Preamble

Knowledge is power, wrote Bacon, *quia ignoratio causae destituit effectum*: To achieve, you need to understand cause and effect. More generally, knowing the context and relationships involved makes the difference between seeing and understanding. Credible scholarship in the Humanities ordinarily depends on years, even decades, of deep immersion in the relevant primary sources, secondary resources, and surviving material culture. For the most part, the laboratory of the humanities scholar is the library: the personal library, the campus library, and archives, libraries, and museums elsewhere. The price of this necessary immersion in context is high in terms of time and effort – and of travel since all libraries are more or less incomplete. Researchers, students, the general reader, editors, . . . everyone is necessarily concerned with the context and relationships of whatever is studied: What other documents relate to this topic? Where and when did this happen? What else was going on around that time and place? Who were the people and institutions mentioned? How were they related? What else did they do?

Over time genres of auxiliary resources have evolved: dictionaries and encyclopedias; bibliographies and library catalogs; place name gazetteers and maps; time-lines and chronologies; biographical dictionaries; and so on. In a print environment the reference collection of the library provides a carefully constructed environment of auxiliary resources well-designed for exploring context.

In a well-stocked library one can consult a diversity of documents and reference works to build up an understanding of any topic. *Bibliographies and library catalogs* provide documentary context by listing documents associated with any author or topic. *Place name gazetteers and maps* provide geographical context by identifying places, indicating where they are, what kind of place they are, and how they are spatially related to other places. *Chronologies and Time-lines* place events in temporal context by listing them in calendar time and showing what happened in any period and what else was going on at, before or after any given event. *Biographical dictionaries* and *Who’s Whos* provide personal contexts by explaining who people were and something about what they did, where and when they did it, and who else they were associated with. *Encyclopedias* provide a wide range or information.

These tools, the essence of a library reference collection, are well-developed, well-understood, and heavily-used in the Humanities in the pre-digital environment. The problem we address is that in the digital environment the need for the functionality of a well-stocked, well-organized reference collection has been neglected. The development and availability of digital versions of individual tools such as place name gazetteers has been uneven and often ill-suited for the Humanities use and the organizational genius of well-selected, easy-to-use *reference collections* has been neglected in digital library development.

In the potentially better-stocked Internet environment, one should be able to be able to enjoy an even better and more supportive “reference” environment. Digital reference resources have been emerging but they are not yet very accessible. They are usually protected by copyright and exist in technically isolated and/or restricted proprietary services. Open access is increasing, however, and there is a strong and widespread desire for sharing material and for developing resources in a open, shared manner (“peer production”).

The convenient structure of a library reference collection is still effectively absent from the digital library environment. What makes the library reference room effective is the ability for humans to “manually” link the resources through proximity and well-placed signs in the library. Neither the proximity nor the “linking” process has been figured out in the digital realm, which is what this project aims for.
New digital corpora are emerging steadily, often on a large scale. How are they going to be studied effectively until convenient auxiliary resources are provided that are also conveniently available in the infrastructure of the digital humanities environment? Until then, the full return on the investment in creating the large corpora will not be reaped.

The logic of the Internet is not simply that remote resources can be accessed from your desktop, but that remote sources can be used in conjunction with each other.

Two very important developments, together, provide the means. One is the rise of “federated” search and retrieval tools (such as the ISO 23950 protocol, aka NISO Z39.50) which enable remote and disparate systems to “talk” to each other. The other is the development of methods for mapping between the multiplicity of vocabularies used in indexes, thesauri, category codes, classification schemes, ontologies and the like. Crafting pairwise mappings by hand is generally not affordable, but techniques using statistical association and natural language processing can generate very useful connections or prompts when moving between different subject organizing systems.

What could now be done

1. Dynamic hyperlinks. The searching previously done by hand in a paper environment can increasingly be done online – and then documented within the mark-up now routinely applied to text documents. It is a digital analog of noting a source in a footnote, but with a dynamic link, as pioneered in the ECAI Atlas of Iraqi Antiquities (ecai.org/iraq) where for each archaeological site there is a dynamic hyperlink that can generate a live search to, say, the Library of Congress catalog or COPAC, the union catalog of all British research libraries, for materials relating to that site and so the result is as up-to-date as the libraries’ cataloging.

2. Advanced search techniques are needed for certain purposes and purely text-based search and retrieval techniques are insufficient. For example, places and their boundaries change over time and place names are multiple, ambiguous, and unstable. So the usual practice of relying on the names of contemporary political jurisdictions is simply inadequate for many purposes in the Humanities. In a reference collection one would look in an atlas: The fine print in the back is a place name gazetteer which serves to identify and to disambiguate place names and also constitutes an index to the printed maps. Maps show the place in context, in relation to other places and to geographic features, such as elevation, land use, and boundaries. In a digital environment, given the right tools, one could do the same, invoking a gazetteer server and a map display.

3. Once the desired functionality is defined one can build support tools for the search for context; for the saving of notes; for the embedding of notes (and, better, hyperlinks) as mark-up; and for sharing the results either as marked-up text or as “stand-off” mark-up (separate from the text but linked to it) with others to use. Then others can enhance the mark-up and share it, building, if desired, a productive community.

4. Enrichment of auxiliary resources. Better yet, once such mark-up has been created, it can also be re-purposed. Suppose that an historical or literary text mentions a hundred different places and has been enriched by mark-up that links the text’s mentions of places to entries in a suitable place name gazetteer. A copy of this mark-up, necessarily in the order in which the places are mentioned in the text, can then be rearranged by place name and added to the gazetteer records. A hundred place names in the gazetteer now have pointers to where that place is mentioned in a text. As a byproduct of marking up the text, the gazetteer as well as the text has also been enriched! The gazetteer becomes more powerful because now it also shows where places are discussed in a text. The gazetteer serves as high quality geographical index to the text. A map could then display the geographic distribution of place names in the text and show all the text mentions within any area. The same could be done with additional marked-up texts. This allows the geography of the two texts to be compared and texts to be ranked with respect to their coverage of any given area. These webs of relationships can be extended indefinitely.
Other kinds of auxiliary resources could be treated similarly. For the sake of exposition we focus, in this proposal narrative, on geographical context and the role of gazetteers and maps because doing so illustrates our purposes well. But in the work to be done we are will be equally concerned with topical, temporal, and biographical aspects of contexts.

