they use and the way they use them. It is a matter of dialect. In a paper environment variations of this type are not very difficult to handle. One can see the options displayed. One learns the foibles of different indexes. In a digital, network environment, the situation is different. First, the visual overview than we depend on in a print environment to familiarize ourselves with the vocabulary used is largely absent, even though some browsing may be supported. Second, the whole point of a network environment is to provide access to an ever-increasing range of distant resources. In a digital environment, both the collections and the indexing, although accessible, are more or less invisible behind the glowing screen, which leads to a paradox: the increase in network-accessible resources ensures that a growing number can be used, but also that a growing proportion of these resources are unfamiliar and so, if searched, will not be searched economically or effectively. For example, someone searching major resources for the topic *automobiles* will need to know to search under:

TL205 in the Library of Congress Classification,
180/280 in the U.S. Patent Classification,
3711 in the Standard Industrial Classification, and
PASS MOT VEH, SPARK IGN ENG in the U.S. federal import/export statistics.

This situation has a very important practical consequence. In teaching and in practice the emphasis is placed heavily on how to make cross-references *within* a thesaurus, but in a network environment where many resources are available, the situation is changes. With a larger population of resources, the logic of the situation is to want to harvest from that larger and less familiar pool with less familiar categorization schemes and a third form of mapping, cross-references *between* thesauri, becomes more important relative to the cross-references *within* thesauri. You want to know not only the US Patent Classification number for, say, *making peanut butter*, but also the (very different) corresponding International Patent Classification number, since moving from one to the other should be easier and smoother than in a paper environment.

A digital environment differs from a paper environment in its ability to support links *directly* between entries in different resources. It is no longer necessary to climb out of one work and then drill down into another to relate two entries. The whole point of a digital, network environment is to support search horizontally across many different resources. For this, cross-references *between* the different vocabularies of different databases become very important.

The outstanding example of mapping between different vocabularies is the Unified Medical Language System developed by the National Library of Medicine, a detailed topic mapping between numerous vocabularies in health and medicine. The problem, however, with the expert crafting of mappings between topic vocabularies is that it requires considerable expertise, is slow, expensive, and obsolescent, and does not scale. Fortunately, software techniques have been developed using statistical association and natural language processing which provide imperfect but useful mappings within minutes and at negligible cost if suitable training data are available. (See Plaunt & Norgard 1998, Buckland & others 1999).

Searching Across Different Media Forms

In a digital environment media forms other than text, such as images and statistical data series, are digitized. Here keywords and text searching are no longer feasible, as we found when we tried to bridge the gap between text resources and socio-economic numeric data series. It would be a welcome amenity if one could find both writings and statistical facts on the same topic. Any time you read an article it would be nice to be able go check how well the statistical evidence matched the written assertions. And if you found a startling statistical shift, it would be good library service to enable anyone to find out easily and quickly whether any one written about and explained the anomaly. But the simple keyword techniques of the text universe don't extend to other media forms. Suppose you find an intriguing number in a cell in a statistical table. You could copy that number into a Google search box and get a result, but it is very unlikely that the retrieved set would have anything to do with the statistical datum. The best that can be done is to use whatever textual description may be available in row labels, headers, captions, and elsewhere. Keyword searching of those words might work, but assigning topical headings from a well designed thesaurus will work much better – especially if the text resources also to be searched also use the same thesaurus or one that is or can be mapped to it.

Although different media cannot be linked directly, they can be linked indirectly through descriptive metadata. And, since different descriptive metadata are used in different environments, mapping between thesauri becomes all the more important as resources in multiple media become more common.

WHERE – Places -- Maps

For some purposes, *what* is not enough. For socio-economic statistical data series, for example, it is also ordinarily necessary to specify where. In bibliographies, as with catalogs, *place names* are typically used, which immediately creates two problems. First, "place" is a cultural construct and so inherently difficult to define; second, place *names* being part of natural language, have several problems. They:

- have different forms, St Petersburg, Петербург, Sainte-Pétersbourg, etc.
- Are multiple: Cluj in Romania / Roumania / Rumania, is also Klausenburg and Kolosvar.
- Change: Bombay is now Mumbai.
- Are ambiguous: Numerous places named Beijing or Lafayette.
- Can be anachronistic. There was no country called Germany before 1870, and
- May be vague, e.g. The MidWest, Far East, or Silicon Valley.

