White Paper Report

Report ID: 98475

Application Number: HD5044008

Project Director: Andrew Jewell (ajewell2@unl.edu)

Institution: University of Nebraska, Lincoln

Reporting Period: 9/1/2008-3/31/2010

Report Due: 6/30/2011

Date Submitted: 6/27/2011

White Paper

Grant #HD5044008

The Crowded Page

Andrew Jewell and Edward Whitley

University of Nebraska-Lincoln (sub-award to Lehigh University)

June 30, 2011

Description of Project

The Crowded Page is an Internet-based humanities computing project whose goal is to create data-mining and visualization tools that will allow researchers to map out the intricate connections between the members of artistic and literary communities. In most accounts of literary and art history, a work of art or literature is said to be the product of a single creative mind. In an effort to make visible what is often obscured in traditional histories of art and literature, *The Crowded Page* seeks to take advantage of the unique capabilities of the digital medium to foreground the ways in which a complex network of friends, editors, neighbors, lovers, and fellow artists and writers informs the creative process.

The Crowded Page has created relational databases that contain information about discrete creative communities and designed an interface for this database that allows researchers to discover and visualize the different kinds of connections that link the members of communities together. Scholars can use the digital tools at *The Crowded Page* to negotiate thousands of pieces of data and then visualize the relationships between this data in ways that would uncover how creative communities function as a collective organism to produce works of art and literature. The visualization and discovery tools on *The Crowded Page* will allow students and scholars alike to understand the nature of authorship in a truly revolutionary way.

At this stage of the project, we have created a "proof of concept" webpage at http://www.crowdedpage.org that provides a glimpse into the kind of powerful interaction with the data that we imagine in the fully formed site. We have two major datasets represented in the current instantiation of the web page: the artist community centered around Greenwich Village in Manhattan, New York, between 1910-1920, and

the community that gathered around Charles Pfaff's beer cellar in lower Manhattan in the mid-nineteenth century. Of these two, the Pfaff's dataset, drawn from Whitley's *The Vault at Pfaff's* (http://digital.lib.lehigh.edu/pfaffs/) is more complete. The Greenwich Village dataset, though reasonable rich with hundreds of people, is not complete enough at this time for one to use it to do actual research on this community. Instead, the initial version of *The Crowded Page* is designed to share some of the promise of this methodology and built interest for future instantiations.

A description of the activities and products during the grant period are listed below.

A. Project Activities

In September 2008, Jewell traveled from the University of Nebraska-Lincoln to Lehigh University to meet with Whitley and Heflin and to hold preliminary discussion about how Heflin's prior work on the Semantic Web could help to accomplish the goals of *The Crowded Page*. This initial discussion set the stage for a series of phone conferences and email exchanges that would continue throughout the academic year of 2008-09. One of the main points of discussion in these conversations was the data model that Whitley and Jewell would need to create in order for their data to be compatible with Heflin's Semantic Web search technologies. It was determined that an important component of this data model would be a controlled vocabulary for describing the relationships between the members of the literary communities that Whitley and Jewell were studying. Existing vocabularies for collecting such data are both thin and flat, defining relationships almost exclusively on the basis of a single factor such as "knows" or "is friends with," and it is important for us to be able to

capture a much more sophisticated representation of the varied relationships in our focus communities.

We dedicated a period of time to researching existing models for controlled vocabulary in defining relationships, and found "Relationship: A Vocabulary For Describing Relationships Between People" <http://vocab.org/relationship/.html>, which provided us with a starting point for creating our own controlled vocabulary. Through an iterative process that involved circulating this vocabulary list among different groups of people at both Lehigh and Nebraska, we settled on a list of controlled vocabulary. (See Appendix One.) For the next stage of the process, we discussed the best way to collect and record the data. Some of our options included: working natively in RDF (the language of the Semantic Web), working in XML and then exporting to RDF, working in a MySQL database and then exporting to RDF, and expressing the data in TEI P5 Prosopographic XML. After discussing our options, we determined that the best course would be for the both teams to work with their own, different MySQL databases for data collection and export the data directly in RDF or in XML for delivery to Heflin (and probable transformation to RDF). There were a number of reasons for this decision, the most compelling of which was that, given the exploratory nature of this stage of the project, it would be important for us to see the strengths and the weaknesses of a number of different approaches. We also wanted to see how flexible Heflin's Semantic Web technologies could be when working with data sets created in different contexts by different users.