Topical searching is relatively well understood and the project team has extensive professional experience dealing with it, decades of published research and teaching in the case of Prof. Buckland, Prof. Larson, and Dr. Gey. ¹

Handling temporal context has been given less attention. When people write or speak about time they often do not use calendar dates. Rather, they use events and named time periods in phrases such as “Elizabethan England,” “During Vietnam,” “during the reign of Henry II.” This use of events to mark time resembles the use of place names in being culturally and locally situated; ambiguous, e.g. a “civil war pamphlet” would date from the seventeenth century in an English context, from the nineteenth century in the U.S., and from the twentieth century in Spain; or multiple, e.g. “during the Great War / First World War / World War I.” Our solution has been to build a Named Time Period Directory modeled directly on a place name gazetteer and which relates named time periods with their associated dates and places. Our first example, populated with 2,000 examples derived from the Chronological Subdivisions used in Library of Congress Subject Headings allowed us to access events through date, event names, or locations. Just as gazetteers relate names to geospatial markers (latitude and longitude) and displayed on maps, named events used temporally can be related to calendar time and displayed in time-lines and chronologies.

The traditional tools for biographical context include biographical dictionaries and also bibliographies and library catalogs are used to find out what has been written by or about any particular individual. One problem that is well understood is the multiplicity and ambiguity of personal names (“name authority control”). Two unresolved neglected problems are the lack of standards or best practices for describing the events in people’s lives and for effectively coding types of the many interpersonal relations (e.g. mentor, business partner, fellow student) beyond the immediate family relationships. This messy situation was well documented in a 2006 TEI report.² ECAI is currently addressing these problems directly in a project entitled Bringing Lives to Light: Biography in Context, led by Prof Ray R. Larson and supported by an IMLS National Leadership grant. The approach being taken is to express the biographee’s life events as a number of events (or activities or changes), often happening concurrently and at any level of granularity. Each event is then encoded as a more or less complete 4-tuple: What kind of activity, where, when, and with whom else. The expectation is that this would achieve two goals: First, every instance of what, where, when, and who-else could be formally linked (hyperlinked, when online) to explanations in auxiliary resources (encyclopedias, bibliographies, time-lines, biographical dictionaries) with the other aspects steering the inquiry (e.g. for a weaver, not just but weaving, but weavers and weaving in eighteenth century New England); second, consistent encoding provides an immediate basis for the identification and analysis of groups of people with any characteristic in common


(prosapography). This project runs through September 2008 and would be tightly coordinated with the proposed project through the active participation of Prof. Larson, Prof. Buckland, Fred Gey, and Kimberly Carl in both.

It is not simply that each of What, Where, When, and Who can and should be treated in its own way, but that these four facets can then be combined as needed to form a well-structured framework for showing context and exploring relationships. That is what we propose to do. The time is ripe and appropriate techniques are at hand. We propose to show what could be done by building the tools and demonstrating a proof-of-concept prototype service in the case of Irish Studies.

A convergent collaboration

A convergence of four kinds of specialists will create, demonstrate, and evaluate techniques to enable anyone to search out the background for topics, places, events and persons encountered easily, inexpensively, and rapidly when reading Humanities texts.

1. Advanced Search Techniques. The Electronic Cultural Atlas Initiative (an informal international collaboration of, mainly, humanities scholars) and the School of Information (formerly School of Library and Information Studies) at the University of California, Berkeley, have been working together for some years on the development of advanced search tools suitable for geographic search, searching by time period, and searching across different fields and different genres.

2. e-Resources on Ireland. At The Queen’s University, Belfast, a collaborative project between the Information Services (which includes the University Library) and the Centre for Data Digitisation and Analysis is creating “A digital library of core e-resources on Ireland: Developing a comprehensive electronic library of materials of interest to Irish Studies scholars.” They will digitize the entire back files of 100 carefully selected journals and some 200 books and 2,500 pages of manuscripts on Irish culture and history. This corpus will be made available through the [U.K.] Arts and Humanities Data Service, through JSTOR, and in other ways. See http://www.qub.ac.uk/cdda/dlcmi/summary.html.

3. Irish Studies. Berkeley’s degree-granting Celtic Studies Program brings together faculty concerned with different aspects of the culture and history of Ireland and related Celtic cultures. The faculty are eager to participate in an initiative that could bring better access and improved analytical tools to support their interests. These collaborating humanities scholars and students will provide guidance concerning their needs, assist with the cultural, linguistic, and historical complexities of the material, and participate in the evaluation for their particular purposes of what is developed.

4. Archives and Scholarly Editing: The Emma Goldman Papers Project at the University of California, Berkeley, is typical of many archives in preparing scholarly editions of historically important papers, an undertaking that is hard to fund and slow to complete. The Emma Goldman editors, already actively collaborating with ECAI on improved practices for marking up biographical texts, are motivated to collaborate in designing and testing any tool that could facilitate and speed up the editorial process, thereby reducing the time and cost. The many papers relating to Goldman’s Irish contacts and interests provide a suitable test case.

Major Issues

The underlying issue is to show how advanced search techniques (largely developed with IMLS support) and wider adoption of standards and best practices could improve the scholarly infrastructure in and for the Humanities. The specific issues addressed are:
1. The acute importance of knowing context in the Humanities: The central objective is to show how the digital infrastructure could be enhanced to make it easier to find out more about any topic, place, period, person, or institution. We have no desire to create any new digital resources (“content”), only to show how current and future resources, especially auxiliary resources, could be used more easily and effectively.

