



---

Theses and Dissertations

---

2010-04-20

## Historical GeoCollaboration: The Implementation of a Scoring System to Account for Uncertainty in Geographic Data Created in a Collaborative Environment

Anthony D. Contreras  
*Brigham Young University - Provo*

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>



Part of the [Geography Commons](#)

---

### BYU ScholarsArchive Citation

Contreras, Anthony D., "Historical GeoCollaboration: The Implementation of a Scoring System to Account for Uncertainty in Geographic Data Created in a Collaborative Environment" (2010). *Theses and Dissertations*. 2108.

<https://scholarsarchive.byu.edu/etd/2108>

This Selected Project is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact [scholarsarchive@byu.edu](mailto:scholarsarchive@byu.edu), [ellen\\_amatangelo@byu.edu](mailto:ellen_amatangelo@byu.edu).

Historical GeoCollaboration: The Implementation of a Scoring System to  
Account for Uncertainty in Geographic Data Created in a  
Collaborative Environment

Anthony David Contreras, Jr.

A Master's Project Report submitted to the faculty of  
Brigham Young University  
in partial fulfillment of the requirements for the degree of  
Master of Science

Brandon S. Plewe, Chair  
Perry J. Hardin  
Mark W. Jackson

Department of Geography  
Brigham Young University  
August 2010

Copyright © 2010 Anthony D. Contreras, Jr.

All Rights Reserved

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a Master's Project Report submitted by

Anthony David Contreras, Jr.

This project report has been read by each member of the following graduate committee  
and by majority vote has been found to be satisfactory.

\_\_\_\_\_  
Date

\_\_\_\_\_  
Dr. Brandon S. Plewe, Chair

\_\_\_\_\_  
Date

\_\_\_\_\_  
Dr. Perry J. Hardin

\_\_\_\_\_  
Date

\_\_\_\_\_  
Dr. Mark W Jackson

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the Master's project report of Anthony David Contreras, Jr. in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

---

Date

---

Dr. Brandon S. Plewe, Chair, Graduate  
Committee

Accepted For the Department

---

Date

---

Dr. J. Matthew Shumway, Chair, Geography  
Dept.

Accepted For the College

---

Date

---

Susan S. Rugh, Associate Dean College of  
Family, Home and Social Sciences

## ABSTRACT

### Historical GeoCollaboration: The Implementation of a Scoring System to Account for Uncertainty in Geographic Data Created in a Collaborative Environment

Anthony David Contreras Jr.

Department of Geography

Master of Science

A Geographic Information System (GIS) is an existing tool to create, manage, analyze and visualize data with a spatial component, and is used by many types of organizations in many fields. For most of the tasks and projects within these fields, a GIS provides highly accurate results. Under certain circumstances (Massive Scope, Widespread Expertise and Multivalency), a GIS fails to provide adequate results in the field of historical geography, for example. Crowdsourcing tools like Wikipedia and Open Street Map (OSM) address some of these issues, but not all, and introduce new problems. This project focuses on geographic data dealing with historical events, places and people. This project uses a wiki software package, extensions for added functionality, and customized tools to implement a scoring system to rate the accuracy of each assertion made by members of the contribution community. The scoring system addresses the ambivalence of the data created by a community.

Keywords: collaboration, volunteered geographic information, historical geography

## ACKNOWLEDGEMENTS

There were many people who contributed to the completion of this project and deserve to be recognized for their efforts. My wife Kristi sacrificed countless hours providing an environment for this work to be accomplished, as well as supporting and motivating me to complete this project. Thank you to my children Kelsey, Kaiden, Isla and Alex for your patience and understanding throughout this journey. I express my appreciation to Dr. Plewe, my graduate committee chair, for his guidance during this experience. I would also like to thank Dr. Hardin and Dr. Jackson for their participation in my graduate committee. Very sincere thanks go to Geography department secretary, Laurie Weisler, who provided invaluable encouragement and assistance with administrative issues every step of the way.

# Table of Contents

<b>LIST OF FIGURES.....</b>	<b>VII</b>
<b>I. INTRODUCTION.....</b>	<b>1</b>
CURRENT GIS USE.....	1
CONCEPTUAL FRAMEWORK AND EXPLANATION OF THE PROBLEM.....	2
PUBLIC PLANNING.....	4
COLLECTING REGIONAL DATA IN THE PRIVATE SECTOR.....	5
HISTORICAL GEOGRAPHY.....	5
POSSIBLE EXISTING SOLUTIONS.....	7
PURPOSE OF WORK.....	7
SIGNIFICANCE OF THE PROJECT.....	7
OVERVIEW.....	8
<b>II. PREVIOUS WORK.....</b>	<b>9</b>
PUBLIC PARTICIPATION GIS.....	9
CROWDSOURCING.....	11
WIKIPEDIA.....	13
CROWDSOURCING GEOGRAPHIC INFORMATION.....	15
EXISTING POTENTIAL SOLUTIONS.....	17
<b>III. PROJECT DESIGN.....</b>	<b>22</b>
STRATEGY.....	22
APPLICATION CONTEXT.....	22
GIS FUNCTIONALITY.....	25
SOFTWARE ARCHITECTURE OVERVIEW.....	25
DATA MODEL.....	27
QUALITY SCORING.....	29
CONTRIBUTOR REPUTATION.....	30
ASSERTION QUALITY SCORE.....	33
DATA DISPLAY.....	35
<b>IV. IMPLEMENTATION OF PROJECT.....</b>	<b>36</b>
STRUCTURE OF THE CONTRIBUTED DATA.....	37
DATA CREATION.....	39
VISUALIZING THE DATA.....	40
VOTING.....	44
VIEWING THE SCORES.....	45
PLACE PAGE LAYOUT.....	47
<b>V. PRESENTATION OF RESULTS.....</b>	<b>48</b>
<b>VI. CONCLUSION.....</b>	<b>50</b>
<b>REFERENCES.....</b>	<b>52</b>

## List of Figures

Figure 1. Number of edits per article. ....	14
Figure 2. Proposed workflow of the web site. ....	26
Figure 3. Project data model. ....	28
Figure 4. Data model implemented for a settlement entity.....	38
Figure 5. Properties of the Name template. ....	39
Figure 6. A view of the location section of the data input form. ....	40
Figure 7. The project's home page with a map showing all contributions. ....	41
Figure 8. A view of the most recent assertion with map showing location. ....	42
Figure 9. The embedded map showing the locations of all assertions for the entry.....	43
Figure 10. Map viewer displaying entries with the Building category.....	44
Figure 11. The flagged Revs toolbar, the user interface for submitting a rating. ....	44
Figure 12. Database tables used for storing and keeping track of ratings. ....	45
Figure 13. One property of a place with score and link to all assertions.....	46
Figure 14. Comparison page to view all assertions. ....	46
Figure 15. The page layout of a settlement place. ....	47

## **I. Introduction**

### **Current GIS use**

A Geographic Information System (GIS) is a tool for dealing with information that has a link to geography or some kind of spatial relationship (Armstrong, 1994). GIS is commonly used in many types of organizations including business, education, non-profit and government (MacEachren *et al.* 2005). Typically, a GIS is used to create, manage, analyze and often visualize spatial data, or data with a geographic component (Wood, 2005). This project focuses on the aspect of data creation rather than analysis or visualization.

The widespread use of GIS in so many organizations with varying goals and interests shows that it is a generally accepted solution for solving problems that deal with the aspect of location. Over time, GIS software has evolved to be able to address problems or questions that it could not previously. One example of this is the ability to analyze 3-D data so that things like elevation and terrain can be considered or accounted for.

However, there are situations where GIS has not been able to adequately address the issue in a way that would result in a solution that is “good enough” to obtain the desired answer to the problem or question. The area of interest in this project is historical geography.

## **Conceptual framework and Explanation of the problem**

Some of the scenarios in which traditional GIS fails include collaborative data collection and data management in public planning (Wood, 2005), the private sector and historical geography (MacEachren, 2000). In the previously mentioned fields, it is sometimes required or just beneficial to allow multiple people work on the same project simultaneously, including the job of creating data. The traditional GIS methods (Siffer, 1995) for collecting geographic data are not satisfactory when the following situations apply simultaneously:

### **1. Massive Scope**

A large project can often require large amounts of information and the work of hundreds of people, when few are available. There may not be enough time or other resources to complete the project with the tools or employees that an agency has at their disposal. This prevents organizations from being able to solve the problems that they are expected to and that makes them seem inefficient.

### **2. Widespread Expertise**

Most people have an expert level understanding of something, even if the breadth of their expertise is very narrow. Allowing experts to contribute to a project or the process of data creation is desirable since their expertise is highly valued. They do not need to be experts in the same field to increase the quality of the results. The real benefit in this situation is that all of these participants are experts in different areas within the scope of the project. If a community of project participants consists of many experts, they can collectively span most of the subject with their knowledge. The opinions of the public can have a

similar effect. Members of the public have knowledge of events, places and people specific to their local area which can also benefit certain projects.

Another aspect of widespread expertise is that of widespread contributors, where there is also a large number of participants making contributions. An increasing number of contributors will also increase the likelihood of someone having valid and useful information (Rashid *et al.* 2006).

### **3. Multivalency**

Properties that can have multiple values are considered *multivalent*. Multivalency occurs when the subject and its data will inherently contain uncertainty and subjectivity that cannot be removed or resolved, resulting in multiple opinions or alternative properties. There are many subjects whose data will always contain uncertainty or subjectivity. One of the best examples is historical data.

It is difficult to know the absolute truth about historical events, figures, and locations, etc. because sources may be questionable, incomplete, or even contradictory (Gregory, 2001, p. 17); thus multiple theories may exist that are considered equally valid. Other subjects may rely on opinions, which can be debated and are unlikely to be generally accepted as fact. Research in these fields commonly results in data that is multivalent.

When a project that would normally use GIS has a massive scope, would benefit from widespread expertise and has multivalency, an alternative solution is needed. These types of projects are common, so producing a solution as previously mentioned would be a great benefit to many fields of study and problem solving. Some of the fields that would

benefit from this solution are public planning, the private sector, and historical geography.