Once we determined the data models that we would be working with, we set out to gather the data. Staff at Nebraska's Center for Digital Research in the Humanities built a complex database for collection of data specific to the needs of the *Crowded*

Page, and Jewell applied for and received an internal grant from Nebraska to hire a research assistant to gather data about the Greenwich Village bohemians of the early-twentieth century in the summer of 2009. Jewell and his research assistant built a database with hundreds of people, groups, works, relationships, and sources; though not a full enough database to use as a research dataset for study of Greenwich Village in the 1910s, it is sufficiently large and robust for the proof-of-concept goals of our Start-Up grant. Whitley similarly applied for and received an internal grant from Lehigh to hire a research assistant to modify the data about the Pfaff's bohemians of the mid-nineteenth century that had already been gathered at an earlier date. As fall semester 2009 began, both Whitley and Jewell had completed their data samples, converted them to XML or RDF for delivery to Heflin (see Appendix Two for a sample RDF file from Whitley), and passed them off to Heflin so that he could run a number of sample inquiries through the data using his Semantic Web search technologies.

In late 2009, Jeff Heflin led a team of computer scientists at Lehigh in formatting the two different data sets from Whitley (on the Pfaff's bohemians) and Jewell (on the Greenwich Village bohemians) such that the data could be made usable for Heflin's Semantic Web search tools. In addition to making this data searchable with his Semantic Web tools, Heflin's team prepared the data such that it could then be handed off to the team at Nebraska for use in a proof-of-concept visualization tool.
Heflin's team needed the Pfaff's and Greenwich Village data to be in well-formed RDF. The Pfaff's data was already in RDF, but the provenance information needed to be modified in order to work with the Semantic Web tools. In conversation with Whitley and others at Lehigh, Heflin's team determined that the Pfaff's data needed to add a new category to the ontology to cope with the author, editor, and creator fields in the data. As

such, the ontology "hasCreator," with the sub-properties "has Author" and "hasEditor," was added. In order to make the vocabulary more complete, the property "pubYear" was also added.

In order to work with the Greenwich Village data, Heflin's team took the MySQL XML files from Nebraska and transformed it into RDF such that it could be processed by the search tools. Then Heflin's team took both the Pfaff's and Greenwich Village data--both of which had now been formatted as well-formed RDF--and loaded them onto their database, at which point they were able to use their Semantic Web API to retrieve the provenance of the Pfaff's and Greenwich Village data from the database.
Heflin's team then communicated their accomplishments with the Nebraska team such that the Nebraska team could begin work on a visualization tool that would draw from the Semantic Web API created by Heflin.

In 2010-2011, the team at Nebraska worked with the tools created by Heflin's team to generate a proof-of-concept web page demonstrating the power of the Crowded Page concept and design. This site, available at http://www.crowdedpage.org, is an effort to build an interface for the powerful RDF query tools that ingested the Pfaff's and Greenwich Village data. This process has involved considerable work from Karin Dalziel (Digital Resources Designer at UNL's Center for Digital Research in the Humanities) to construct an initial interface that made the complex interactions with the data comprehensible and attractive. Simultaneously, Keith Nickum (Programmer at UNL's Center for Digital Research in the Humanities) has worked to take the API provided by Heflin's group and get it working on a server at the University of Nebraska-Lincoln. It has taken considerable trial-and-error effort to get the system working in a new environment and to make the interface designed by Dalziel operate with the migrated

database. The last several months of this stage of the project have been dedicated to solving these problems.

B. Accomplishments

In our proposal to the NEH, we stated that *The Crowded Page* would "create a relational database that will contain information about discrete creative communities and then design an interface for this database that would allow researchers to discover and visualize the different kinds of connections that link the members of communities together." Pleasantly, that is precisely what we have accomplished during the grant period.