2. Traditional text-based search techniques, though powerful, are simply not sufficient. For example, places (a cultural concept) exist in space (a physical concept) – and places (and relationships between them) cannot be effectively understood without a map. A place name gazetteer addresses this duality by serving as a kind of “bilingual dictionary” between place names and geo-spatial coordinates (latitude and longitude). In a digital environment a dynamic map interface can show geographic dispersion, relationships between data of different sorts (“layers”), and changes through time (Zerneke 2006). Given the limitations of using current names and contemporary political boundaries, one should be able to draw any loop on a map to indicate an area as the region of interest.

3. The emergence of digital versions of traditional reference works is important. A portal with links to well-selected resources, as at Irish Resources in The Humanities (irith.org), is very useful. We address the next step of programming the interface for forwarding searches into listed resources.

4. The Humanities are heavily dependent on a range of cultural agencies, especially archives and special collections of papers, libraries, university presses and other scholarly publishers, and curators of material culture (museums and historic sites). All have economic problems. As a by-product of our primary concern with Humanities scholars this project could contribute to the efficiency or cost-effectiveness of the first three.

How grounded in the Humanities

This project is centrally grounded in the Humanities because Humanities scholarship requires understanding complex cultural developments and this requires deep awareness of context. The intellectual effort of examining context in the Humanities research cannot be avoided or delegated, but the time and effort of finding and using auxiliary and explanatory resources could be greatly reduced in a better-designed digital humanities environment. Our primary objective is show how scholars and students in a range of disciplines and the general public could be helped. As a by-product of that concern we will also contribute to the effectiveness of cultural agencies upon which Humanities scholars depend.

Scope

The scope is defined on three dimensions: by the corpus of 100 journal back files, 200 books and 2,500 manuscript pages concerning Irish studies; augmented by others resources on an opportunistic basis as would benefit the project’s purposes.

Relationship to other uses of digital texts

The availability of digital humanities corpora can be seen in terms of three uses. First, having resources available at the desktop instead of dependence on travel and interlibrary loan is a huge advantage. Second, computational operations within a corpus (search, natural language processing, text-mining, etc.), now a well-established field, allows a variety of analyses that were not previously practical.

Our interest is in a third and different direction: Computational techniques for relating texts and text fragments to external resources outside the corpus, seeking to contextualize by relating what is inside it to what is outside.

The difference between these three uses can be illustrated with an important resource for early Irish history: Edmund Hogan’s Onomasticon Goedelicum : Locorum et Tribuum Hiberniae et Scotiae =

An Index, with Identifications, to the Gaelic Names of Places and Tribes, which indexes place names from 175 sources. The preface states: “This book contains, firstly, the Gaelic place-names found in the many unprinted and printed volumes . . . secondly, their identifications taken directly from the cited text and the context, or from several texts and contexts combined . . . Such a book, if fairly well done, will save Celtic scholars from exhausting and time-wasting search . . . and set them free for the nobler work of editing and translating Irish texts.”

This resource exists only on paper. A preliminary test on the passage shown indicates that it could be reliably digitized and made available to all over the Internet (The first use). Ideally, it could also be enhanced by expanding the abbreviations and adding latitude and longitude where known. Once available as a digital text file, a variety of text-mining analyses by county, by source, and other features are readily imaginable. (The second use).

The third use, that interests us, would be to link what is in the text to what is outside: the place names could be searched in bibliographies and library catalogs for related documents; references to the sources could become hyperlinks to such sources as have also been scanned or digitized; the place names could be displayed as points on new or ancient maps and the geographic context (boundaries, topography, climate, current land use) examined; mentions of persons, such as Niall Caille, could link to corresponding entries in biographical dictionaries; the place names could link to drawings, photographs, and virtual reality reconstructions; and numerous other possibilities. But, even better, every such outward link could also be used in reverse as an inward link: Each resource mentioned could be annotated so that there is an entry for this place or person in Hogan’s Onomasticon Goedelicum.

Study of the Medieval Irish secular saga tradition involves not only decoding Middle Irish language, but identifying individual characters and scribes, relating the time of manuscript production to the presentation of stories (ancient even in their own time), identifying place and tribal names, ancient and modern, knowing about traditional and early Christian legal systems in Ireland and more. Simply identifying the physical route of the contending forces in the epic Táin Bó Cuailnge (“The Cattle Raid of Cooley”) requires most of the disciplines just mentioned. Closer to the present, Yeats’s poems, “Song of the Wandering Aengus” and “The Lake Isle of Innesfree,” for instance, require a good deal of non-literary information to decode.

For example, we find a reference to the ancient place name Emain Macha in an article in Studia Hibernica. Clicking on the name could give us access to various kinds of contextual information: 1) A zoomable satellite view of the site with information on latitude, longitude, and spatial relationship to other places, 2) A link to Hogan’s Onomasticon Goedelicum which gives the modern name of the site (Navan Fort, Co. Armagh [which could have its own links to the local chamber of commerce or whatever]) plus a set of references to it in medieval texts, 3) A link to the Dictionary of the Irish Language giving...
etymological and usage information, 4) A link to images of some of the manuscripts in which the name occurs, 5) Links to the name in bibliographies, such as the already searchable Celtic Studies Association Bibliography [<www.humnet.ucla.edu/humnet/celtic/csanabib.html>], 6) Links to relevant map and atlas sites, 7) Direct links to the name in other publications in the CELT corpus and soon to the Belfast Irish journals corpus.

Relationships to other work

The project involves two really tight collaborations: First, with the ECAI project “Bringing Lives to Light: Biography in Context,” led by Prof. Ray Larson, supported by IMLS, to develop better practices for marking up biographical text; and with The Queen’s University, Belfast’s digitization of the core literature on Irish culture and history, project funded by the [U.K.] Joint Information Systems Committee (JISC) and explicitly charged by JISC to collaborate with initiatives in North America.

The methodological approach is compatible with the intellectual goals of the project and the expectations and needs of users because we develop tools and infrastructure to enable learners at any level to understand context and relationships more easily and more quickly. A research and demonstration project of this type requires a specific focus to test its effectiveness – in this proposal Irish Studies – but, once the techniques have been developed and tested they would be applicable across the Humanities and beyond.

