

Bringing Lives to Light: Biography in Context

Excerpts from Proposal Narrative

Assessment of Need

Cultural heritage, history, and social sciences are fundamentally about human activity. Everyone is interested in what other people do and have done.

Life-stories are hard to beat as a basis for narrative and for engaging interest, especially among young people. Biographies are regularly among the best-sellers. Not only History, but also Geography and most other subjects can come alive in the travelogues, journeys of discovery, and the life-stories of those involved. Science can be explained through the work of scientists. Engineering is routinely explained through the heroic struggles of inventors. Even natural history is often taught through the unfolding drama of the activities of an individual animal during its life-cycle or through the seasons of the year.

But mere narrative is not enough. *Understanding the context* differentiates education from memorizing. Building and supporting a community of learners needs more than facts. It is understanding the *circumstances* of people's actions that illuminates their lives, but there is a significant gap in the infrastructure developed by libraries, museums, and publishers in this area. We have standards for handling people's *names*, but not for their *lives*. There is, quite simply, no established standard or "best practices" for encoding what people *do*, nor for helping them to search out the resources that can provide the *context* to understand their actions and experiences.

Our objective is to design, demonstrate, and evaluate techniques that would bring lives to light by revealing them in their contexts.

There is a large literature on biography as a form of historical writing, and on the several different genres, such as hagiography (lives of saints); autobiography; oral history; biographical dictionaries; prosopographia (collective biographies); and on how these genres have evolved and been used over the years (e.g. *Auto/biography studies*, 1985- ; *La biographie historique*, 1990).

Biographies vary in their focus, sometimes with an emphasis on actions (as in hagiography), or on trajectories in time (as with careers); or on geographical movements from place to place, as "lifepaths" (Lifepaths 2004); or on personal relationships, as intellectual history and studies of social networks.

Libraries and museums and the kinds of projects and publications they undertake depend heavily on the construction of biographical records resembling biographical dictionaries, which usually have an underlying narrative structure visible in "Who's Who" style (Petras 2003). There has been a universal move to use structured, marked up records to characterize and to identify details of "content". So what is the best way to mark up the underlying structure and content of biographical records? Consider this example:

Emanuel Goldberg, Born Moscow 1881. PhD under Wilhelm Ostwald, Univ. of Leipzig, 1906.

Director, Zeiss Ikon, Dresden, 1926-33. Moved to Palestine 1937. Died Tel Aviv, 1970.

Note that almost every word denotes *what* he did, some action or activity, directly (born, moved, died) or implicitly (PhD, Director), *where* he did it (Moscow, Leipzig, Dresden, Palestine, Tel Aviv), *when* it happened (1881, etc.), or *who else* was involved (Wilhelm Ostwald, Univ. of Leipzig, Zeiss Ikon). The pattern suggests two conclusions: That representations of lives could be structured as a series of (more or less complete and complex) events; and each event could be represented by a more or less complete four-facet "tuple" composed of an *action* (WHAT) in *time* (WHEN) in a *place* (WHERE) in relation to *others* (WHO). For example, "Born 1881", "Born in Moscow", "Born in Moscow in 1881", "Born to Grigorii and Olga Goldberg at 32 Miasnitskaia, Moscow on August 19, 1881" are differently complete descriptions.

In descriptions so terse, there is much that could, and should, be learned about these activities, times, places, persons and institutions, if we are to understand what was involved. A good paper-based library provides a well structured environment in which a student or researcher could find out more about each element, especially through the structure of the reference collection with its sections on biography, geography, history, and other topics – and, as the terminology becomes clearer, using the catalog and specialized bibliography. A human with pen and paper can build up a dossier through these varied resources, but, by and large, in a digital environment, no such structured environment is provided. In a digital environment where all is beyond the screen, structure has to be provided digitally.

The Larger Context: Metadata as Infrastructure

Biographical summaries for individuals cannot carry much contextual content. They would become large and obsolescent - like an encyclopedia! Nor can the user's workstation hold all the contextualizing resources. The only workable and cost-effective solution is to have access to an "intermediate environment" containing access to basic reference tools and the relationships between them: gazetteers; time period directories; biographical dictionaries; as well as bibliographies and collections; and also the relationships between them. We need an environment that makes it easy to move between specialized resources, following up clues and links – and to compile a selective dossier of excerpts from them. Now the resources would be network-accessible reference resources, e.g. gazetteer servers such as the Alexandria Digital Library gazetteer, the NGiA GEOnet Names Server (GNS), and the *Getty Thesaurus of Geographic Names On Line*, with suitable "service protocols."