More specifically, we have:

1. Developed an ontology of relationships designed for gathering data about creative communities in the past
2. Created a dataset of relationships centered around the Greenwich Village neighborhood of New York City between 1910-1920. This work was funded thanks to a grant-in-aid by the University of Nebraska-Lincoln Research Council. Sabrina Ehmke Sergeant, who at the time of work was a Ph.D. candidate in English at UNL, was hired with the Research Council funds and was central to researching and building this dataset.
3. Provided both the Greenwich Village dataset and the already-existing Pfaff's dataset to Jeff Heflin's group at Lehigh University for transformation into an implementation of RDF based on the *Crowded Page* ontology.

4. The combined datasets (both Pfaff's and Greenwich Village) contain about 607 different individuals and 1845 editorially declared relationships between those individuals.
5. Modified existing RDF query tools to allow for sophisticated analysis of the datasets, specifically for finding and articulating the complex web of relationships implied in the data.
6. Migrated the API created by Heflin's team to servers at UNL for web delivery of the data and the data analysis.
7. Built an interface to allow users to browse and search the data in customized ways.

In accomplishing these tasks, we have, like all such projects, been repeatedly faced with challenges, both expected and unexpected. For example, despite the additional funding from the UNL Research Council, we were never able to build a dataset on Greenwich Village big enough to allow for real research on that topic. We knew that this would only be a proof of concept and not a published research tool, and yet we learned more explicitly how much time and effort it would take to develop the data for a truly robust and usable dataset. Such a project would need to be the major focus of a group of scholars for an extended period.

Another challenge we have faced is integrating kinds of data relationships we thought crucially important into existing RDF frameworks. For example, we were committed to providing a source for every claim made in the database for a relationship between two people. As humanities scholars, we are fully aware that backing up editorial claims is extremely important both to demonstrate authority and to provide the

audience with a mechanism for checking the claims. The existing RDF frameworks aren't designed with historical sourcing in mind, and we had to work to find strategies for attaching a source to the declared relationships in the database.

A final challenge has been just getting the systems to work. Given the dispersed team and varied servers, and, of course, the innovation inherent in the project, it has taken a lot of effort on behalf of programmers and designers to make www.crowdedpage.org operate effectively, even as a proof-of-concept.

C. Audiences

The primary audience we wish to attract are scholars who work in the digital humanities as well as literature scholars more generally. As a "proof-of-concept" publication, we expect the primary audience of *The Crowded Page* to be those interested in how digital technologies can enhance the research of historical communities. However, we will also be interested in reaching scholars in literary and history who have an interest in the content of our datasets. We have also been approached by Laura Mandell to write an article about *The Crowded Page* for *The Poetess Archive Journal*, an online scholarly publication that attracts scholars equally from the fields of the digital humanities and nineteenth-century literature. Our involvement in Micki McGee's NEH-funded Compatible Databases Initiative (described in greater detail in "Continuation of the Project" below) will also allow us to have access to a broader group of scholars both in the digital humanities and in fields of literary study outside the nineteenth-century United States.

D. Evaluation

Once the website for *The Crowded Page* goes live in July, there will be a series of internal and external evaluations. Teams from both institutions--Lehigh and Nebraska--will be asked to use the site and comment on its strengths and weaknesses. The Lehigh team, which developed the Semantic Web search technologies running on the back end of the site, will evaluate the effectiveness of the visualization and interface designed by the Nebraska team, particularly with respect to how the visualization and interface make use of the Semantic Web search technologies. The Nebraska team will also have the opportunity to respond to the Lehigh team regarding the strengths and weaknesses of the Semantic Web technologies that they were provided with.

External evaluation of *The Crowded Page* will take place on both an informal and a formal basis. Informally, the co-PIs will confer with colleagues in the digital humanities who are a part of their academic networks, such as those involved with NINES: Networked Infrastructure for Nineteenth-Century Electronic Scholarship, and the Digital Americanists. (Jewell is on the board of NINES, and both Jewell and Whitley are former presidents of the Digital Ameicanists scholarly group.) Jewell will also seek out commentary from colleagues in the Center for Digital Research in the Humanities at the University of Nebraska who have not been heretofore involved with *The Crowded Page*.