2. Importance of this collaboration

The project is a collaborative effort of the following organizations [. . . ]:

1. The Electronic Cultural Atlas Initiative (ECAI, ecai.org) is an informal international collaboration of several hundred scholars, mainly in the Humanities, seeking to advance learning and research in the humanities and social sciences through increased attention to time and place. Administratively, ECAI is a unit reporting to the Dean of International and Area Studies, University of California, Berkeley. ECAI has ten years experience in outreach promoting Humanities-appropriate digital techniques, the development of a clearinghouse of network-accessible geo-referenced resources, software development, exemplary electronic publications, research and development and 18 international conferences on four continents. (The P.I. is Co-Director of ECAI.)

2. In this, as in prior work, ECAI draws heavily on the technical expertise in Berkeley’s School of Information (formerly School of Library and Information Studies) in the areas of librarianship, information retrieval, metadata research, digital library development, and the like. (The P.I. is emeritus professor and a former Dean).

3. At the Queen’s University, Belfast, the Centre for Data Digitisation and Analysis (CDDA) and Information Services (which includes the University Library) is funded to create a major digital library. The Principal Investigator, Dr. Paul Ell writes, “Under the JISC funding Belfast is charged, possibly quite unusually, with demonstrating that the new e-Library results in enhanced research outputs. We are not, therefore, simply distant providers of resources, we need to engage with scholars and the project is led by an academic.”

4. Berkeley’s Celtic Studies Program, almost unique in the nation in being a degree-granting program, is highly motivated to increase the quality and quantity of digital resources available. Prof. Daniel Melia, an authority on Irish Gaelic and Irish folklore will lead the participation of scholars and students.
5. The *Emma Goldman Papers Project* is a long-established, traditional historical papers archive engaged in collecting, preserving, and encouraging the use of papers by and relating to Emma Goldman (Kovno 1869 – Toronto 1940) an influential anarchist. Like ECAI, it reports to the Dean of International and Area Studies at Berkeley. The archive is engaged in a long-term process of editing historically-important papers for formal publication. Barry Pateman, an experienced editor, is already working with ECAI on biographical mark-up practices and will participate in this project in two ways: He will bring expert professional expertise in editorial best practices and use the archive’s documents relating to the early years of Sinn Fein and other Irish topics to establish whether the tools we will develop could, as we hope and believe, reduce the time, effort, and cost of editorial preparation of scholarly editions. Celtic Studies students and others studying aspects of Irish culture and history would gain first-hand (supervised) experience in scholarly editing.

The work proposed could not be done properly without such a collaboration.

3. **History and Duration of the Project**

This project would start October 1, 2007. It is a new project without previous funding. However, it builds directly on a series of projects, more fully described in Attachment 6: History of Grants.

- **Seamless Searching of Numeric and Textual Resources** (IMLS, 1999-2002) and two DARPA grants developed techniques for computer supported mapping of related terms in quite different vocabularies when searching across multiple vocabularies (category codes, classification systems, indexes, ontologies, subject headings, thesauri, and the like), across different domains and document genres, and cross-lingually. [http://metadata.sims.berkeley.edu/GrantSupported/seamless.html](http://metadata.sims.berkeley.edu/GrantSupported/seamless.html)


- **Bringing Lives to Light: Biography in Context** (IMLS, 2006-2008) is in progress. Personal name disambiguation (sorting out different people with the same name and different names for the same person) is a well-understood problem for which good practices are well-established among librarians and others. But as TEI has documented, practices for encoding biographical text concerning the events, activities, and personal relationships in people’s lives are in a mess. We are developing and will soon demonstrate what we believe will be a generally acceptable and widely interoperable “best practice” approach based on encoding each event and activity as a 4-tuple of kind of event / activity (what); where; when; and with whom, in each case with dynamic links to external resources. [http://ecai.org/imls2006.](http://ecai.org/imls2006.)

- **Support for the Learner: What, Where, When, and Who** (IMLS, 2004-2006) devised techniques for search by time period and an interface supporting differentiated search by topic, place, time and persons. [http://ecai.org/imls2004.](http://ecai.org/imls2004.) The work outlined in this proposal would build directly on this recently completed project. The project would be fully completed by September 30, 2009. No subsequent phases are intended, but, if all successful, the project can be expected to stimulate or influence a number of new initiatives and improved “best practices.”

4. **Enhancing the Humanities through Innovation and Emerging Technologies**

*The need in the Humanities is for context*

As already noted, context matters a great deal in the Humanities.
Why Irish Studies makes a good case-study

Irish Studies as a particularly challenging area of humanities research. It is challenging for a number of reasons. It involves scholars in a range of Humanities disciplines - history, archaeology, human geography, Irish Gaelic, English, education, literature, performing arts, music, politics, and more. Publication of Irish Studies research tends to be equally disparate. While there are a few key journals, most notably those published by the Royal Irish Academy, articles of interest to Irish Studies scholars could appear almost anywhere and library holdings of these resources are comparably sparse and dispersed. This results in the over-use of some materials, and the significant under-use of others and the bulk of the potential resources for scholars are unavailable in the USA. They are in Ireland and, unevenly, in the U.K. Hence, US scholars are starved of research materials. Finally, there is a significant corpus of electronic material which would assist Irish Studies scholars if they could access it easily, but these materials tend to be in bespoke dissemination systems, if online at all, and searching across and between them is impossible without the kinds of search tools and linking methods that ECAI and the School of Information have been developing.

Prior the JISC’s £1.3 million grant to the Queen’s University for the large-scale digitization of Irish studies materials a study was conducted to ascertain whether the investment would likely be justified by future use. A thousand faculty worldwide were identified and surveyed. The results showed very clearly that Irish Studies was not necessarily a discrete subject area/department in its own right, even in Ireland and the UK. Irish Studies may exist as a department with clearly identified staff and education programs, or it could be part of Area Studies, Cultural Studies, Celtic Studies, Literature, History, Politics, Sociology, Conflict Studies, Peace Studies or British Studies and so on. It was necessary to think laterally to try and discover where interested parties might be located or with whom they might be associated.