Further, there is a tendency to use political jurisdictions to define places, but political jurisdictions and their boundaries are themselves bit unstable. Consider the Balkans and the former USSR.

Space, however, is a scientific construct and can be specified using the coordinates of latitude and longitude. Place and space constitute a dual naming system. Places can be defined in terms of the space they occupy and an important reference genre, the place name gazetteer, records these relationships, linking places with spaces.

A good gazetteer is a list of place names that also says what kind of place (geographic feature type) and gives spatial coordinates (latitude and longitude). These records show when similar names are for different places, when different names refer to the same place, and, most importantly, allow places to be shown on a map. In library parlance, gazetteers are place name authority files. If the place names in reference works were linked horizontally to their corresponding entries in online gazetteers, then there could be map-interfaces both for search results, showing visually the geographic dispersion of any set of records, but also as a device for expressing the geographic scope of a query, expressed as being within any hand-drawn area.

So, for example, a place name can be dot on a map, the dot can be a link that can connect to a webpage about that place or generate a live query to search in other resources concerning this place. (For some examples see ECAI Iraq (2003) and Going Places (2004)).

WHERE – Time Period – Timeline

Just as people tend to refer locations in terms of *places* rather than *spaces*, so also, in both speech and writing, people tend to discuss time in terms of *events* rather than calendar dates. We use phrases such as, “after college,” “during Vietnam,” and “under Clinton.” Time and events are mutually defining. Physical events are used to calibrate calendars and clocks. Calendars and clocks are, in turn, used to express the sequencing, duration, and intervals between events. There is, in effect, a dual naming system for *when* as well as for *where*, but, curiously, named time period directories, analogous to a place name gazetteers, are hard to find. The analogy is close, however, so we designed and built a gazetteer-like time directory with the following components:

- Period name (e.g. Clinton administration, Weimar, Civil War);
- Period type (e.g. Reign, dynasty, war, natural disaster);
- Calendar time, specifying both calendar and dates; and
- Where this named period occurred. (This aspect is analogous to when a place name was in use).

The symmetry in design and relationship with place name gazetteer is shown in Figure 2.

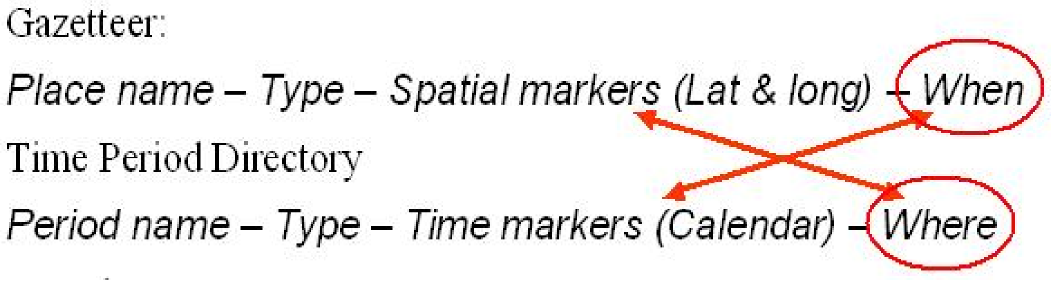
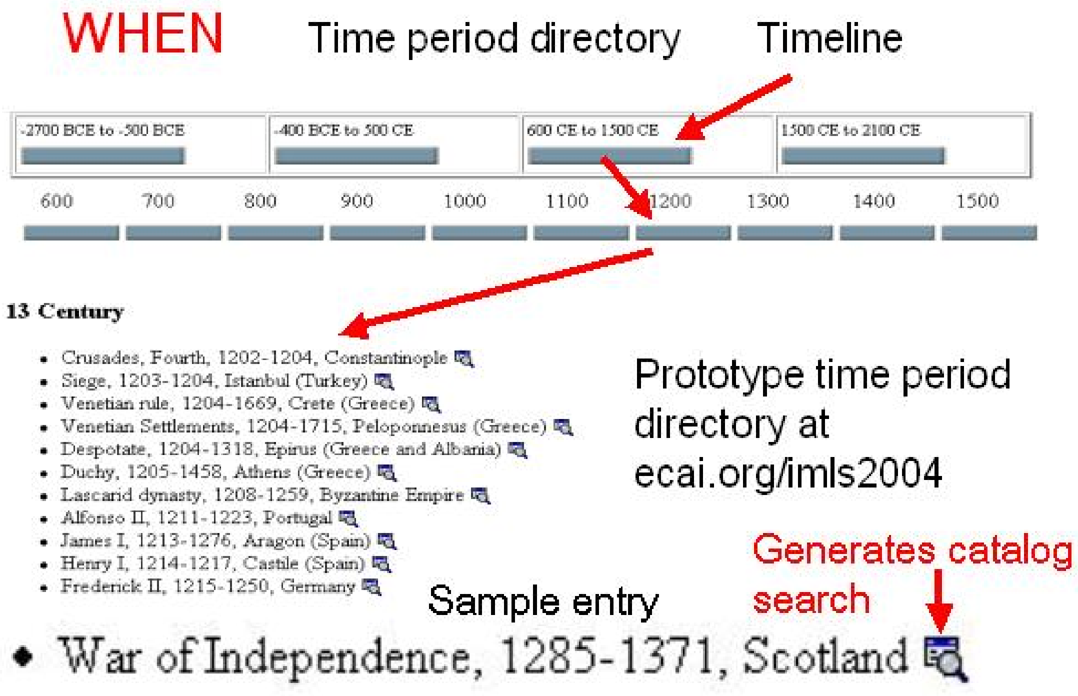