More formally, both Jewell and Whitley will be attending an NEH-sponsored meeting of The Compatible Databases Initiative (see below in "Continuation of the Project"), where they will have the opportunity to receive feedback on *The Crowded Page* from a group of scholars in the digital humanities who are also working to develop tools for visualizing and analyzing literary communities. This intense, three-day meeting

with scholars who share the same goals as Jewell and Whitley promises to provide the best, most critical evaluation of *The Crowded Page* that we could hope for.

E. Continuation of the Project

The short-term plan for continuing the project involves working with Professor Micki McGee of Fordham University on her NEH-funded project, The Compatible Databases Initiative: Fostering Interoperable Data for Network Mapping and Visualization. Professor McGee has invited us to be a part of a working group of scholars in the digital humanities who are developing visualization and discovery tools for mapping out the networks of relationships in literary communities. We will be joining Professor McGee for a September 23-25, 2011 meeting that will be held at the Yaddo artists' community in Saratoga Springs, NY. We are looking forward to sharing what we have learned about how to use digital tools to study creative communities with this group and we hope to learn a great deal as well.

There has recently been an increase of interest in social networks, visualization, and data mining in the digital humanities recently--with projects like the Berkeley Prosopography Services and LOD-LAM (Linked Open Data in Libraries, Archives, and Museums), for example--and we want to make sure that *The Crowded Page* can benefit from the insights of this community of scholars. We are confident that working with Professor McGee and the other members of the Compatible Databases Initiative will help us not only to identify the best practices in the field, but to point us in the right direction to take for the next stage of our work. Specifically, it has always been our goal to have *The Crowded Page* be more than just a profile of the bohemian communities of nineteenth- and early-twentieth-century New York: we have wanted to create a space

where scholars working with literary communities of any time period or location can use our digital tools to help them explore the relationships in these communities (see "Long-Term Impact" below for more detail). Our hope is that meeting with scholars from the Compatible Databases Initiative will give us a better idea for how to realize this goal.

F. Long-Term Impact

The long-term goal of *The Crowded Page* is to have the site provide literature scholars with a suite of digital tools that will allow them (a) to create a database of information about a literary community, (b) to use a controlled vocabulary to describe the relationships between the people and texts in that community, and (c) to use a visualization and text exploration tool to make new discoveries about the workings of that community. We imagine that the site will be useful on two levels: (1) individual scholars will be able to learn more about the particular literary community to which they apply this suite of tools; (2) visitors to the site can look at the various communities profiled on *The Crowded Page* and draw larger inferences about the workings of literary communities as a whole. That is to say, the site will have both micro and macro applications: the discovery tools will be powerful enough to identify the meaningful connections in a single community, and the collective impact of having multiple literary communities featured on the site will reveal patterns about the workings of literary communities that cut across broad historical and geographical boundaries.

G. Grant Products

The major grant product is www.crowdedpage.org, which is a interface illustrating the work we've done and providing a proof-of-concept for this kind of digital humanities research. More abstractly, another product of this grant is the strategy, experience, and knowledge generated by completing the work of the grant. Most immediately, these "products" will be shared as we contribute information and insight to Micki McGee and her team as part of the NEH-funded project Compatible Databases Initiative: Fostering Interoperable Data for Network Mapping and Visualization.

Appendices

1. Controlled Vocabularies Developed for *The Crowded Page*
2. Sample RDF/Provenance Process
3. Database model

Appendix One:
Controlled Vocabularies Developed for *The Crowded Page*

Controlled Vocab for Relation of Person to Person

Term	Description
acquaintanceOf	A person having more than slight or superficial knowledge of this person but short of friendship.
antagonistOf	A person who opposes and contends against this person.
collaboratesWith	A person who works towards a common goal with this person.
colleagueOf	A person who is a member of the same profession as this person.
worksWith	A person who shares a workplace or specific professional association with this person.
friendOf	A person who shares mutual friendship with this person.
lifePartnerOf	A person who has made a long-term commitment to this person
livesWith	A person who shares a residence with this person.
neighborOf	A person who lives in the same locality as this person.
immediateFamilyOf	A person who is either the parent, child, or sibling of this person.
romanticWith	A person who is involved romantically and/or sexually with this person
familyOf	A person is part of the same family as this person, but is neither parent, child, sibling, or spouse.
spouseOf	A person who is married to this person