Topic categorization systems, such as the *Library of Congress Subject Headings (LCSH)*, are suitable for the WHAT aspect (e.g. **Wool-combing Search Also Under Woolen and worsted manufacture**), but there is the problem of mapping between different topic vocabularies (*LCSH*, Dewey Decimal Classification, *Art and Architecture Thesaurus*, and so on). Methods exist to help searchers with this task (Buckland and others 1999).

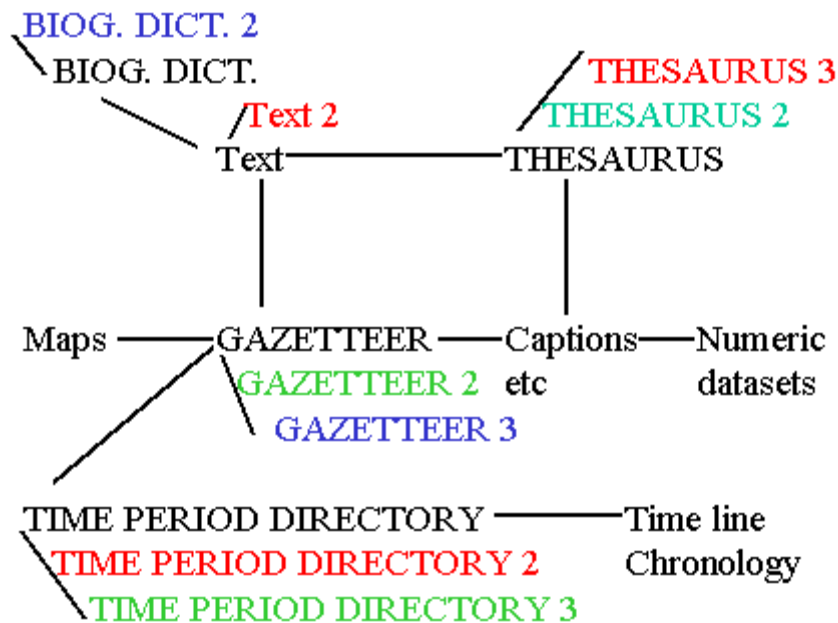
Place names are usual for the WHERE facet, but places' names tend to be multiple, ambiguous, and unstable. A special genre exists: the gazetteer of place names, which is an authority file of place names. A good gazetteer will include relationships between different place names, e.g. *Leipsic USE Leipzig and Leipzig IS-PART-OF Saxony*; a coding for type of place (aka Feature type, e.g. "Ancient site"); an indication of when that name was in use; and, most importantly, a geospatial reference in terms of latitude and longitude. The geospatial specification has several potential uses. It can locate the place, it can differentiate two places with the same name, it can link different names for the same place, indicate spatial relationships between places (e.g. near, between, etc.). Further, latitude and longitude serve as a lingua franca across disciplines and languages to bring together diverse material relating to the same place. Also, very importantly, latitude and longitude, allow places to be depicted on maps. Geographic subject headings, geographical subdivisions, classification numbers for places and tables of geographical suffixes all need to become or, more sensibly, link to gazetteers. (For our work on this issue see Buckland & Lancaster (2004) and the *Going Places in the Catalog* project website: ecai.org/imls2002).

Similar issues arise with time. In speech and in writing, people commonly refer to periods of time by name rather than by calendar date, e.g. "Civil War", "during Vietnam", "under Clinton", and so on. Named Time Period directories, comparable to place name gazetteers, are needed to provide this function, and would include, minimally, the period name, type of period (reign, war, dynasty, natural disaster, etc.), and a specification in calendar dates. And just as place names have a temporal aspect, period names have a geographical aspect. The "Civil war" period is in different centuries in the U.S.A., England, and Spain. Petras (2005) describes a

prototype online Named Time Period directory with 2,000 records based on enriched *LCSH* chronological subdivisions. (See ecai.org/imls2004/timeperiods.html).

The WHO facet can be supplied by reference to other entries in the same or other biographical directories to find text and portraits.

These four aspects WHAT, WHERE, WHEN, and WHO, each with its own distinctive tools and display requirements is shown in the following diagram in which the metadata structure is capital letters provides the framework for search and navigation.



National Impact and Intended Results

IMLS envisions a seamless infrastructure for learning where schools and cultural and community organizations work in concert. But this cannot be achieved without adequate and interoperable tools. Resource *mark-up*, such as XML, provides a means to identify content in a digital document.

There has been a huge investment in the development of digital resources and certain kinds of infrastructure. The benefit, the return on this massive investment in hardware, software, and digital “content”, lies in the quantity and kinds of the *use* made of it. And if the right tools are not at hand, or they are not convenient and intuitive, the resources will be little used.