Interested parties were located throughout the world where you would expect to find them, in the UK, USA, Canada, and Australia, but also in France and Spain (several centers), Scandinavia (very strong interests), Eastern Europe (besides the literature interests, a great interest in politics and economics and Ireland's development in the later half of 20th century seen as an example to new EU entrants), Japan (literary interests), and South America (history and literature). Where scholars are pursuing research interests or even supporting teaching programs in institutions and departments without a clearly defined Irish Studies agenda collegial and even special subject support in their institution's Library will also be lacking, thus exacerbating their isolation. Some academics (US based) commented on the lack of credibility for certain subject areas in Irish Studies in the US because of the lack of resources to support the subject for teaching or for research.

Resources relating to the Irish Studies are not neatly tied up in one or two electronic database “libraries” (which are standard resources now in most disciplines), they simply don't exist. A few references may be found in the standard bibliographic databases, emphasizing the interdisciplinary and multidisciplinary nature of the subject. Checking COPAC www.copac.ac.uk for titles proposed for scanning revealed that quite a few of them were available in UK institutions - but not in any meaningful concentrations. They were widely scattered in location and coverage wasn't comprehensive. At The Queen’s University, they are all on site, together, with comprehensive runs, but this does little for the rest of the world, unless they travel to Belfast. Resource discovery is a major problem as these journals do not have indexes of any kind - in the journals themselves, published cumulatively or represented in any bibliographic database (standard tools now not just for researchers but for undergraduates too).

Respondents usually asked whether small society and local history publications will be included because important information can be found in these publications, and not just recent publications but also those of the 19th century. They will be included and this confirms the importance of place and time for bibliographic access.
Cultural agencies supporting the Humanities

Our primary concern is with scholars at all levels, but we see, as important by-products, significant potential benefit in aiding and supporting the work of editors of scholarly editions of historical papers; in examining how the functionality of the library reference collection could be restored in the digital library environment; and in making scholarly monographs more useful and more valued.

Strategic considerations for Humanities Initiatives

1. Tools and techniques will be developed for very wide potential application.
2. This is an important contribution to infrastructure design;
3. Standards and “best practices” are adopted after their advantages become apparent. This project, if successful, would provide an increased motivation for the digitization of auxiliary resources across the Humanities and for the adoption of standards, such as Z39.50 and structured URLs, that can support remote search.
4. Would move the use of important tradition resources genres in contemporary peer production modes.
5. Provides motivation to adopt open access reference resources.

Research Facilities available

The Queen’s University, Belfast, has outstanding research facilities for digitization at the CDDA and the University Library has exceptional resources for Irish studies. ECAI and the School of Information have good computing support and will acquire the substantial additional storage required.

5. Methodology and Standards

The project is designed in components, but, first, an introduction to what we mean by faceted searches and advanced search techniques, named entity extraction, and federated search.

Faceted searches and advanced search techniques

The phrase “advanced search techniques” ordinarily refers to complex operations on text using statistical techniques, natural language processing, and the like. Well-known examples include Latent Semantic Indexing and Probabilistic Retrieval Models and many variations on these and other approaches fill the pages of Information Processing and Management and similar journals. We mean no criticism of these techniques, but observe that purely textual approaches are insufficient, as we will show in the example of searching by place.

The standard practice in libraries and elsewhere of using places names has limitations. A place may have multiple names; names are often ambiguous; names change; places themselves and their boundaries change with time. A place name gazetteer (an authoritative list on named places, indicating type of place and spatial coordinates (latitude and longitude) is important not only for disambiguation but also because it enables map display. Figures 1 & 2 show what can be done when a library catalog is connected with a gazetteer. In Figure 1 a catalog search for “Folklore” initiates a map that shows the locations that the folklore books retrieve. Figure 2 does the reverse, by allowing the user to specify (draw) a region and specify some geographical feature (here capital cities in South America) which will initiate a search in a library catalog for these kinds of places in that region.
Figure 1. Geographical display of search results.

Figure 2. A map display for specifying a geographical query.
Just as maps are important for dealing with places, time-lines are valuable for dealing with time periods. A combination of place and time can be handled using both. In Figure 3, a searcher interested in events has zoomed in geographically and chronologically (using a time-line) to central Ireland in the late 17th century. An icon at Limerick signifies an event represented in the Time Period Directory that we had constructed in a previous project. Clicking on this icon brings up a record for the Siege of Limerick in 1690. A further click generates a live search of the Library of Congress online catalog for material on this event. The first item retrieved is shown. In just this one record, the subject headings identify the military context “War of 1689-1691” and someone importantly involved, William III. Clicking on those underlined subject headings or the author or the Wikipedia icon would automatically extend the search accordingly to provide more background information.

There are two neglected problems in dealing with people: Finding common ground and some best practices for encoding and marking-up text concerning the events and activities in an individual’s life; and vocabulary and display methods for characterizing the many different interpersonal relationships that exist beyond the basic family and biological relationships. (Both are being addressed in our current project Bringing Lives to Light). Dealing with Institutions raises problems resembling those of persons.

Our view is that this kind of contextualizing assistance for any word, name, date, event, person or institution should be a standard feature in the digital humanities environment. The kinds of advanced tools we have above are needed and we propose to demonstrate their use in the case of Irish studies.
Fig. 3. Geo-temporal search for events: The siege of Limerick, 1690.