Gazetteer:
Place name – Type – Spatial markers (Lat & long) – When
Time Period Directory
Period name – Type – Time markers (Calendar) – Where

Figure 2: Symmetry of place name gazetteer and time period directory.

A prototype time period directory with some 2,000 entries was constructed by extracting chronological subdivisions from Library of Congress Subject Headings in library catalog records and adding the geographic field when not already present in the record. Three interfaces have been provided: Using lists of names of countries, of major cities, and, for some countries, states; a map with clickable countries, cities, and states; and a timeline interface (See Fig. 3). When any named time period has been selected, users can click on a link to generate a live (Z39:50) search of the Library of Congress catalog to retrieved records for books concerning that period. (The searches work because the time period names are derived from properly formed subdivisions of subject headings.) Further, titles and, more especially, the subject headings of the catalog records retrieved indicate what topics and which individuals were important enough in that period to have books written about them (Petras, Larson & Buckland 2006). (See Fig. 3).



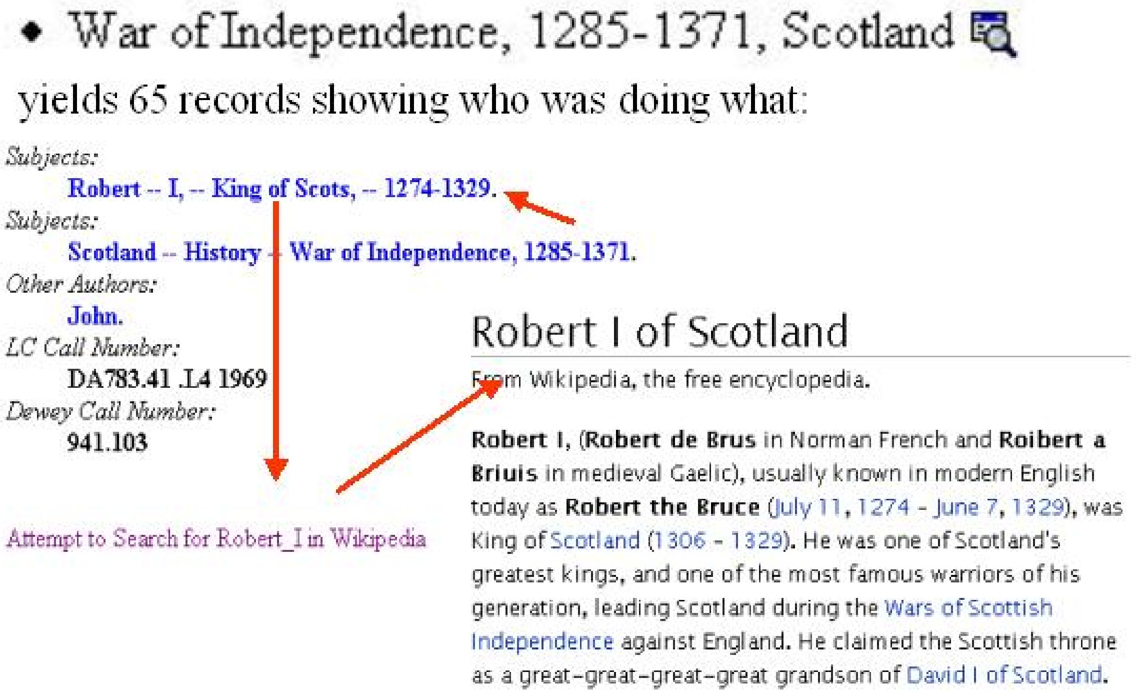
13 Century

- Crusades, Fourth, 1202-1204, Constantinople
- Siege, 1203-1204, Istanbul (Turkey)
- Venetian rule, 1204-1669, Crete (Greece)
- Venetian Settlements, 1204-1715, Peloponnesus (Greece)
- Despotate, 1204-1318, Epirus (Greece and Albania)
- Duchy, 1205-1458, Athens (Greece)
- Lascarid dynasty, 1208-1259, Byzantine Empire
- Alfonso II, 1211-1223, Portugal
- James I, 1213-1276, Aragon (Spain)
- Henry I, 1214-1217, Castile (Spain)
- Frederick II, 1215-1250, Germany

- War of Independence, 1285-1371, Scotland

Figure 3: Time line interface to time period directory, sample entry, and live query link.