Hierarchical Breakdown of Controlled Vocabulary for Person to Person Relationships

acquanitanceOf

- professionallyRelatedTo
    - worksWith
    - collaboratesWith
    - colleagueOf
- familyOf
    - immediateFamilyOf

        - spouseOf
        - lifePartnerOf
- socialAcquaintanceOf
    - friendOf
    - romanticWith
- proximityAcquaintanceOf
    - livesWith
    - neighborOf
- antagonistOf

Controlled Vocab for Relation of Person to Work

We tried to pull from Dublin Core whenever possible, though the language is altered to reflect differences in our data arrangement (ie, instead of DC's "creator," we have "creatorOf" to model relationship controlled vocab). The limited number of DC elements means that others (all except "creatorOf," "contributorTo," and "publisherOf") are created for The Crowded Page.

Term	Description
creatorOf	This person is primarily responsible for making the work.
editorOf	This person had a formal editorial role with the work, including acquiring it for publication, soliciting its creation, and/or working with the creator to revise the work for publication, and/or is responsible for making the work available to the public.
curatorOf	This person is responsible for making the work part of a public exhibit.
contributorTo	This person made contributions to the work in a way not otherwise identified in the list of controlled vocabulary.
illustratorOf	This person provided original visual art to be integrated as part a textual work.
translatorOf	This person translated a work from one language into another.
adapterOf	This person adapted a work of one medium into a work of another medium.

Controlled Vocab for Relation of Person to Place

These are all created for The Crowded Page.

Term	Description

residence	This place is the residence of the person it is related to.
workplace	This place is the workplace of the person it is related to.
socialSpace	This place is a regular social gathering spot of the person it is related to.

Controlled Vocab for Relation of Person to Group

We've decided that we really need no list of controlled vocab for relationships between people and groups. Instead, we will just say someone is a "memberOf" a group (or not).

Term	Description
memberOf	This person belongs to an established group.

Appendix Two: Sample RDF/Provenance Process

In this document, we describe how the source data (book, page numbers, etc.) can be associated with the relationships documented by the Crowded Page project. We call such data the "provenance data" of the relationship. There are three major issues:

- How to represent the source data for a relationship in RDF. This essentially defines the syntax for exchanging information among systems.
- How to load and store this information in our knowledge base (recall, a knowledge base is similar to a database, but has a capability to draw additional conclusions from the data).
- How to retrieve provenance data from the knowledge base.