### **Public planning**

In public planning, the opinion of the community should be valuable and highly desirable to decision makers. Traditional GIS used by public agencies do not adequately acquire nor include the public opinion. Opinions are often sent to the planners via an office memo, email, phone call, or some other traditional method, but none of these can be easily incorporated with a GIS. When the public submits opinions there might be a text description of the location, which can be misinterpreted. Citizen-contributed geographic information in a GIS format produces data that is more easily incorporated into the planning process.

The properties of proposed building zones, roads and parks will have multiple values, representing the varying opinions of the community. Each citizen will have their own opinion, and it may or may not be different than that of another. Some members of the community that share their opinion may have more useful information to contribute about a specific place because they live or work in that area. Reaching out to the citizens for information used in the planning process can result in obtaining data from people who are “experts” in their community and should result in effective planning.

### **Collecting regional data in the private sector**

The three problems previously listed commonly appear in the private sector. An example of this is a company that wants to know the demographics of the possible customers in the service area of each branch office around the country. The central office could not acquire this data within a reasonable amount of time, and there are employees who live in every area that the customer intends to learn about. These employees have direct access to the population of interest, and have general knowledge of the area (Cinderby, 1999). Most companies have small GIS organizations, so the number of employees that have a medium to high level of experience with GIS is small. Any type of survey or other data acquisition technique for demographic data will have uncertainty (Seeger, 2008) since people who take surveys have the opportunity to give false information or omit important details in their responses. The ideal solution would allow all employees to contribute to a master database that can be used by everyone in the company.

### **Historical geography**

This project focuses on the field of historical geographic information, which more often than the previously mentioned fields consists of the criteria that make GIS a less effective tool for data collection. The scope of historical data, by nature, is too large for an individual to address. There have been so many events, civilizations, people, places and so on that a single person or agency cannot research and describe them all. There is a wide range of options when it comes to subjects within history, so it is common that experts of history only have a high level of knowledge in one or two specific areas, and possibly geographic regions too. Many college professors teach courses dealing with a

specific time period or place, which usually reflects their area of expertise. There are also many amateur historians, usually having a very specific focus, typically in a small region. These resources are often underutilized when collecting historical information when they have useful and valid information to contribute (Showers and Malahleha, 1992; Olsson and Folke, 2001; Robertson and McGee, 2003). While their knowledge of historical events and people is great, historians generally do not have much experience with GIS and its methods and should not be required to perform advanced analysis on geo-historical information (Gregory, 2001).

Studies in various fields have presented problems with making historical data useful (Mladenoff, 2002; Shelton, 2007). Many aspects of historical information are open for debate, since documentation can be incomplete, or multiple sources can present different and sometimes conflicting information (Gregory, 2001; Shelton, 2007). Plewe describes the possible causes of uncertainty that arise when using historical records to create data (2002, pp. 441-442). In this project, places had alternative names, dates, location, and even different opinions about whether or not two places were unique or alternative names for the same place. As a result, there are often differing opinions about the properties of the data. Because of these circumstances, geo-historical information is a prime candidate for a method other than GIS to provide a more meaningful outcome.

## **Possible existing solutions**

Before creating a solution to the problem, existing solutions should be evaluated.

*Crowdsourcing* is a method that gathers large numbers of users in a community to contribute information and edit information in a common database, typically using the internet. When a project has a massive scope, this solution can get many people to take part in the project. Even if each person only has expertise or opinions on a small part of the project, their collected knowledge can be substantial. For example, Wikipedia (English) has roughly 300,000 contributors and 3,221,964 articles.

## **Purpose of work**

The purpose of this work is to build a database of historical places, using existing crowdsourcing technologies, augmented by new tools to handle multivalency and quality control. This should result in better information through a larger knowledge base as well as a form of quality control.

## **Significance of the project**

This project will benefit the research community in the field of Historical Geography and Historical GIS, as this implementation seeks to provide a better method of creating such data and provide a usable way to account for the uncertainty of the data contributed by a community of users. This project can also be extended to other fields that deal with geographic data with inherent multivalency such as planning. Since the project leverages the efforts of some of the existing solutions that will be described later, it can be easily

reproduced and extended in the future. This work does not provide a complete solution for this issue, however it advances this field of research and can be extended in the future with improvements.

## **Overview**

The remaining sections will be presented as follows. The Previous Work chapter highlights the publications to date and existing projects with the relevant information they each provide to this study. Then I will discuss the deficiencies in these earlier efforts that I intend to solve in this project. The next chapter describes the desired strategy that can be used to provide a solution to the problem. The implementation of the project is explained with specific examples. A presentation of the results and suggestions for improvements follows. Finally, the conclusion addresses the changes made based on the findings of the preliminary testing, conclusion of the project itself, and recommendations for those who extend this work in the future.

## II. Previous Work

There have been many attempts to address at least one of the problems described in my conceptual framework. Discussion of these solutions and studies in detail provides an understanding of both their contributions to the improvement of the data creation process, and the problems associated with each method.

### Public Participation GIS

Public Participation GIS (PPGIS) is a field of study that provides a way for many users to contribute to the GIS process (Jankowski *et al.* 1997). It provides the many answers to complex problems that have many possible solutions (Cinderby, 1999). PPGIS is also referred to as “GeoCollaboration” (MacEachren, *et al.* 2005). This specific area is one of great interest to the geographic community.

Much of PPGIS research deals with the participation of a community in GIS analysis, not data collection, but there have been some studies into data contribution and to a certain extent, quality control. Elwood and Ghose (2001) point out that in some cases, the data and processes that result from PPGIS projects can be reused for similar or sometimes unrelated projects, making them more valuable to the organization that has implemented the PPGIS. Since the purpose of this work is to create general-use data, to be used and reused multiple times, it would benefit from research. Many of the problems that analysts use GIS to help solve are usually solved by groups; this creates a growing disparity between what the systems are built for and how they are actually used (Armstrong, 1994).

Paez *et al.* (2006) identifies three possible reasons that the results from public participation may not be considered credible by the decision makers. While the 2006 study deals with cost benefits analysis, two of the characteristics appear with geo-historical data collection through PPGIS as well. When collecting information from the public, it will be subjective since contributors will either use biased models or a lack of scientific background (Paez *et al.* 2006). The other reason is that the results infrequently contain spatial information (Paez *et al.* 2006). Adding a geographic input component to the application will provide spatial information so that it can be used by geo-historians (Gregory *et al.* 2002, p. 11).

Early PPGIS research began with projects in which contributors physically met together (Jankowski and Nyerges, 2001), however Wong and Yang (2004) extend this format to the web, which could attract more participants who have the option of participating at home on their own schedule (MacEachren, 2001). Carver *et al.* (2001) also mentions the benefit of an online forum for collaboration as not limited to location and time.

Contributions can be made by participants in any location with internet access, at any time of the day. Exposing a project to the World Wide Web (WWW) increases the number of users that can participate and can utilize the knowledge of many contributors. Included with the added benefit of the larger community contributing to the project is the problem of multivalency. Al-Komandy (2000a, 2000b, 2001) has shown that PPGIS has not been successful in solving many problems because this problem is present but not addressed.

Public Participation GIS attempts to address issues that are best solved by a group of people, though most of the studies have had their shortcomings. The most common of these is the problem of multivalency and quality control. Collecting data through crowdsourcing does not include the analysis of data and will also result in multivalent properties. PPGIS studies generally have not utilized large numbers of contributors making crowdsourcing more desirable.

### **Crowdsourcing**

There are over 1.73 billion internet users in the world, roughly 25% of the global population. (<http://www.internetworldstats.com/>) With the increase in the number of internet users (Carver *et al.* 2001), there has been a movement (Baum *et al.* 2003) towards community knowledge websites. A new term called “crowdsourcing” has gained popularity and attention from various fields and industries (Howe, 2009). This term defines a method of taking a task or job that is usually assigned to one person and allowing a community of participants to collectively complete the task. Derived from the term outsourcing, this method, also called “open source collaboration”, has many advantages for different types of projects. Some of the benefits of this method are faster completion of the project, lower costs, higher quality output, and a more robust source of information.

Because there can be more than one user working on the same job or task, more is being accomplished during that time. Since more work is being done within the same time period, the task will be completed sooner than if one person was doing it. Costs are

usually reduced since the participants are usually working for free, because in many cases, they benefit from the result or end product (Alsever, 2007). Those who take part in this type of process often have a passion for the subject so they will contribute to the project without monetary incentives being offered (Windsor, 2009). For the same reason, those in the community often have a feeling of ownership, either of their own contributions or the project as a whole, which is one possible incentive to encourage high quality results (Howe, 2008, 14-15). Finally, with the higher number of participants, comes a larger body of knowledge and experiences. In many cases, the collective knowledge of ten people is significantly greater than that of one individual. Leveraging this type of system would help obtain widespread knowledge and contributions that could produce useful historical data.

Another benefit to this solution is that participants do not need a strong knowledge of GIS technology or methodology. Participants of crowdsourced projects are usually able to submit basic information, which an expert can use for analysis or decision making. Since crowdsourcing applications often have simple tools to input information, this can increase the number of potential users that can efficiently use this type of application. While crowdsourcing addresses two of the three issues described earlier, it does not address the issue of multivalency (only the most recent edits are shown), and introduces a problem with quality control, since anyone can change info without verifying it. For the purposes of historical data, this is particularly problematic, since there is no way to classify the validity of data reported.

## **Wikipedia**

The most common of crowdsourcing use “Wiki” – based sites that allow any user (anonymous or registered, depending on administrator’s settings) to contribute their knowledge about a subject that the site is concerned with. This input is usually in text form, not GIS, like most websites. The most popular and well-known is Wikipedia.org. The Wikipedia project is an online reference system to provide internet users with a resource for obtaining information about most things they have questions about, something that has massive scope. The Wikipedia project is founded upon an assumption that each potential internet user has his or her own unique skill set and knowledge base (Cinderby, 1999). It is also available to almost everyone with internet access, so that widespread expertise can be utilized.