The crucial next step is to move beyond content mark-up to embedded links that generate connections to additional explanations in network-accessible resources. Links recreate the “library” by relating the digital document to the digital context.

The intended result is the development and demonstration of tools to enable learning resources -- libraries, museums, and scholarly publishing – to move the biographical content of their resources toward a seamless infrastructure. Tools that shed light on people’s lives are also of widespread interest far beyond education in government, industry, law enforcement, health care, and other contexts.

Project Design & Evaluation Plan

Project design

This project is structured as a set of tasks:

Task 1: Encoding biographical content. The task is to design, develop, and validate standards and best practices for describing lives. The approach builds on XML, TEI, CIDOC, EAC, etc., but requires a specialized language, comprising:

- *Grammar (Syntax)* ,e.g. a XML markup schema, building on existing standards or proposed standards ; or, comparably, “microformats”; and
- *Vocabulary (Semantics)*, with three aspects: (i) A robust basic, general-purpose set of terms or categories for life events, including birth, education, marriage(s), employment, . . . and death; (ii) Specialized vocabularies and categories for atypical applications, e.g. for lives of medieval Tibetan Buddhist monks; and (iii) Mappings between the general and specialized vocabularies, as is needed in any situation with multiple thesauri.

The pivotal challenge here is to move beyond the well-established routines for establishing *name authority control* and into ways to encode and to understand the *events* (activities, episodes) of life. Building on our prior research (summarized in an Attachment), a simple repeatable 4-tuple appears adequate: WHAT activity or event; WHERE did it happen; WHEN was this; and WHO [ELSE] was involved.

Task 2: Interoperability through embedded links and queries. If encoded mark-up of content is central to the use of digital documents, encoding dynamic links providing connections through the infrastructure of scholarship is crucial to using documents in a digital environment. Static links are familiar. We emphasize links that generate queries to seek out more information. In the *Wikipedia* embedded queries are now provided for ISBNs that, when clicked, can fetch related bibliographical data and search for the ISBNs of related editions. In the *ECAI Iraq* portal (ecai.org/iraq/) we embedded special links that initiate Z39.50 searches of major libraries’ catalogs for holdings relevant to the page. In the Robert the Bruce example below, the biographical subject heading in a retrieved catalog record searches the *Wikipedia* (Buckland 2005).

Task 2 is importantly different from Task 1. We need embedded links that generate well-formed queries from the encoded content to harvest additional contextual resources:

WHAT queries to libraries’ subject catalogs, then on to texts and images.

WHERE queries to place name gazetteers servers, then on to map displays.

WHEN queries to time period directories, then on to time lines and chronologies.

WHO queries to name authority files, biographical directories, and encyclopedia entries.

As an example, each entry in our named Time Period Directory has a live link generating a Z39.50 search of the Library of Congress online catalog. The link for “War of Independence, 1285-1371” (pertaining to Scotland) retrieves numerous records revealing what and who was important enough during that period to become the subject of books. The biographical subject headings in retrieved catalog records are transformed into searches for biographical articles about the person named:

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](#).

Task 3: Tools for Effort-Effective Editing:

Adding mark-up is tedious. Two kinds of tools can make a large difference:

1. Tools to identify and disambiguate names, places and times, generically known as “named entity extraction”; and
2. Tools for inserting mark-up more or less automatically.

The third task is to assemble and provide both kinds of tools as editing aids.

Task 4: Testing in Practical Applications:

Rigorous design and realistic stress-testing requires practical needs and diverse materials. Partners needing better methods of handling very varied biographical material will provide a realistic range of specialized needs. They have each agreed to (i) Define a real need; (ii) Provide a range of biographical texts; and (iii) Assist with design and evaluation.

These “editorial partners” and their corpora (described in the Attachments) constitute a carefully selected range of needs in three complementary sectors, in each a mainstream application and a more challenging environment to provide “stress-test.”

(a) Biographical description in Archives and Historic Texts

(a.1). Our mainstream test bed is the U.K. *Archives Hub*, a national gateway providing free access to descriptions of 18,000 archives and manuscript collections held in UK universities and colleges. The collections range from small sets of an individual's correspondence to the archives of large businesses and include collections of films, audio recordings and digital collections as well as more traditional archival materials. The 2002 international standard Electronic Archival Description (EAD) includes elements for biographical information, but EAD records seldom follow any detailed structure in providing this information. The proposed Electronic Archival Context (EAC) standard may provide elements to permit detailed biographical mark-up, but so far all examples have been based on bibliographic authority records and lack detailed biographical information.