1. RDF Representation of Provenance Data

The RDF data model consists of triples, each with a subject, predicate and object. This is sufficient for recording a specific relationship (e.g., <Thomas Bailey Aldrich, friend of, Launt Thompson>), but makes it difficult to attach source information to this relationship. However, RDF provides a reification method, which allows us to identify a particular triple and then provide additional information about it. We use this method, which although verbose, is semantically correct. To illustrate, we start with an example of the RDF currently produced by Rob Weidman for the Vault at Pfaff's data.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:vp="http://digital.lib.lehigh.edu/pfaffs/schema#">
<vp:relationship rdf:about="http://digital.lib.lehigh.edu/pfaffs/r1">
<vp:title>Thomas Bailey Aldrich friend of Launt  Thompson</vp:title>
<vp:rID>1</vp:rID>
<vp:p1ID>1</vp:p1ID>
<vp:p2ID>30</vp:p2ID>
<vp:type>friendOf</vp:type>
<vp:sources>
<vp:source rdf:about="http://digital.lib.lehigh.edu/pfaffs/w1052">
<vp:sourceNote>
Mrs. Aldrich refers to Launt Thompson as "Mr. Aldrich's chum" (22). She also mentions
that Thompson gave a plaster medallion of Aldrich to him as a wedding gift (58).
</vp:sourceNote>
<vp:sourcePages>22,25,56-57</vp:sourcePages>
</vp:source>
<vp:source rdf:about="http://digital.lib.lehigh.edu/pfaffs/w1464">
<vp:sourceNote>
```

A collected edition of Aldrich's poetry to be published by Ticknor & Fields was "embellished with an exquisite steel engraving of the poet after the medallion by his friend Launt Thompson" (63).
</vp:sourceNote>
<vp:sourcePages>38,63,73</vp:sourcePages>
</vp:source>
</vp:sources>
</vp:relationship>

The equivalent data in the proposed format is given below.

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cp="http://swat.cse.lehigh.edu/resources/onto/crowdedpagebin.owl#">
<rdf:Statement>
<rdf:subject rdf:resource="http://digital.lib.lehigh.edu/pfaffs/rdf/p1"/>
<rdf:predicate
rdf:resource="http://swat.cse.lehigh.edu/resources/onto/crowdedpagebin.owl#friendOf"/>
<rdf:object rdf:resource="http://digital.lib.lehigh.edu/pfaffs/rdf/p30"/>
<cp:provenance>
<rdf:Description>
<cp:source rdf:resource="http://digital.lib.lehigh.edu/pfaffs/rdf/w1052"/>
<cp:sourceNote rdf:datatype="&xsd;string">
Mrs. Aldrich refers to Launt Thompson as "Mr. Aldrich's chum" (22). She also mentions that Thompson gave a plaster medallion of Aldrich to him as a wedding gift (58).
</cp:sourceNote>
<cp:sourcePages rdf:datatype="&xsd;string"> 22,25,56-57</cp:sourcePages>
</rdf:Description>
</cp:provenance>
<cp:provenance>
<rdf:Description>
<cp:source rdf:resource="http://digital.lib.lehigh.edu/pfaffs/rdf/w1464"/>
<cp:sourceNote rdf:datatype="&xsd;string">
A collected edition of Aldrich's poetry to be published by Ticknor & Fields was "embellished with an exquisite steel engraving of the poet after the medallion by his friend Launt Thompson" (63).
</cp:sourceNote>
<cp:sourcePages rdf:datatype="&xsd;string">38,63,73</cp:sourcePages>
</rdf:Description>
</cp:provenance>
</rdf:Statement>

</rdf:RDF>

2. Load reified RDF data into knowledge base

For every reified statement, the process of loading the transformed RDF reification format into the underlying knowledge base is as following.

A. determine the subject (e.g., p1), predicate (e.g., friendOf) and object (e.g, p30) of this statement;
B. get internal IDs for these URIs from the dictionary table;
C. use these URIs' internal ID to get an internal statement ID for this statement in the knowledge base by querying the triples table; (Alternatively, step 2 and 3 can be combined in one sql query, which may improve the performance)
D. get internal ID for the work (the resource in the cp:source element) of this provenance from dictionary table;
E. convert all associated source information contained in the cp:provenance element into a tuple with internal ID of the statement and id of work together as the key. The associated source information includes strings of source note and source pages.
F. store the tuple of source information in the provenance table. The DB schema design gives detailed table relation.

3. Retrieving provenance information

During the query process, first we only retrieve normal social chain information, including name, RDF id, web URL etc for each entity (e.g., person, work, or organization) and relation (e.g., friendOf, authorOf, memberOf, etc.) between them. When necessary, for example Link.getMetadata() method is called, we will retrieve corresponding provenance information based on the triple information, i.e. the subject as rdfId of predecessor link, predicate as rdfId of this link and object as rdfId of successor link. The process includes the steps similar to the first three in above loading process. Then using the internal statement ID, we retrieve all provenance information related to this statement, including the work, the source note and the source pages. After that, we fill in metadata, a java object, with this information and return it. This is a lazy solution which means we retrieve provenance only on request. However, after the first request, the provenance information is still available to the application as long as it continues execution. Another alternative can be an active one, which means associated provenance data is queried and returned when the chain is constructed.. When the Link.getMetdata() method is called on an entity instead of a predicate, it returns null. This is because it is unclear what the provenance of an entity should be.

Appendix Three: *Crowded Page* Database Model