The accuracy of crowdsourced data will depend on the knowledge and experience of the contributor. In the case of Wikipedia, this is only monitored by administrators and users, from which not all subjects can be covered at an expert level. Because of this, the accuracy of information submitted by users cannot be guaranteed (Stvilia *et al.* 2005).

The system relies on users regularly checking all the information being submitted by other users and either removing it, or replacing it with the correct information (Stvilia *et al.* 2005). Theoretically, the greater the number of users of the site, the more likely a user with a correct understanding of the subject will happen upon a page and ensure the validity of its contents. This is the case with Wikipedia, whose pages are regularly viewed and modified constantly. It is very common to post something (especially if it is

most definitely not true) to an article and see that it has been removed later in the day.

Figure 1 shows that over time, articles are being edited more often.

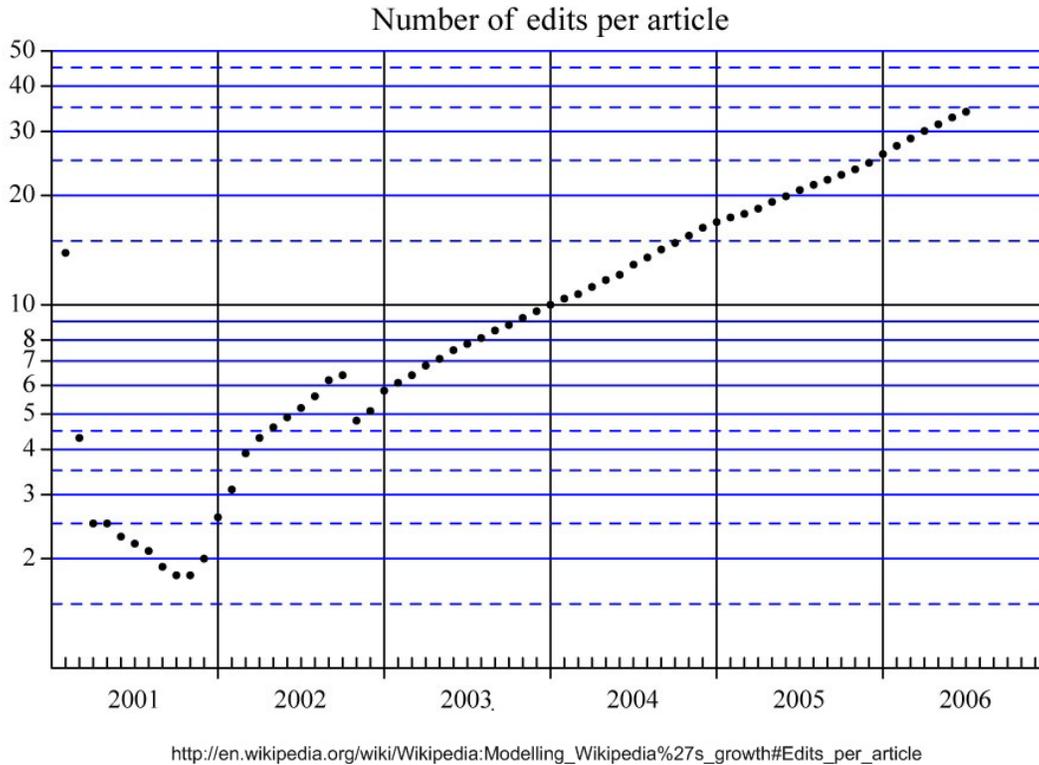


Figure 1. Number of edits per article.

Articles continue to be published by field experts (Shiffer, 1995) and updated information is made available. Each update has the potential problem of erroneous assertions where people make bad edits because they think they are right, even when the previous information is correct. Additionally, edits made by users can be opinions rather than verifiable assertions. Thus the peer review method that Wikipedia includes by itself is not always an effective method of quality control (Hu *et al.* 2007; Stvilia *et al.* 2008). One possible solution to this issue is to make all edits available for contributors to provide feedback, which is part of the feedback system this project implements.

It is now possible for anyone with access to a web server to pick one of the many wiki software packages available and install it. Many of these packages are based on open-source technology, and the administrators are free to make changes to the software according to the purpose of the website.

### **Crowdsourcing geographic information**

At this time there is a strong awareness of geospatial technology (Wood, 2005), (MacEachren, 2000). Two of Google's technologies (Google Earth and Google Maps) have helped increase this awareness, by allowing users to easily create information and share it (Seeger, 2008).

A site called OpenStreetMap (OSM) provides online mapping functionality similar to Google Maps, but it also includes crowdsourcing tools for geographic data creation. A project which began in 2004 (Haklay and Weber, 2008, p 13), OpenStreetMap is a web site that allows registered users to upload street data to their data repository that is served over the internet to web and desktop clients. The OSM interface and concept borrows from the Wikipedia site, having a similar looking "Edit" tab and keeps track of versions of uploaded data. Once data has been uploaded, the contributor or other participants have the ability to edit the data to add more information or clarify existing information, making it more useful. While this is a powerful tool for creating data within a collaborative framework, it lacks a method for evaluating the quality or reliability of the data. This project is often cited as a prime example of "crowdsourced" GIS.

Another term that is often synonymous with crowdsourcing geographic information is Volunteered Geographic Information (VGI). In a study about using VGI for landscape planning, Seeger (2008) concludes that the information from the volunteers with local knowledge helped decision makers to make more informed choices before taking action. However, it is difficult to interpret the credibility of VGI contributions, and Flanigan (2008) suggests further research and testing on this subject.

The map-based wiki site Wikimapia (<http://wikimapia.org/>) adds a geographic component to what Wikipedia already has. This site allows contributors to map the location of Wikipedia articles, and all the features have links to the related Wikipedia article and other web sites. This site still leverages some crowdsourcing techniques by allowing users to create geographic features and assigning some descriptive properties, however there is no tool or functionality that solves the problem of the ambivalence and data quality of the contributions.

The Histoatlas project (<http://histoatlas.org>) seeks to be the accurate geo-historical encyclopedia. This site is relevant as it also has the intent to collect geo-historical information. It utilizes crowdsourcing to address its massive scope and intends to have contributors that specialize in all aspects of historical geography. This site does not expose alternative assertions or opinions so it is difficult to tell what others have proposed in attempts to describe the entity. Because there is no ability to track assertions

submitted, Histoatlas also lacks a method to handle the multivalency that results from managing so many assertions.

As previously shown, there are many crowdsourcing tools available. After evaluating these, each has shown the ability to address the need for widespread expertise. These tools are deficient, however in effectively handling multivalency and quality control.

### **Existing Potential Solutions**

#### **The New FamilySearch.org**

There are many billions of people who have lived on this earth and it would be impossible for a single person or group to document each person. Genealogy is one of the largest applications on the internet. In this field, there is a large set of contributors with focused expertise, each researching their own information. The LDS church has created a new version of its FamilySearch.org website, which is intended to assist its users in finding information about ancestors, including birth date and place, marriage, baptism and death date and place. The new version has a massive common database combining all of the genealogical data the church has collected, but also leverages crowdsourcing for data collection. The user can make modifications to any records they have submitted, or contribute an opinion for records they did not submit. Being a family history site, it fits into the geo-historical field, since there is a geographic aspect to the historical data, because it features places that can be shown on a web map. This project benefits from having contributors from all regions of the world, as they each have access to detailed information about people from their area, and many have expert knowledge about their own families.

Genealogical information is inherently multivalent, for the same reasons as other historical information. When data comes from sources that are incomplete, unreliable, or conflicting, multiple theories of any fact are possible. To handle this, the NewFamilySearch project uses a system of adding alternative assertions. Users can add their own assertions of a fact or mark information that others have submitted and they disagree with as “disputed” and provide sources and justification for their assertions. These disputations and assertions are visible to all users.

It is a sign of good progress that the multivalency problem is recognized and that an effort has been made to account for this. Unfortunately, it is not enough to be able to mark an assertion as disputed, since only the author of the data and the editor are the only ones whose feedback is considered, and all assertions are shown, even ludicrous ones and typos. The entire community should be involved in this process and people should be able to interpret the results so that they are useful.

### **Pleides**

The University of North Carolina’s Pleides project (<http://pleiades.stoa.org/>) attempts the same issue as this project, geo-historical data, but for the classical- era Mediterranean (Elliot, 2006; Talbert, 2006). Anyone can contribute information about historical places, including ancient Roman and Greek civilizations. To handle the quality control issues, after submitting a contribution it must be reviewed by an editor before it is published and available to all users. This attempt at quality control may work in some circumstances, although it requires editors with a lot of time available to review every submission. This

system of quality control is not scalable to the size of Wikipedia because of the high demand this would put on the limited number of editors available.

In the case of Pleiades, the quality control starts when the contributor assigns a level of confidence for the contribution. Along with the scaling limitation, using a centralized editorial staff keeps data quality in the hands of recognized experts, but even they may not know as much as an amateur historian. This system also prevents the community from submitting feedback, which can be a better way to address the problem that is the result of multiple assertions about the same thing.

Changes are logged in a history in an attempt to account for multiple values for the same place and the owner can view these revised versions, but this is a poor way to handle multivalency, because people only see the most recent, not the best, alternative, and they only see one at a time.

### **Citizendum**

Citizendum is a wiki-based encyclopedia project that has its own method of quality control that attempts to remove the limitation found in Pleiades. The site accommodates for this by giving users the opportunity to become editors based on real world credentials (collegiate degrees, work experience, publications, etc.). In the last two years of its existence, Citizendum has produced only 121 final articles, and more than 13,000 still in the editing process. The benefit of this system can only be seen when the number of editors is large enough to review and revise the large quantity of new and existing articles.

In summary, there have been many solutions that try to address the problems of massive scope, widespread expertise, and multivalency. These websites often solve the first two problems and introduce the third without providing a solution. This project seeks to provide a solution for all of these issues that is user friendly and provides easy to use GIS functionality at a very basic level. The next section describes the design of the project in a way that addresses the aforementioned three problems.