For example, EAD record #GB 0532 HOL for “Papers of Sir Isaac Holden (1807-1897) and Family, West Riding wool combers” contains a <bioghist> record element:

```
<bioghist encodinganalog="JISC-HUB07"><p>
```

Sir Isaac Holden was born in Paisley in 1807. He worked in a cotton mill, as a shawl-weaver, and as a teacher before taking the post of book-keeper in Townend Brothers worsted mill in 1830. He stayed with the firm until 1846, becoming works manager and introducing mechanical

woolcombing methods. In 1847 he took out a joint patent with Samuel Cunliffe Lister for <title>Improvements in carding preparing combing and spinning wool, and also in making heald and genappe yarns</title>. In 1848 Holden and Lister opened a wool-combing factory at St. Denis, near Paris ; this was soon replaced by factories at Croix and Rheims. About this time the two were involved in work on the square-motion comb, which Lister later patented. In 1858, after a number of disagreements, Lister sold his share of the business to Holden. Leaving the running of the French factories largely to his nephews Jonathan Holden (1828-1906) and Isaac Holden Crothers (1830-1908), in 1864 Holden opened a large new wool-combing plant in Bradford, the Alston Works.</p>
<p>Holden was elected M.P. for Knaresborough 1865-68, for the Northern Division of the West Riding 1882-85, and for Keighley 1885-95. He was created a baronet in 1893, and died in 1897.</p></bioghist>

A more detailed XML mark-up would identify each person and place. We propose to design and demonstrate mark-up that would:

- relate place names to a place name gazetteer server and show them on a map display;
- combine place names with dates to generate a geographical “lifepath” on a dynamic map;
- also identify key actions and link them to explanatory encyclopedia entries and to LCSH subject headings for library catalog search;
- display the dates as a chronology or time-line and, through an online time period directory;
- find external resources, such as the text of the patent and mention in Parliamentary papers.
- and so on.

The <bioghist> element in both EAD and EAC permit the use of a chronology of events in a person’s (or corporate entity’s) life, such as (in EAD):

<bioghist>

<head>Biographical Note</head>

<chronlist>

<chronitem>

<date>1892, May 7</date>

<event>Born, <geogname>Glencoe, Ill.</geogname></event>

</chronitem>

<chronitem>

<date>1915</date>

<event>A.B., <corpname>Yale University, </corpname>New Haven,

Conn.</event>

</chronitem>

<chronitem>

<date>1916</date>

<event>Married <persname>Ada Hitchcock</persname>

</event>

</chronitem>

<chronitem>

<date>1917-1919</date>

<event>Served in <corpname>United States Army</corpname></event>

</chronitem>

</chronlist>

</bioghist>

Detailed chronologies like this are very rare in EAD records, however, and there is no standard way to differentiate *types* of events.

We will work with the Archive Hub personnel to show how <bioghist> elements could be enriched with automatically derived mark-up and links. One approach is to use EAD/EAC specific *microformats* to mark-up data without requiring changes to the existing EAD or EAC

schemas, or to append an automatically generated <chronlist> to a more conventional <bioghist> narrative.

(a.2) The *Centre for Document Digitisation and Analysis* (CDDA), at the Queen's University, Belfast, is a research unit with interests in temporal Geographical Information Systems, the development of electronic research resources, e-Science and Grid technologies. It provides a comprehensive digitization service and is active in research, both in terms of the development of new digitization methodologies, and in using computerized resources in traditional scholarship. The Centre has a particular interest in GIS in the humanities and social sciences, has specialized in the digitization of exceptionally varied form of digital documents, and emphasizes mark-up by place, time, and map visualizations. We will work with CDDA on a selection of unusual and challenging resources.

(b) Scholarly Editing

(b.1) Our mainstream test bed for scholarly editing will be our partnership with the editors of *The Emma Goldman Papers Project*, part of a national initiative to retrieve the papers of individuals whose life work has had a lasting impact on the course of American history. The EGPP, at the University of California, Berkeley, has collected, organized, and edited tens of thousands of documents from around the world by and about Emma Goldman (1869-1940), a leading figure in American anarchism, feminism, and radicalism. The papers provide a window into social and cultural movements in late-nineteenth- and early-twentieth-century America, Europe, Asia and Latin America. (See sunsite.berkeley.edu/Goldman/Project/project.html)

An important component of the editorial work is the compilation of biographical summaries (See http://sunsite3.berkeley.edu/Goldman/Samples/sample_bios.html), e.g.