**Named entity mark-up**

During the past decade, within a field of Linguistics called “computational linguistics,” tremendous strides have been made in the automatic recognition within text of the names of persons, places, and organizations. This area of research and development is called Named Entity Recognition (NER) and it has advanced through evaluations such as the Automatic Content Extraction (ACE) conferences [http://www.nist.gov/speech/tests/ace/](http://www.nist.gov/speech/tests/ace/) run by the National Institute of Standards and Technology. Groups within the Association for Computational Linguistics undertook in 2003 a Named Entity Recognition “shared task” wherein groups attempted to do language independent NER [http://www.cnts.ua.ac.be/conll2003/ner/](http://www.cnts.ua.ac.be/conll2003/ner/) against a set of standard texts. The automatic recognition results from each group were then matched for correctness against manual markup of the named entities within

the texts. Libraries of software for automatic named entity recognition are already available in the public domain for scholars – one such is alias-lingpipe (http://www.alias-i.com/lingpipe/) which was used by some participants in the 2006 GeoCLEF geographic information retrieval from text evaluation (http://ir.shef.ac.uk/geoclef/)3 chaired by co-PI Fredric Gey. The potential for NER to augment text is described by Crane.4

“In a sample 300 volume, 55 million word collection of nineteenth-century American English, automatic named entity identification has added 12,000,000 tags. While this collection focuses on name rich historical materials and includes several reference works, this system already discovers thousands of references to named entities in most book length documents.”

The question immediately arises as to what tests have been applied to see if these 12 million tags are correct? What we envision is not only automatic “suggested” named entity markup, but also an environment in which Irish scholars will be able to view the named entity tags within the digital source document and validate their correctness, as well as add additional markup as needed with the automatic editing device which contains embedded markup tools.

Federated search

Search and data access of remote systems is made possible thanks to several established protocols for distributed search and metadata harvesting. We plan to use the Z39.50 information retrieval protocol, which provides direct search of many (if not most) library catalogs available today, as well as to other interesting resources such as the UK Archives Hub. Some newer systems now also support the new descendent protocols of Z39.50, known as SRU (Search / Retrieve via URLs) and SRW (Search / Retrieve Web Service), utilizing CQL (Common Query Language), a standard query syntax for representing queries. The Library of Congress serves as the maintenance agency for these standards (http://www.loc.gov/standards/sru/). Some collections of interest (Such as the eScholarship editions at the California Digital Library) use SRU for access.

Other collections are available for metadata harvesting using OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), including the eScholarship editions from CDL, the Museums and Online Archives Collaboration (MOAC) and the Online Archive of California(OAC), and the collections of full-text books and other materials housed at the Internet Archive.

Project Plan

The project has been planned with the following components:
1. Needs Assessment and Evaluation
2. Tools and Techniques.
3. Test-bed.
4. Use and applications.
5. Documentation and dissemination.

1. Needs Assessment and Evaluation

Needs assessment and evaluation will repeat the methods that were found satisfactory in the Support the User project. This project, like that one, is exploratory and so exploratory approach will be


used. We make little distinction between needs assessment and evaluation, since the latter should be in terms of the former.

At a narrowly technical level we will report on what techniques were used, whether they worked and how well they performed in a technical sense. The emphasis will be on functionality achieved rather than speed and efficiency of performance, unless performance issues interfere with use.

We will learn from users through lightly structured interviews and conversations to ascertain the needs, skills (skills in using computers for their work and familiarity with auxiliary resources), and the preferred working habits of actual and potential users. Questionnaires will also be used sparingly in the evaluation.

The nature of the contents of the test-bed limits the likely range of volunteer users. All use will be entirely voluntary. We will start with the faculty and students in the Celtic Studies Program and then gradually expand outwards as time and resources permit, first to interested faculty and graduate students on the rest Berkeley campus, then include anyone on or off campus, at Berkeley and elsewhere, interested in the materials in the test-bed and/or the tools being developed. We will work to include a range of different people and make a point of including some users in the general public, “amateurs” in the best sense.

We envisage making use of prototype tools operating on the test-bed openly available on the Web as soon as they become fit for use.

2. Tools and Techniques

We will develop two tools and a technique, which we are calling a Context Finder, a Context Builder, and a Context Provider, respectively.

2.1. Context Finder

The Context Finder is a search interface, with some special features. It would sit in front of or next to any text being read. (Fig. 4 shows one initial design.)

(a) A reader is examining a text in a window. If s/he is curious to know more about any word or phrase encountered, the cursor is used to click on it or block them, and a separate search box opens up with the blocked word(s) in it.

(b) The reader can modify, augment, or replace the word(s).

(c) The query is categorized by “facet,” i.e., whether this search for a topic (e.g. Weaving); a place (e.g. Rathlin Island); people (i.e., an individually named person); a time period (a calendar date, a named time period, e.g. War of 1689 – 1691, or an event used temporally “after the Famine”); or a named institution (e.g. The Royal Irish Academy). (These facets resemble the sections of a reference collection). The choice, made by clicking on a button, determines the kind of search and the type of auxiliary resource. (Some resources, notably encyclopedias, are suitable for multiple facets, of course.) Depending in the facet chosen, a ranked menu of available auxiliary resources appears in a window, one is chosen, and the interface generates and sends off a search. For places, one or more gazetteers which could retrieve a descriptive record for that place(s) with that name and display locating it in map. The map display could be used to extend a geographic search to nearby places by indicating the areas of interest with a cursor. For events and named time period, one or more time-lines, chronologies and/or named time period directories would be offered for search and the result of the search offered as a list and/or as a time-line display. Likewise, separate options for Persons (Biographical dictionaries); Institutions (directories); and Topics (library catalogs). Of course encyclopedias, library catalogs, and web search engines are useful for multiple purposes. The menu of resources can be expected to improve as new resources become available. The options and ranking can be expected to be different for different kinds of users – just as the selection of resources in a reference collection is situational. When the text
already contains TEI or other XML compliant mark-up the interface would use the mark-up to pre-set default facet and resource buttons.

(d) The search result appears in a window below. A fragment of context, helpful, we hope, has been found.