### **Uncertainty and multivalency in data**

Uncertainty is a part of almost every aspect of geographic data (Cinderby, 1999); from the accuracy and precision of the location to the description information contained within the metadata (Liang *et al.* 2005). While many research endeavors have focused on the best way to handle uncertainty in the location of entities with a spatial component, little effort has been directed towards studying the uncertainty of other attributes, many of which are used for analysis in decision making (Wang *et al.* 2005). This is important to recognize and handle when there are many resources at risk and the use of those resources are based upon this information, however correct or incorrect, certain or uncertain.

A form of uncertainty that often appears in historical data is when historical records are the primary base of the model (Plewe, 2002). Historical records are frequently incomplete, (Ott and Swiaczny, 2001, 113-114), Historical quotes or records are sometimes referenced incorrectly to support an argument or asserted to have different meanings by different people (Powell, 1987; Flaherty, 1995). Although historical texts

answer many questions, it can often create more questions and result in multiple views of the past (Boonstra *et al.* 2004). When multiple alternatives and assertions are available, there must be some way to account for them in a way that allows people to make decisions about how to use these alternatives.

### **III. Project Design**

#### **Strategy**

The overall strategy of this project must address all of the three problems described in the conceptual framework section of the Introduction chapter and address the data quality issues that crowdsourcing introduces. We leveraged existing crowdsourcing technology with two added features. It includes a feedback system that addresses the problem of data quality by allowing all of the community members perform the duties of an editor; in essence, crowdsourcing quality control. The application will also handle multivalency by storing and allowing access to the alternative properties that will be entered by the users. The strategy is to create a solution that can be easily applied to other situations that all of the same issues.

#### **Application Context**

Our study focuses on the historical geography of the Church of Jesus Christ of Latter-day Saints (LDS), the location and time of significant events, which people were there, what they did, and other related topics. The events and actions that took place in the past have a profound effect on the current state of the LDS church, so most members are keenly aware of and interested in Church History. This project, called “Mormon Places,” covers historical events, places, and people relevant to Mormonism from the early 19<sup>th</sup> century to the present time. The explanation that follows demonstrates how this topic meets the criteria outlined in the conceptual framework.

**1. The scope of the project is too large for an individual (person or organization) to complete.**

Since the number of relevant places and events span a long time and wide space, it is not feasible for a graduate student and his Professor to accomplish the task of researching and contributing enough information to create a robust database. Historians commonly spend years researching a small segment of space and time and still have unanswered questions. Thus the scope is massive and would benefit from an increased number of contributors. The project should be available on the internet as a web site, where the application can be made available to any number of users. These users, preferably in large numbers, can provide the work needed for the project that an individual or research team could not accomplish on their own.

In the early years of LDS history, many of the members moved in collective groups from one region to another, many times to avoid persecution by local residents. For example, LDS members were driven out of Independence, Missouri in 1833 and temporarily settled in Clay County, Missouri. During this same time, there were still many members living in Kirtland, Ohio, two states away. There are many periods of time like this where significant events, places, organizations and people existed in multiple different locations at the same time. This creates a wide scope for both the area and time of the subject being researched. Some of the types of places that will be included in this project are Church buildings (meeting houses, temples, houses, etc.), ward organizations, and settlements of LDS members.

**2. There is a significant benefit to having input from many experts on different subjects and/or different geographic areas.**

Since the project has a massive scope, it also requires expertise in a large number of varying aspects of LDS church history and specific regions. The author and his advisor do not have enough knowledge of this subject to create data that adequately covers the spectrum of LDS church history, because professional and amateur historians each have great knowledge of the past and can create a large information pool. Members of the LDS church study its history in Sunday school meetings, and many become amateur historians, often narrowly focusing on the history of the Church in their area, or on the history and regions relevant to their own ancestors. This creates a very large community of people who can contribute to knowledge about LDS Historical Geography. Because these historians live in many different countries, people in all parts of the world should be able to access the site and provide their experiences and local knowledge of events, places, people, etc.

**3. The subject and its data will inherently contain uncertainty and subjectivity that cannot be removed or resolved.**

Since this is a study of historical places, much or most of the contributions will rely on historical sources such as land records, personal journals, news documents, etc. These types of sources have the potential to be questionable, incomplete, or contradictory, which has frequently led to hotly debated alternative scenarios of past places and events.

To address this problem, the solution must have a way to handle the multivalent nature of the data. The ideal method to address this is a mechanism that records and always keeps

every assertion that is made, similar to the new FamilySearch.org application. Every assertion should be available to all users for their evaluation. While every assertion may be relevant, some will be better than others. If members of the community have a way to provide feedback on the assertions, those viewing the data can use the results of this feedback to help evaluate the level of acceptance by the community as an indirect indicator of their truth for each assertion recorded and displayed in the application.

### **GIS functionality**

Although there is geographic information, the GIS functionality and tools needed are minimal, since they will be used by historians who have little or know GIS experience (Gregory *et al.* 2001). Tools and functions should be easy to use and a participant should be able to enter and edit geographic information with little or no training. GIS has a vast array of geometric functions and tools; however the data will not be used for spatial analysis yet, so that won't be necessary for the purposes of this project. The geometry of the data will be represented as points, which are easy to implement in non- GIS systems. The server software used will have to include a way to create, store and display the location of points, attributes, and dates when these properties change (including the possibility of alternative entries).

### **Software architecture overview**

To address the issues of massive scope, widespread expertise and multivalency, we designed the following architecture:

This project will use an existing crowdsourcing software package (wiki) since it already provides a solution to the issues of massive scope and widespread expertise, discussed in the introduction. By using this type of existing solution, contributors can live in varying geographic regions and will potentially have expertise and knowledge in varying subjects. Because the standard wiki software is designed to manage text pages, we need to add the following functionality:

- Location of each entity represented as points
- Store entities as structured elements with multiple properties
- Store each assertion separately so that all versions are available to represent multivalency
- Raw feedback
- Use feedback to evaluate the quality of assertions
- User interface that masks complexity of data

The proposed workflow for data entry is shown in Figure 2.

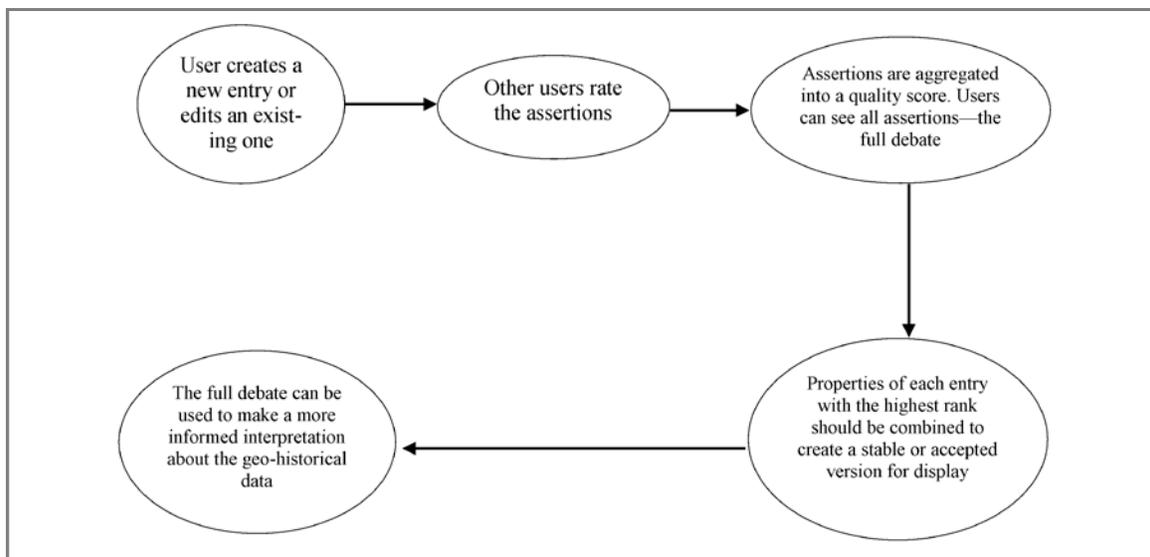


Figure 2. Proposed workflow of the web site.

The proposed workflow was designed by the authors of this project based on the flow of similar crowdsourced projects, with modifications to accommodate the data structure and desired output of the process. The benefit of this workflow is that the output is dynamic and will change to reflect assertions that are added or modified and feedback that is provided.

### **Data model**

The data model is built around a central concept of a historical place, which could be any of a number of kinds of entities, such as a settlement, a building, a congregation, or an event. The place can be made up of zero to many location properties, and zero to many descriptive properties that are applicable to the place, such as name, population, existence, etc., shown in Figure 3. Each property will also have time information associated with it to keep track of changes over time, which is crucial in historical data. Each assertion of a property (e.g. “name”), includes not only the asserted value, but also the name of its contributor, the start and end dates between which it was valid, and some text justification and sources. Another property for the location that specifies the coordinates of a point that represents the location of the settlement, along with its contributor, start and end dates and justification are also included. Other properties could be added to further describe that entity. Any feedback on these assertions are stored as another entity and contains a reference to the assertion it describes and the rating associated with that assertion. Entities representing contributors store properties about his or her account and contributions.