Timmermann, Claus (1866-1941), German-born anarchist, editor, immigrated to the United States around 1883. In St. Louis he edited and published *Der Anarchist* from 1889 to 1891. In the summer of 1891 he ceased publishing the paper and moved to New York. The following year, according to E[mma] G[oldman], she and A[lexander] B[erkman] confided in him about their Homestead plan, and he helped them write the manifesto to the striking steelworkers, "Labor Awakens." Timmermann was tried on 1 September 1893, and sentenced to six months on the charge of inciting to riot for his speech at the 21 August rally in Union Square, the political gathering that prompted EG's arrest, trial, and imprisonment. In New York, he edited the anarchist papers *Brandfackel* (1893-1894) and *Sturmvogel* (1897-1899).

(b.2). As a more severe challenge, we propose to work with the University of Southern California Shoah Foundation for Visual History and Education. The Shoah Foundation collected testimonies from nearly 52,000 survivors and other witnesses of the Holocaust, and at the same time collected biographical data from these interviewees. We hope to work with the partially edited transcripts. The sheer complexity of the geography of central Europe, the multiple languages, and the tumultuous events make this collection an excellent challenge, especially with respect to place names and chronology. Following informal discussions, we had hoped to attach a letter giving formal permission to use this material. However, the administrative changes associated with the Survivors of the Shoah Visual History Foundation becoming a part of the University of Southern California this January have delayed a formal response. The interviews do not present themselves in the linear fashion in time and place. They are replete with flashbacks, ambiguous references to friends and relatives, discussions of possible courses of action from which only one was finally chosen, so the assignment of metadata descriptors to interview sections is challenging. We attach a letter of support from Prof. Doug Oard who has worked with development of the XML marked up data and Automatic Speech Recognition

transcription of interviews under funding from the National Science Foundation. 4,000 of these Holocaust survivor interviews have been manually indexed by trained indexers using a custom-built thesaurus for their controlled vocabulary. The thesaurus is rich in geography (“Warsaw (Poland)”), special geographic types (“Concentration Camp”, “Hiding Place”) and geo-temporal references (“Berlin 1939-1944”). A “Persons” file gives the essential information about date-of-birth, birthplace, residences during WWII (ghettos, concentration camps) and post-war destination.

For example, survivor SL was born in 1928 and raised in Przemysl, Poland, near the Ukraine border. The Shoah thesaurus entry for Przemysl would need to be augmented with latitude (49°47′ N) and longitude (49°47′ N) locator information available from another source, such as a gazetteer. One might wish to add contextual information for Przemysl. A web search “Przemysl Jewish population” yields these grim statistics “*Before WW2 about 24,000 Jews lived in the town,*” and “*It is estimated that only 400 Przemysl Jews survived the Holocaust in Przemysl itself, Russia, Poland and in other countries.*” (<http://www.deathcamps.org/occupation/przemysl%20ghetto.html>)

(c) Educational Publishing

(c.1.) Our mainstream test bed will be the biographical articles in the *Wikipedia*, which is an irresistible resource because it is accessible, text can be downloaded, and articles can be submitted for experimental purposes.

(c.2.) To ensure that the techniques are really robust we will work with the [Religious Atlas of China and the Himalayas](#), being developed by a team of scholars throughout North America, Asia, and Europe. It will include names, dates, coordinates, and associated information for several thousand historic religious places in China and the Himalayas, including mosques, churches and temples; sacred mountains; religious kingdoms; monumental statuary, and other categories. Users will be able to make maps and time lines, conduct queries to learn how places are linked by sect, place, personage and other characteristics; and link from the brief records in the gazetteer to rich documentation such as images, scholarship, and additional information about each place. At least one-third of the data is biographical. This brings complexity of languages and scripts (Chinese, Tibetan, and numerous others).

These diverse corpora and partners provide a strong “laboratory” for our work.

Task 4. Evaluation.

This is a research project attempting to demonstrate that hoped-for improvements are feasible. The evaluation methodology is therefore in two parts:

1. Proof of Concept: Are the Tasks 1-3 enumerated above feasible? This is mainly an either-or evaluation. The onus is in the researchers to demonstrate that what is proposed can in fact be done. This can only be achieved by doing it, by building, demonstrating, and documenting a functioning prototype. It is our intention to provide an openly web-accessible prototype of the systems developed – for all to see – as we have in prior projects.

2. Testing and Evaluation: If what we propose is feasible, *how well* does it perform? In addition to technical evaluation through TREC style conferences, notably CLEF and INEX, we have designed an Outcomes-Based Evaluation and a “Logic Model” is attached. Our partners are committed to assisting with data collection for evaluation.

Task 5. Documentation and dissemination.