2.2. Context Builder

The Context Builder is an enhanced Context Finder. The difference is that the Context Builder can save the choice of facet (place, topic, person, . . .) and the search path (when an acceptable search result has been found) and then can store both as XML mark-up with, minimally, two components. The first, encodes the word or phrase of interest as being a place, a person, or whatever; and the second a dynamic hyperlink capable of generating the same search on demand, when worth saving. The former can be viewed as the lightest level of conventional content mark-up; the latter can be thought of as analogous, in a digital environment, to a footnote citing a source at the bottom of a printed page. A difference is that a hyperlink doesn’t merely cite, it can take you to a source. (We have embedded dynamic hyperlinks that send federated search queries to a specific auxiliary resource in prior projects, notably ECAI Iraq, a temporal-spatial portal into digital resources about history, cultural sites, archaeological excavations and heritage preservation initiatives using interactive, time-enabled thematic maps. See, for example, the first three links at http://ecai.org/iraq/SiteName.asp?SiteID=7).

So, the next time (or the next reader) would have a pre-prepared live search for the same auxiliary source, perhaps by now more up-to-date.

This markup can be embedded in the text being read, if the reader has the right and the means to do that; or it can be stand-off mark-up stored separately but with a pointer to the location in the text. We assume that text already marked-up could be improved with additional or substitute mark-up; that both forms of mark-up could be shared and enriched collaboratively in a “peer production” mode; and that different communities of users would want and need dynamic hyperlinks to different resources according to their knowledge and the nature of their interests. So texts enriched in this way (or stand-off mark-up) could be shared, but not all readers would want the same choice of auxiliary resources.
2.3. Context Provider

The Context Provider, more a technique than a tool at this stage, would take the mark-up applied to a text, rearrange it, and the apply the details to enrich the target auxiliary source with links back to the text. In the example given above under “Enrichment of auxiliary resources,” dynamic links to a gazetteer added to the place names in a text are copied, sorted by place name, and then added to the gazetteer. The gazetteer now has the additional amenity of citing mentions of places in a text. We would create some examples to show what could be done. This process is highly cumulative as the web of connections would begin to build up in a way hardly imaginable in the print environment.
3. Test-bed

Resources to be used – Test-bed texts

The primary test-bed will be a digital version of the core literature on Irish studies. The Queen’s University, Belfast, has been funded by JISC to select, find, and digitize (scan and OCR) an Irish Studies corpus of 100 journal back-files, 200 books, and 2,500 manuscript pages. Once digitized, these will be made publicly available world-wide by JSTOR, an organization set up by the Mellon Foundation specifically to make digitized versions of back files of leading scholarly journals available. In the meanwhile, this corpus will become available as a test-bed for this project for the purposes of the project and for at least the duration of the project. [. . .]

During the course of the project The Queen’s University, the University of California Press, the California Digital Library and possibly other sources will make additional resources available to provide diversity as and when desirable for the purposes of the project.

Resources to be used – Auxiliary resources

Auxiliary resources are essential for this project. Existing openly accessible resources will be used whenever possible. The immediate difficulty is that as of early 2007, the range of specialized auxiliary resources for Irish studies that are openly available on the Internet and willing and able to respond to federated searches is small but growing. We don’t need much for technical proof of concept, but we do for credibility among scholars. There is a chicken-and-egg problem, here: the utility of resources cannot be demonstrated until they become available, but they are unlikely to become available until their utility has been demonstrated. (One of our expectations is that even a modest demonstration will inspire a more rapid development of accessible auxiliary tools.)

We will address this problem by using more general resources accessible through federated search (such as the Library of Congress online catalog, the Wikipedia, and The Royal Historical Society Bibliography) pending the use of more specialized resources, and where we can find no other feasible option we will create resources as minimally necessary for proof-of-concept purposes. (Prof. Melia and Ms Wildy and their contacts are well-qualified to provide guidance.)

For example, geographic search requires a place name gazetteers. We will start with the records made available by the U.S. Board on Geographic Names (BGN) through the National Geo-Intelligence Agency (previously NIMA) with >25,000 place names in the Republic of Ireland and more for Northern Ireland. For Northern Ireland we will have use of the Ordnance Survey place name records. We expect to need additional resources for historic Irish place names, so we expect to draw on the gazetteer prepared for the 1861 Irish Census of Population, half of which has already been scanned and OCR’d by CDDA into an Excel file. What remains to be done is to scan and OCR the rest, proof-read and correct, and the addition of spatial coordinates. Maybe, if we run under budget and/or other funds were found, we or others could digitize the Hogan Onomasticon Gaelicum. (An initial experiment was encouraging.)

We may need to develop or adapt a named Time Period Directory for reigns, administrations, wars, and other events used to indicate time, at least sufficient for demonstration purposes. (In the event that it would be cheaper or otherwise advantageous for such work to be performed at Belfast or elsewhere, we would seek prior permission from NEH to arrange a subcontract from within the existing project budget.)

Through our collaboration with the Bringing Lives to Light: Biography in Context project, we have access to the entire contents of the UK Archives Hub, a distributed database comprised of EAD records of the archival collections of over 150 institutions in the UK, including the National Libraries of Scotland and Wales. These are being indexed and "text mined" by the biography project and will be made available for our use as well.

Plain text, mark-up, and mark-up editing tools.
Initially the project will use plain text, i.e. without any mark-up or ignoring any mark-up. Our interest in mark-up arises because a very light level of mark-up, merely identifying place names, personal names, time markers (dates, named time periods) and names of institutions allows the search portal to identify the facet and prompt for search accordingly. After experimenting by hand and making a concordance, Named Entity Extraction techniques and mark-up editing software would be used to prepare the text by identifying proper names and other recognizable terms by a combination of natural language processing and reference to lists. This is a relatively undemanding requirement because we envisage computer-aided use not fully automated use, and it is doubtful that deeper differentiation or dealing with the difficult problem of anaphora need be addressed.

The different facets (what, where, when and who) could be indicated by color coding the text or its background, possible with the degrees of confidence indicated by the darkness of color or texture to generate very light mark-up to plain text.