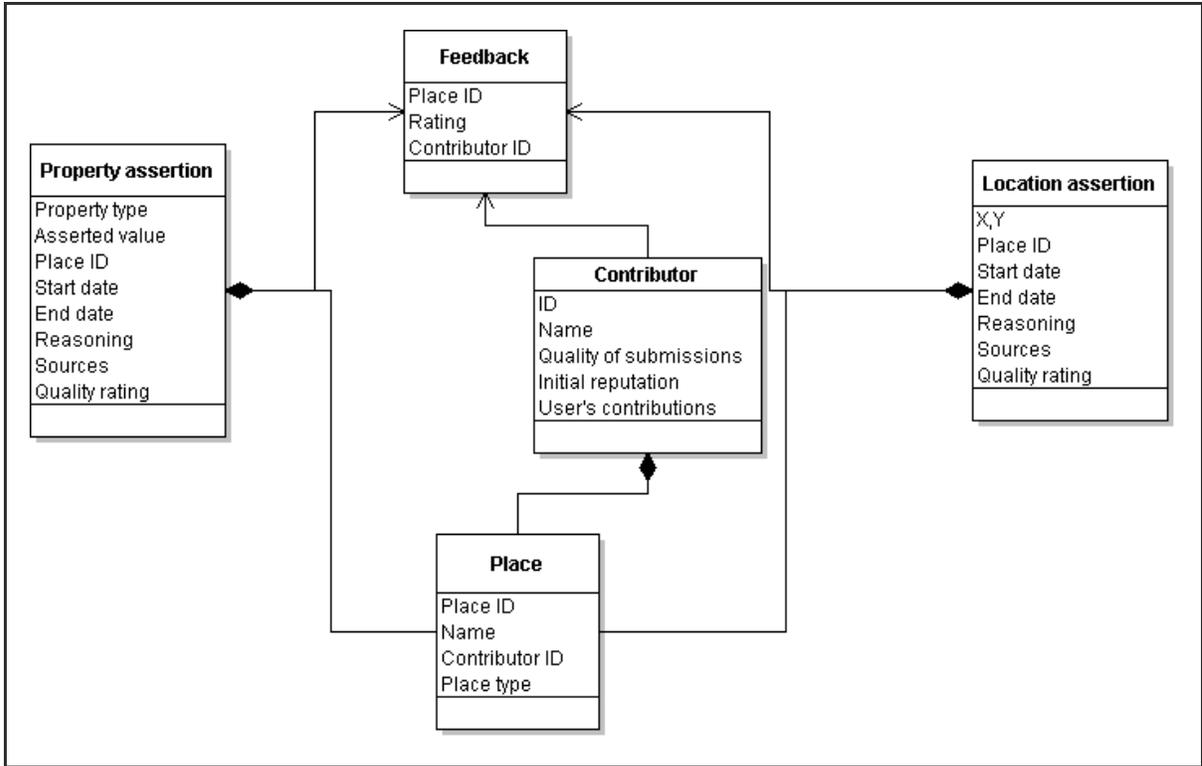


Figure 3. Project data model.

Here is an example displaying this model in use. This demonstrates the kind of historical and spatial information that the database will store.

**Place: Plum Hollow** by jdoe [*The entity record has a common name to make it easier to refer to*]

**Property: Existence="certain"** (3/5/1848-12/9/1852) by jdoe quality 9/10 [*the asserted lifespan of the place*]

**Location: -102.389, 34.7389** (3/5/1848-5/6/1848) by jdoe quality 8.5/10

**Location: -102.883, 35.1002** (5/6/1848-12/9/1852) by cooper quality 8.5/10 [*Cooper believes the settlement moved*]

**Location: -105.632,32.5548** (dates unknown) by jones quality 6/10 [*someone else's opinion*]

**Property: Name="Plum Hollow"** (3/5/1848-1/1/1850) by cooper quality 7.6/10

**Property: Name="Apple Ridge"** (3/19/1849-12/9/1852) by cooper quality 8.5/10 [*note that Cooper believes the two names were both in use for a while*]

## Quality Scoring

This project will allow contributors to act as editors by making edits to assertions and providing feedback by submitting a rating for each assertion. Since the community consists of other historians, most members will be qualified to evaluate the accuracy of the information others submit and those evaluations can be used to assign a score to the assertions. The score generated by the system is interpreted as a rating of quality with a scale of zero to four and considers relevant factors. Two types of scores are generated by the system, one for the quality of a contributor, and the other for the quality of the assertion. In theory, the value of each is based off of the other so that the quality of a user's contributions will help determine the reputation of that contributor, and the reputation of a contributor is used to calculate the quality of their assertions.

The scoring for contributors that results from the data creation uses their participation record, properties that show how they have contributed and how many times. This assumes that members of the community that create more new entries and edit existing ones more often have a stronger reputation and their assertions have more weight (Kazai and Milic-Frayling, 2009; Alonso *et al.* 2008; Wilkonson and Huberman, 2008).

Contribution scores factor the feedback of participants and the reputation of the user who created that version of the entry. The contributor score of the assertion's owner must be calculated before its score can be determined.

### **Contributor Reputation**

The reputation of a contributor is calculated separately as shown below:

$$R(p) = N(u) + \sqrt{N(e) - N(u)} * 2$$

Where:

p = person/contributor

R(p) = Reputation of a contributor p

N(u) = Number of new (unique) places created

N(e) = Number of total edits (including new places)

This calculation accounts for the number of new entries submitted by the user (one point each) and also considers the number of existing entries the user has edited. In this way, users are rewarded more for creating new entries and still receive recognition for adding content or value in some way to the contributions of others.

For example, a user who has created 55 unique entries and made a total of 125 edits the calculation

$$72 = 55 + \sqrt{125 - 55} * 2$$

shows that user's score to be 72, demonstrating that the square root and multiplying by two are calibrations to decrease the weight of edits to existing items. The range of this score begins at zero, and continues, potentially to infinity. The score increases with the number of edits and new entries that the user makes, which has no limit, other than time and memory to store the entries. In order to limit the range and have it on the same scale as assertion ratings, the contributor reputation should be normalized.

The score representing the normalized reputation of a user is calculated as follows:

$$N(R(p)) = \frac{4 * (R(p))}{|R_{max}|}$$

Where:

p = person/contributor

R(p) = Reputation of a contributor p

N(R(p)) = Normalized reputation of a contributor p

R<sub>max</sub> = Maximum reputation of existing contributors (this increases with user scores)

The calculation divides the contributor reputation by the highest reputation attained by a user at that point to rank that contributor on a scale of zero to one. That ratio is multiplied by four which is the highest value of the scoring scale, which allows this score to be comparable to the others. The Normalized contributor reputation for previously mentioned mock user, given the user with the highest R(p) is 84 in this hypothetical scenario, is calculated as:

$$3.428 = \frac{4 * 72}{84}$$

The range for this value is zero to four. The normalized score increases as the raw reputation score gets closer to the current maximum score obtained. It would be ideal for this score to be weighted by the quality of contributions rather than just the number of edits, however this was too difficult to implement in version 1.0. Since the quality of each assertion is calculated using the contributor reputation, there is a recursive situation where each score depends on the other and this calculation removes that limitation.

## Assertion Quality Score

The score representing the quality of an assertion can be calculated using the formula below:

$$Q(c) = \frac{k R(p(c)) + \sum_{f \in F_c} r(f) Q(f)^n s(f)^m}{k + \sum_{f \in F_c} Q(f) s(f)}$$

Where:

- k = Constant value, influence of contributors reputation
- R(p(c)) = Reputation of contributor, this is the user score determined by the Contributor score calculation.
- C = Contribution or assertion entered by a user
- F<sub>c</sub> = All feedback for content c
- f = Individual feedback for content c
- r(f) = Numeric rating in feedback f ( scale 1-4)
- Q(f) = Quality of feedback f (feedback can also be rated, since it is a contribution of its own.)
- s(f) = Substantiveness of feedback f (eg. ratings with justification and sources counts for more than ratings alone)
- m = Constant to weight substantiveness
- n = Constant to weight quality

This formula is a weighted average, at its most basic level, with the assumption that variables R, Q and r are using the same scale (0 to 4). It places the weight on the rating submitted by members of the community, taking into account the number of these

ratings. The reputation of the user who created the entry is also considered in this calculation, with a constant which provides the ability to calibrate the importance of reputation to the quality of a contribution. Because of computational issues resulting from the recursive nature of this calculation (it uses the value of the contributor reputation, which is based off of the ratings of the contributor's assertions) and the difficulty associated with obtaining variables  $Q(f)$  (there is no mechanism to evaluate this) and  $s(f)$ , it has been simplified to:

$$Q(c) = \frac{(4 * (\frac{N(R(p(c)))}{N_{max}}) + \sum_{f \in F_c} r(f))}{(4 + |F_c|)}$$

The simplified formula uses some of the same variables as the original with the addition of

$N(R(p))$  = Normalized reputation of a contributor p

$N_{max}$  = Maximum normalized reputation of existing contributors (this increases with user scores)

The constant values were removed because of a lack of testing available to calibrate the values; they will be calculated and included in a future version. Using this formula and a sample assertion from the database that has 32 different ratings (an even mix of high and low) and was created by the user with a normalized score of 3.428 and a maximum user score of 4 would be:

$$2.678 = \frac{(4 * (\frac{3.428}{4}) + 93)}{(4 + 32)}$$

The range for this value begins at zero and continues to the maximum possible value of feedback, which is 4. This calculation normalizes the contributor score by the maximum value to assess that user's overall reputation and applies that value to the average feedback submitted. This makes the score easy to understand and compare. Since this calculation uses a variable of the contributor reputation score, and explanation of this formula is also included.

### **Data Display**

The overall strategy for the place pages is to allow users to see the highest rated assertions associated with the place being viewed to show the site visitors a compiled place entity with the properties deemed most accurate by the community of participants. From the place page users will be able to edit specific properties with a simple interface. To aid users in comparing contributions, users should be able to view all existing assertions (and their scores) for the same property.

## IV. Implementation of Project

At the beginning of this project, the options for implementing a system were building one from scratch or using an existing software package. Creating a new application would require a lot of time and other resources, but only include the features that are necessary for the project. These features would be tailored to meet the specific needs that the project requires. Existing software can save time because it already contains many of the tools needed, though it can require customization to fit the design of the project.

Sometimes customization is not an option, and another solution must be found. It was decided that for this project the benefits of an existing software package outweighed the other costs and benefits. We considered using an existing web GIS application, however none of the candidates could provide the functionality that would address the needs that the wiki already handles. While versioning is available with ESRI's ArcGIS server used with an ArcSDE database, this method of versioning does not allow us to store the multivalent properties effectively.

The base of the website is the MediaWiki package, which is an open source project, freely available to the public, and the basis for the Wikipedia site. This software package uses an open source DBMS, either PostgreSQL or MySQL, to store the data, and PHP pages to manage the data and the user interface/experience. This open architecture makes it easier to alter the wiki code for our purposes. Additional pages were created to assist with creating and visualizing the data.