The Cheshire system interface was modified to display biographical subject headings separately in catalog records and embed an additional link from each name to search for a corresponding biographical article in the *Wikipedia* online encyclopedia. (See Fig. 4).


- War of Independence, 1285-1371, Scotland

yields 65 records showing who was doing what:

Subjects:
Robert -- I, -- King of Scots, -- 1274-1329.
Subjects:
Scotland -- History -- War of Independence, 1285-1371.
Other Authors:
John.
LC Call Number:
DA783.41 .L4 1969
Dewey Call Number:
941.103

Attempt to Search for Robert_I in Wikipedia

Robert I of Scotland

From Wikipedia, the free encyclopedia.

Robert I, (**Robert de Brus** in Norman French and **Roibert a Briuis** in medieval Gaelic), usually known in modern English today as **Robert the Bruce** (July 11, 1274 - June 7, 1329), was King of Scotland (1306 - 1329). He was one of Scotland's greatest kings, and one of the most famous warriors of his generation, leading Scotland during the Wars of Scottish Independence against England. He claimed the Scottish throne as a great-great-great-great grandson of David I of Scotland.

Figure 4: Query from subject heading of retrieved catalog record to *Wikipedia* article.

WHO – Persons – Interpersonal relationships

Biographical dictionaries and "Who's Whos" are a long established reference genre and they tend to follow similar styles, but moving biographical reference works into an online environment is hindered by two major obstacles. First, formal standards for mark-up and even "best practices" appear to be a lacking, or, if used, remain proprietary and unpublicized. Even the experimental links from our time period directory to *Wikipedia* typically fail for lack of a common standard for form of personal name: Librarians invert (surname first in subject headings); and the *Wikipedia* does not in its filenames. A common standard or a standard conversion is required. Second, librarians, archivists, and bibliographers have focused on establishing authoritative files of personal names, not on the events in people's lives.