We would extend the Context Finder and the Context Builder to use already marked-up text (pre-existing TEI or XML, or our additional mark-up) as a prompt.

The addition of mark-up also supports more effective retrieval of books through special retrieval techniques that exploit the structure of the documents in order to identify appropriate elements (such as chapters, pages, and paragraphs) instead of delivering the whole book in response to searches. We will be participating in the Book Retrieval Track of INEX (The INitiative for XML retrieval) where we will be exploring these retrieval approaches using a corpus of over 30,000 books from Open Content Alliance scanning efforts.

4. Use and applications

4.1. Scholarly use. The primary target would be faculty, students, and others interested in the Humanities, in our case Irish studies. Volunteers would be recruited as soon as an early version of the Context Finder becomes available.

4.2. Applications in Cultural agencies

These applications will be addressed flexibly following the development of the tools and techniques. We have identified three. This is a secondary aspect of the project, but we sense the possibility of substantial and widespread benefits.

4.2.1. Editing scholarly editions of historic papers. NEH knows well the high and protracted costs of producing scholarly editions of historically-important papers. At Berkeley, the Mark Twain papers and the Emma Goldman Papers projects are examples. We believe that the Context Finder and, more especially, the Context Builder would provide a small but significant productivity tool for the editorial process. In fact, the idea for a Context Builder originated in a discussion of this need with professional editor Barry Pateman, who will take the lead in helping us to examine this need. He will help us test this application. Even a small increase in efficiency or cost-effectiveness could have a substantial cumulative effect in this hard-to-fund work.

4.2.2. Implications for library services. The functionality of the traditional reference service has not yet made an effective transition to the digital environment and is fading in the print environment as library users shift to working digitally and reference librarians are decreasingly in evidence. The dilemma has been referred to as the “9-to-5” problem: The reference librarians are liable to be available but underutilized from 9:00 a.m. to 5:00 p.m.; students prefer to work from their laptops in their dorms from 9:00 p.m. to 5:00 a.m. Meanwhile, professional library literature focuses on converting reference desks into 24/7 call centers, which may be desirable but does not scale. What is wrong with this picture? The answer is that the functionality of a reference collection of auxiliary resources can and should be designed afresh in the context of digital technology. One answer lies in reversing the relationship between source and contents: With the technology of the codex, you have to find the source and then look at records in it;
Google-style, you find the records, then consider the sources they come from. P.I. Michael Buckland, who has written extensively on the re-design of library services, will take the lead in this.\(^5\) It will be mainly a matter of demonstration, explanation, and discussion, and part of the Documentation and dissemination.

4.2.3. Enhancement of digital publications. The specialized monograph published by a university press has played a large role but publication is decreasingly viable. There are many factors at play here. Our suggestion is that a monograph that is enriched with contextualizing mark-up would be more attractive and more reader-friendly than one that is not. Further, thanks to the Context Provider idea, the mark-up for the book, could also enrich auxiliary tools and make that monograph more visible, more useful and more used. Our project would provide the tools and show the opportunities. We intend to explore this on a small scale.

Controlled comparisons of the reading, by volunteers, of digital texts without the benefit of project deliverables will be compared with texts for which a Context Builder had been used to provide rich contextualizing mark-up had been added. This will be part of the Needs Assessment and Evaluation.

5. Documentation and dissemination.

See next section.

Disclaimers

1. This project is about tools, infrastructure, and the digital humanities environment. It is not a publication project and the purpose is not to provide a permanent public information service, even less to encroach on any existing or emerging service providers. It is about how better use could be made of existing and future resources, with detailed case-study.

2. Relevant digital corpora are becoming available in abundance, but auxiliary resources that are digital and open to federated search access are still scarce. It is a chicken-and-egg situation. Only demonstrating the value of such resources will encourage their development: Demonstrating their value is difficult when few are available. We’ll use what we can for proof-of-concept purposes and improvise resources only as necessary.

6. Dissemination

Dissemination will be through multiple channels: an informative project website; freely web-accessible prototypes; technical reports; papers presented at conferences and published in conference proceedings; and articles in leading professional and technical journals. Dissemination will be aimed at librarians, scholars and educators in the humanities, and developers of digital library and museum projects. The P.I.s and partners have well-established records of dissemination and we invite inspection of the Support the Learner project website at http://ecai.org/imls2004 which includes an openly accessible working prototype search portal and more than twenty different publications. Project presentations were made in at least seven states and also in China, the U.K. and the Netherlands. The Advisory Committee is seen as important for two-way communication.

A “Lessons learned” white-paper will be published. [ . . . ]

Concluding Comments

Digital resources are becoming available for the Humanities on a grand scale. This proposal addresses the challenge of developing better ways to make more effective use of these resources and, especially, helping people to understand the whole by learning about context and relationships.

In this proposal narrative we have concentrated on dealing with the geographical context, which serves well to illustrate the nature of what we intend. But the project is equally concerned with the temporal aspects of context (time, time-lines, and the use of events and named time period and temporal markers), biography (persons, prosopography, and interpersonal relationships), institutions, and topics -- and of combinations (e.g. geographical “life paths”). These other aspects could have been addressed in comparable length, but not within the 25 pages allowed.

We have used the unifying phrase “Irish studies,” but, in practice, scholars interested in Ireland are exceptionally dispersed and isolated. There is no prospect of a unifying disciplinary digital library as in Chemistry or Computer Science; a different, subtler approach is needed, which makes the choice of “Irish studies” more interesting, more needed, and more likely to provide a model for other fields in the Humanities.

We have chosen to focus on Irish studies and a new digital Irish corpus because there is a need, because it seems a very suitable test-bed for the technical challenges of tool-building, and because we have the requisite academic expertise. Nevertheless, the technical approaches and tools to be developed could be applied to other collections and communities and, except in minor details, are not specific to Irish Studies.

If, as Bacon claimed, knowledge consists in an understanding of causes and effects, then seeking to make contextual information easier to find becomes a prudent strategic move in any program concerned with advancing knowledge.