Several extensions for Mediawiki were installed to assist in setting up the website. The Semantic Mediawiki extension is used to add structure to each page or place, so that each type of entry will have a preset list of properties. The *templates* tool from this extension provides the ability to create a structure that uses attribute fields similar to a relational database, with more flexibility which meets the need to store entities as structured elements with multiple properties. Semantic Mediawiki is based on the theories of the semantic web, where computers are able to interpret the meaning of information on the web. The power of this tool is that it provides a way to mimic the data structure from the design in this system, rather than the free text of a typical webpage. The Semantic add-on also has a *form* tool so that users can create an entry with an input form, rather than having to type in the properties manually (using a complex internal wiki syntax). This addresses the need to have a user interface that masks complexity of data. The semantic maps capability of the extension is also used to create and display location information to store the location of each entity represented as points.

### **Structure of the contributed data**

*Categories* are a management feature included in the MediaWiki project to classify entries. The “Category:” tag (with a descriptive name value) can be included in a template so that all entries created with a form that uses that template can be described on the category’s “Special Page”. Categories for each type of place were created in the wiki. When a place entry is created through the form, it is automatically assigned the matching category with wiki code. The categories used for this project were Branch, Building,

Mission, Settlement, Site, Stake, and Ward to correspond with each type of place the application stores.

The project’s website utilizes templates, a feature of the Semantic Mediawiki extension. By using templates, predefined structures can be created and used in entries by participants, similar to fields in traditional relational database systems.

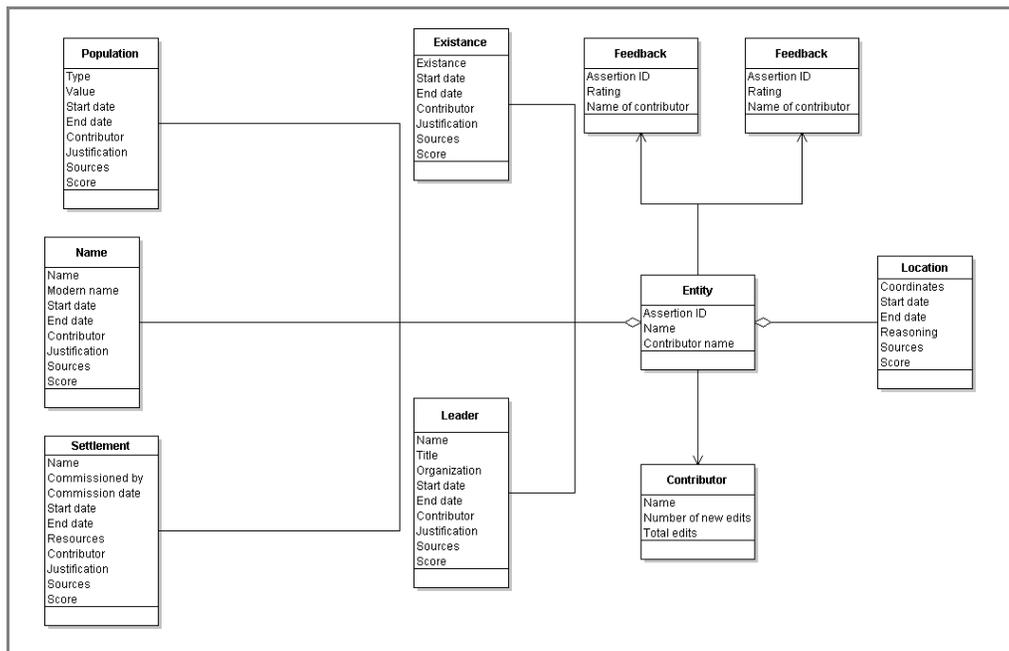


Figure 4. Data model implemented for a settlement entity.

Templates are the equivalent to the “Property” class in the conceptual data model. The template defines the set of attributes (called properties in Mediawiki terms) that are required to properly describe an aspect of the entity. The templates are referenced in the forms using combinations that describe each type of feature. For example, the Settlement form contains the Name, Existence, Settlement, Leader, Population and Location

templates describe this entry (figure 4). Figure 5 is an example of a template for the name property.

```
Owned by: {{{owner}}}
Commissioned By: {{{commissioned_by}}} Commission
Date: {{{commission_date}}} Date construction began:
Name: {{{construction_start_date}}} Date construction
{{{name}}} ended: {{{construction_end_date}}} Date razed:
{{{razed_date}}} Made of: {{{material}}} Used for:
{{{use}}} Contributed by: {{{contributor}}} Justification:
{{{justification}}} Text Sources: {{{textSources}}}
Internet Sources: {{{internetSources}}}
```

*Figure 5. Properties of the Name template.*

### **Data Creation**

The Semantic Forms extension for Mediawiki was used to create entry forms to create and edit new and existing entries. This takes advantage of the templates by creating separate input fields for each property of every template that is referenced. The location property has a map input which displays a Google Maps control (figure 6). The user can click on the map and the location text box will be populated with the coordinates for that location. The ability to search for an address (geocoding) is also available. This feature provides a mapping tool that is easy to use and familiar to the user. No more than a low level of GIS experience is needed to use this functionality. Other properties can be entered according to the templates that are related to the entity that is being created.

Location: 34.126601, -117.218791

Enter address here

Map Satellite Hybrid Terrain

Start\_date:  November 2009

End\_date:  November 2009

Score:

Contributor:

Justification:

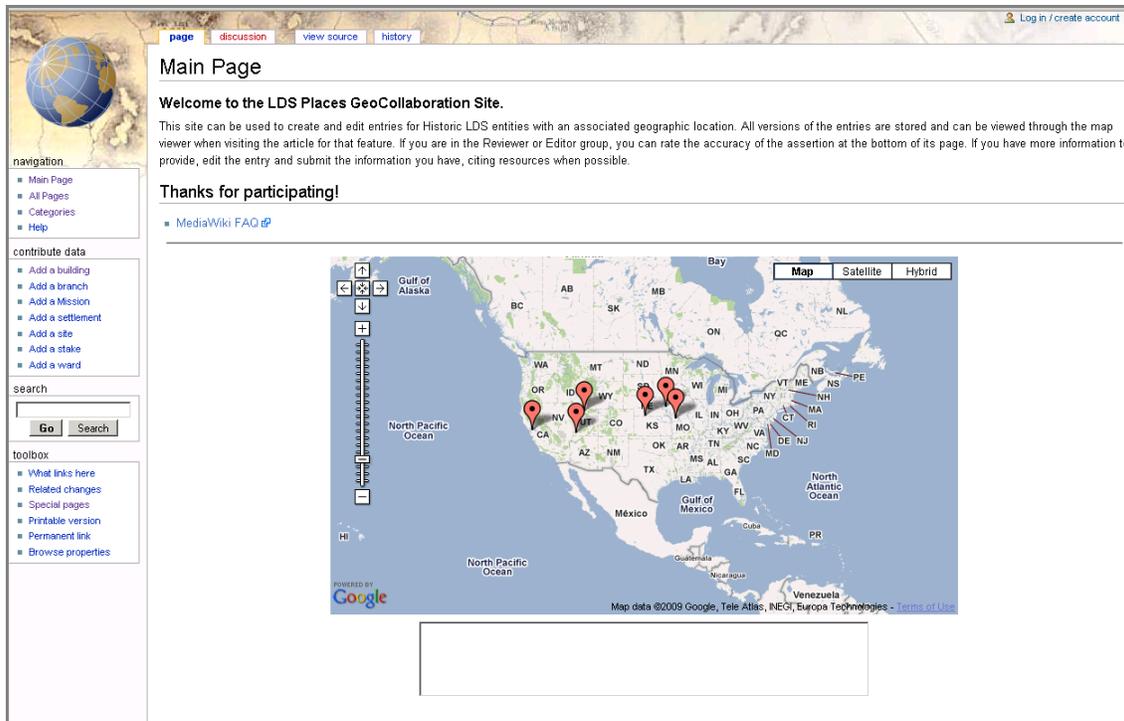
TextSources:

InternetSources:

Figure 6. A view of the location section of the data input form.

### Visualizing the data

The Main page features a customized map that displays the locations of the most recent assertion for every entity (figure 7). This quickly presents the user with an idea of how much data there is and where it is asserted to be located. Not every assertion is displayed so that the user is not overwhelmed by the amount of data. They will also be able to process a lower number of features visually.



*Figure 7. The project's home page with a map showing all contributions.*

The pages created for places will display each of its associated templates and their properties. For each type of assertion, only the one with the highest score will appear on the page. The Semantic MediaWiki add-on also provides mapping functionality through the Semantic Maps portion of the extension, which integrates the ability to use wiki text with latitude/ longitude coordinates (figure 8). When the entry is displayed, a Google Maps control is included in the page with a marker for the location with the highest score.

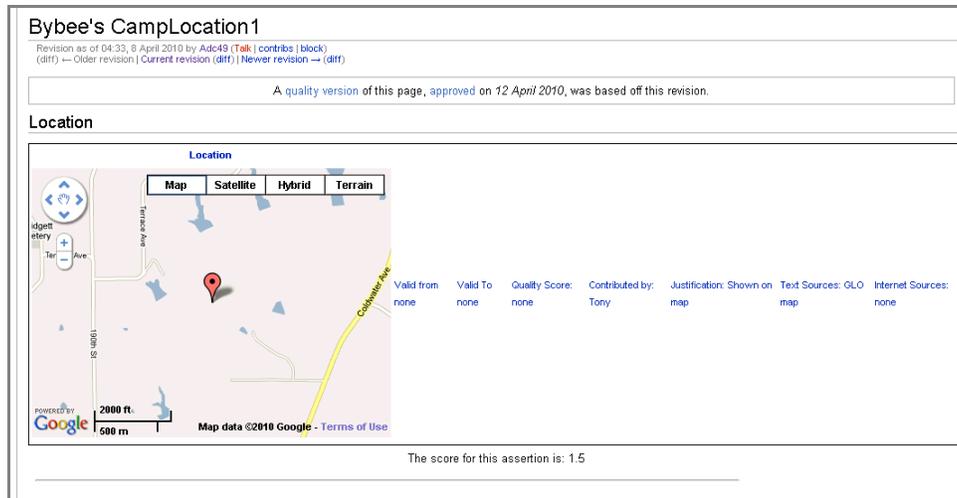


Figure 8. A view of the most recent assertion with map showing location.