Nevertheless, the future use of biographical reference works in an online environment offers particularly exciting possibilities if only format, markup, and metadata could be made interoperable. The reason is that personal lives can be regarded as composed of a series of events (birth, marriage, death) and activities (study, occupation, creative work) each of which occurred during some point or span of time, in some place or places, and often involving other persons. Biographical texts are very densely packed with significant action statements, place names, dates and eras, the names of other persons and institutions, and references to documents. There are many internal links for these various aspects within the *Wikipedia,* but the real logic of digital technology would be to have links to other external resources. Subject headings exist for kinds of activities, gazetteers list places, time period directories can be used for eras, and biographical dictionaries list other people. Biographical text can be seen as epitomizing the potential digital difference in reference collections. When a place is named, it makes little sense to require the user or climb out of the biographical dictionary, locate the geography section, find a gazetteer, and then drill down into it to find the corresponding entry for that place, or to make a note of a mentioned document and take the details to a catalog to locate it. In a world of markup, metadata, and federated search protocols, the challenge is re-design reference works in a digital environment to operate on a basis of lateral links in a formerly hierarchical environment.

We can consider this prospect as a reversal or inversion of the structure of the codex technology of the paper-based reference collection. See Figure 5.

Reference Genre	*Vocabulary*	*Special displays*	*About*
Dictionary, Encyclopedia	Topics	Cross-references	What
Atlas, Gazetteer	Places	Maps	Where
Almanac, Chronology	Time	Timelines	When
Biographical dictionary	Persons	Interpersonal relationships	Who

Reference genres and their features in a paper environment.

About	*Vocabulary*	*Special displays*	*Reference genre*
What	Topics	Cross-references	Dictionary, Encyclopedia
Where	Places	Maps	Atlas, Gazetteer
When	Time	Timelines	Almanac, Chronology
Who	Persons	Interpersonal relationships	Biographical dictionary

Reference features and their genres in a digital environment.

Figure 5: Reference relationships in paper and digital environments

Conclusion

Other speakers have discussed the need to replace or redesign undergraduate libraries and conventional reference desks to serve more effectively a generation of library users who prefer to search from their laptops, in their dorms, during the night, and, as far as they can, independently. The time is ripe for moving the *reference collection* to the second stage of technological change. Reference works are already largely digitized. The challenge is now to adapt the design of the reference works themselves to fit the already existing digital, networked environment of library users. Significant changes are overdue and suitable methods are already at hand to provide the digital difference in reference collections.

ACKNOWLEDGEMENTS

This paper draws on the work of many people, especially Aitao Chen, Fredric C. Gey, Ray R. Larson, and Vivien Petras, in three projects partially supported by the Institute of Museum and Library services through two National Leadership grants: Seamless Searching of Numeric and Textual Resources (award 178, 1999-2002), Going Places in the Catalog: Improved Geographic Access (LG-02-02-0035-02, 2002-04) and Support for the Learner (LG-02-04-0041-04, 2004-06).

REFERENCES

Additional material can be found at and through the “Support for the Learner: What, Where, When, and Who” project website http://ecai.org/imls2004/.

Buckland, M. *Redesigning Library Services: A Manifesto*. (American Library Association, 1992) Available online at http://sunsite.berkeley.edu/Literature/Library/Redesigning/html.html Visited Apr 9, 2006.

Buckland, M. & L. Lancaster. 2004. Combining time, place, and topic: The Electronic Cultural Atlas Initiative. *D-Lib Magazine* 10, no. 5 (May 2004). http://www.dlib.org/dlib/may04/buckland/05buckland.html Visited Apr 9, 2006.

Buckland, M. and others. 1999. M. Buckland and others. Mapping entry vocabulary to unfamiliar metadata vocabularies. *D-Lib Magazine* 5, no. 1 (Jan 1999). http://www.dlib.org/dlib/january99/buckland/01buckland.html Visited Apr 9, 2006.

ECAI Iraq. 2003. http://ecai.org/iraq/ Visited Apr 9, 2006.

Going Places in the Catalog: Improved Geographic Access. 2004. http://ecai.org/imls2002/ Visited Apr 9, 2006.

Petras, V., R. R. Larson & M. Buckland. 2006. Time period directories: A metadata infrastructure for placing events in temporal and geographic context. Forthcoming in the proceedings of the Joint Conference on Digital Libraries.

Plaunt, C., and Norgard, B. A. 1998. An association based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science*. 49, no.10 (August 1998):888-902. http://metadata.sims.berkeley.edu/assoc/assoc.html Visited Apr 9, 2006.