When users want to see the differences in the locations of the various revisions for an individual entry, the View alternatives page meets this desire. This application uses the Google Maps JavaScript API, PHP, and JavaScript HTML Document Object Model (DOM) code and is embedded in each assertion's page. The code searches the text for geographic coordinates within a coordinates property or map display object. If a coordinate pair is found, the marker symbol is displayed on the map using that location. If there is no location, the marker will be set to 0,0 with a unique offset so there is no overlap. When the marker is clicked on, a text box will display the text contained in that version's entry. The map provides an easy way to see all the assertions to that the location of each can be compared easily. Figure 9 shows the map displaying the locations of all the assertions for a building entity, and symbolizing them based on their score. The highest ranked assertion appears in dark blue, becomes increasingly lighter to dark green and gets lighter again until yellow and white for all others.

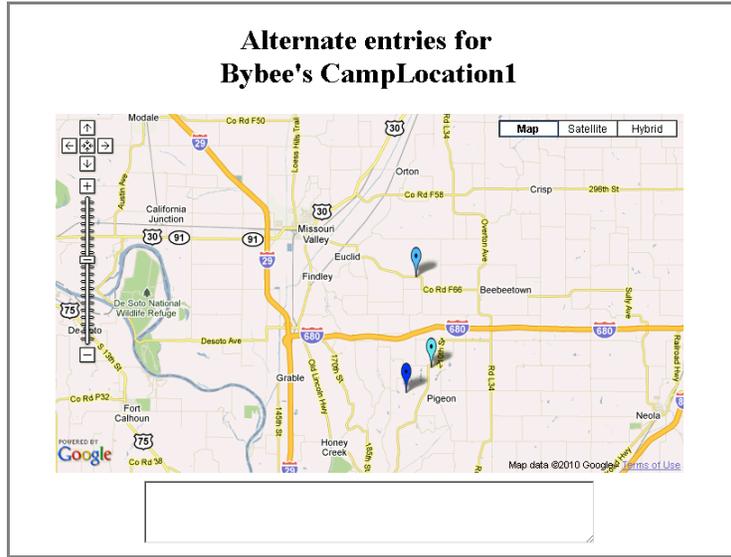


Figure 9. The embedded map showing the locations of all assertions for the entry.

Each category has its own home page that includes a map that displays the location of the latest revision of every entry that the category contains (figure 10). This provides excellent management and organizational functionality, so that entries of a specific type are easy to find and acquire a quick overview of. This also helps by being able to find features that are within a specific category.

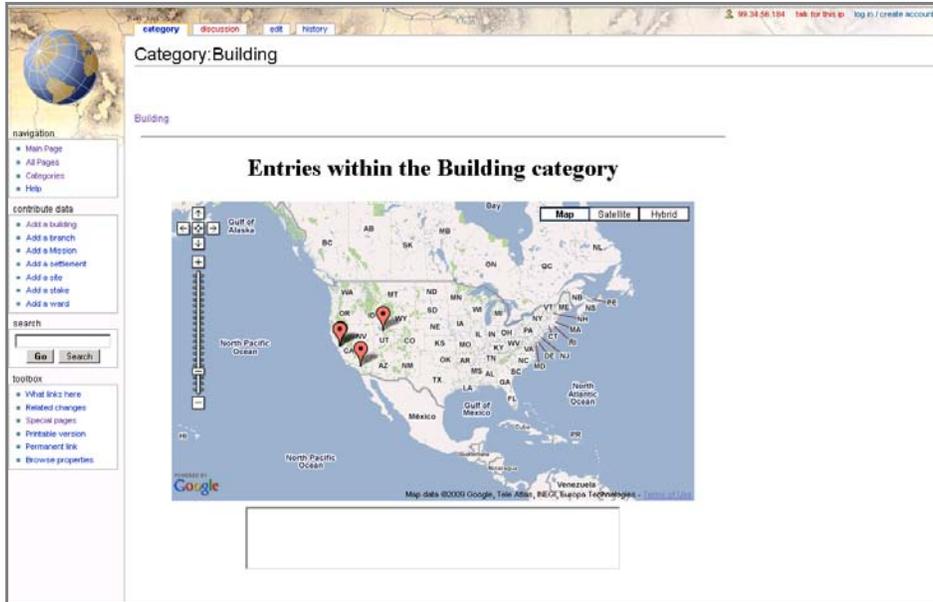


Figure 10. Map viewer displaying entries with the Building category.

## Voting

The voting system used by the site is a modified version of the FlaggedRevs extension for Mediawiki. When a user in the editor or reviewer group is logged in to the site, a flagged revision toolbar is displayed at the bottom of every article page.

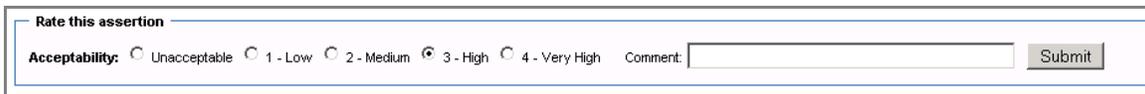


Figure 11. The flagged Revs toolbar, the user interface for submitting a rating.

The toolbar, shown in figure 11, has a set of radio buttons for selecting the rating. When submitted, the information is sent to a table in the database. By default, this table only keeps one record for each article revision, so an archive table was created to hold previous votes. Only the accuracy value is used for calculating the score of each revision, using the calculation from the previous chapter. Figure 12 is a representation of the tables related to the rating functionality.

flaggedpages	flaggedrev_archive	flaggedrevs
fp_page_id: INTEGER	fr_page_id: INTEGER	fr_page_id: INTEGER
fp_reviewed: INTEGER	fr_rev_id: INTEGER	fr_rev_id: INTEGER
fp_pending_since: INTEGER	fr_user: INTEGER	fr_user: INTEGER
fp_stable: INTEGER	fr_timestamp: INTEGER	fr_timestamp: INTEGER
fp_quality: INTEGER	fr_comment: TEXT	fr_comment: TEXT
<b>flaggedpages_pkey</b>	fr_quality: INTEGER	fr_quality: INTEGER
fp_page_id	fr_tags: TEXT	fr_tags: TEXT
<b>fp_pending_since</b>	fr_text: TEXT	fr_text: TEXT
fp_pending_since	fr_flags: TEXT	fr_flags: TEXT
<b>fp_quality_page</b>	fr_img_name: TEXT	fr_img_name: TEXT
fp_quality	fr_img_timestamp: INTEGER	fr_img_timestamp: INTEGER
fp_page_id	fr_img_sha1: TEXT	fr_img_sha1: TEXT
<b>fp_reviewed_page</b>	accuracy: INTEGER	accuracy: INTEGER
fp_reviewed	depth: INTEGER	depth: INTEGER
fp_page_id	style: INTEGER	style: INTEGER
		<b>flaggedrevs_pkey</b>
		fr_page_id
		fr_rev_id
		<b>fr_img_sha1</b>
		fr_img_sha1
		<b>page_gal_rev</b>
		fr_page_id
		fr_quality
		fr_rev_id

Figure 12. Database tables used for storing and keeping track of ratings.

## Viewing the Scores

The map viewer that displays the different article versions also incorporates the scores of each revision. Code inside the page collects the ratings for that revision, the user who submitted that rating. A calculation is performed that takes into account the score for the revision's author and the rating for the revision itself. The revisions are sorted by score and ranked in order. The first seven entries are displayed with a rank-based color marker so the order can easily be visualized. If there are more than seven assertions to display, the remaining revisions will be symbolized with a white marker to show that it exists, but is not a top ranked revision. This provides a quick and simple way for users to see what has been accepted by the community as the most accurate description for that feature. The score of the assertion is calculated using the contribution scoring formula described in the Project design chapter.

For non-spatial properties of a place, the score of that contribution is shown at the bottom of its section (Figure 13).



Figure 13. One property of a place with score and link to all assertions.

Each property section also contains a link to another page that displays all assertions for that property of the place for comparison. Figure 14 illustrates the functionality of this page.

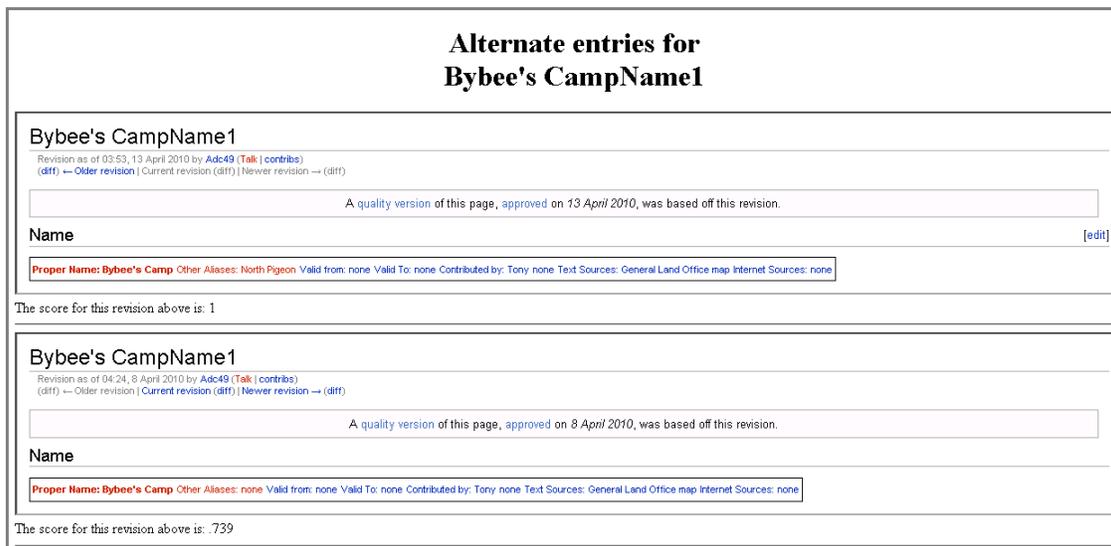


Figure 14. Comparison page to view all assertions.

## Place page layout

Each place page represents a place entity and displays all of the available properties for that type of place. There is a section for each property, which shows the attributes of the assertion with the highest contribution score for that property, its score, and a link to view all assertions for the property. Location properties are shown in a map, the first map displaying the contribution with the highest score, and the second showing all asserted locations on the same map for comparison. Figure 15 demonstrates the layout for a settlement place.

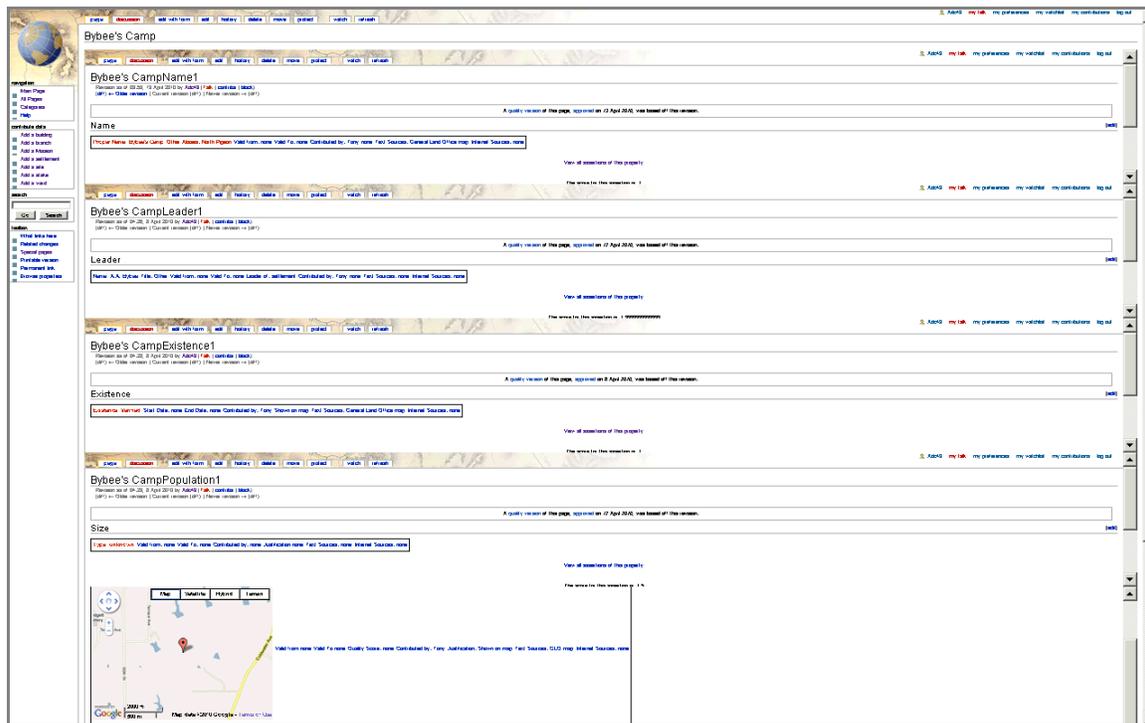


Figure 15. The page layout of a settlement place.

## V. Presentation of Results

After testing the final implementation of this project, qualitative feedback was received about the site and the discussion of this feedback follows. Testing was used by entering a small sample of real historical place data. While it was simple to enter individual properties into the forms, it was time consuming to separate the information into each property and fill out each property individually.

When calculating the contributor scores, the site only takes into account the number of edits made by participants to new and existing entities. While this method of scoring contributes is justified (Kazai and Milic-Frayling, 2009; Alonso *et al.* 2008; Wilkonson and Huberman, 2008), better results may be achieved by adopting the approach taking into account the scores of the assertions they contribute, and the quality or acceptability of their contributions. One difficulty is that the formulas are recursive, since they each rely on the value of other. However, this would place the emphasis on the quality of content rather than quantity of edits that may not provide any benefit to the knowledge of the community.

There are other capabilities and functionalities available with the add-ons installed with the site application. Several of them and others that have not been installed may be beneficial to the process of allowing a community contribute, rate, and discuss geographic and historical-based data. These should also be evaluated for the benefits they may provide while taking into account the circumstances of the individual project.

There are another two additions to the project that would improve the outcome/user experience. The first is to package this project as an installable Mediawiki extension, so that other users can incorporate these capabilities. It would also be useful to include a way to differentiate between edits so that new assertions appear differently from an altered previously existing assertion.

## VI. Conclusion

The creation of the site was successful. The Wiki site allows users to create and edit entries of historical significance to the LDS population from multiple locations with multiple users. Users are able to visualize the geographic properties of each version of each entry along with their descriptions and titles. Members of the community are able to submit feedback for each assertion and collectively score the data so that it can be evaluated and used better. Additionally, users have an interface for entering historical data and locations associated with those entities which accepts simple types of input.

This site gives the user a way to assign geographic properties to each place, and it provides a method to visualize those entries along with their associated description of the property. It also possesses the scoring system that takes into account the opinions of the entire community as well as the reputation of contributors. The wiki structure used provides the functionality to receive information from a community of users and manage all assertions made by each individual contributor. Finally, this is a website which is accessible through the World Wide Web and is available to anyone with internet access. The site is meant to run constantly and is available to the users at all times, from any remote location. However, this site is not perfect, and future work should continue to improve its functionality.

Consideration should be given to the potential users' ability to access the site, or who may not have the technical knowledge to participate, and must travel to a specific location to meet either of those needs. The recommendations include that the site be kept

very user friendly, have a site which is compatible with as many browsers as possible, and that the users be verified as accurate source contributors before their accounts are created (Alonso *et al.* 2008). This will also prevent skewed data which might occur from duplicate accounts made by the same user.

The calculation of the contributor's score should take into account the quality of their feedback, rather than the number of assertions created/edited. There should be a mechanism in place that prevents "back scratching" where users intentionally vote for each other's entries to skew the data in a certain direction. Westerheijden, Stensaker and Rosa (2007) also acknowledge this concern, though not in the context of wikis or crowdsourcing applications.

The completion of this site is a milestone for the entire community wishing to create data with geographic attributes and account for the many possible alternatives that result from using crowdsourcing methods. Now there is a method for allowing users in different locations, at different times to collaborate when creating and editing data, so that the knowledge of all those participating can be used to result in the more accurate data than can be created using any other method. This method also provides the benefit of this data having the possibility of being collected at a much faster rate. Organizations can save or preserve resources by crowdsourcing the data creation process to willing participants who have experience in the designated field. Ultimately this format allows for a more efficient method of collecting data that accounts for multivalency and provides quality control for more useful data.

## References

- Al-Kodmany, K. (2000a). Using Web-Based Technologies and Geographic Information Systems in Community Planning *Journal of Urban Technology*, 7(1), 1-30
- Al-Kodmany, K. (2000b). Public Participation: Technology and Democracy *Journal of Architectural Education*, 53(4), 220-228
- Al-Kodmany, K. (2001). Online tools for public participation *Government Information Quarterly*, 18, 329-341
- Alonso, O., Rose, D. E., and Stewart, B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum* 42, 2 (Nov. 2008), 9-15.
- Armstrong, M. (1994). Requirements for the Development of GIS-Based Group Decision-Support Systems *Journal of the American Society for Information Science*, 45(9), 669-677
- Baum, S., Stimson, R. & Yigitcanlar, T. (2003). Analyzing the Patterns of ITC Utilization for Online Public Participatory Planning in Queensland, Australia *Assessment Journal*, 10(5), 5-21
- Boonstra, Onno; Breure, Leen; Doorn, Peter. (2004). Past, present and future of historical information science. *Historical Social Research*, 29(108), 4-132
- Brewer, I. & MacEachren, A. M. (2004). Developing a conceptual framework for visually-enabled geocollaboration *International Journal of Geographical Information Science*, 18(1), 1-34
- Canepa, C., Hoyt, L., Khosla, R. (2005). Leaves, Pebble, and Chalk: Building a Public Participation GIS in New Delhi, India *Journal of Urban Technology*, 12(1), 1-19
- Carver, S., Evans, A., Kingston, R., Turton, I. (2001), Public participation, GIS, and cyberdemocracy: evaluating on-line spatial support systems *Environment and planning. B-Planning and Design*, 28(6), 907-921
- Cinderby, S. (1999). Geographic information systems (GIS) for participation: the future of environmental GIS? *International Journal of Environmental Pollution*, 22(3), 304-315
- Chua, Y. L., Wong, S. (2001). Data Intermediation and Beyond: Issues for Web-Based PPGIS *Cartographica*, 38(3&4), 63-80
- Elliott, T. (2006). New Classroom Maps for Ancient Geography. *Occasional Papers Committee on Ancient History*, 3, 61-80

Elwood, S., Ghose, R, (2001). PPGIS in Community Development Planning: Framing the Organizational Context *Cartographica*, 38(3&4), 19-33

Flaherty, Martin S. (1995). History "Lite" in Modern American Constitutionalism. *Columbia Law Review*, 95(3), 523-590

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211-221

Gregory, I.N., Kemp, K. and Mostern, R. (2001). Geographical Information and historical research: current progress and future directions. *History and Computing* 13, 7-21

Haklay, M. & Tobón, C. (2003). Usability evaluation and PPGIS: towards a user-centred design approach. *International Journal of Geographical Information Science*, 17(6), 577-592

Howe, J. (2008) *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*, Crown Publishing Group, New York, NY, 2008

Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q. (2007). Measuring article quality in wikipedia: models and evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 243-252

Jankowski, P., Nyerges, T. (2001). GIS-Supported Collaborative Decision Making: Results of an Experiment *Annals of the Association of American Geographers*, 91(1), 48-70

Jankowski, P., Nyerges, T.L., Smith, A., Moore, T.J. and Horvath, E. (1997). Spatial group choice: a SDSS tool for collaborative spatial decision-making. *International Journal of Geographical Information Science*, 11, 577-602

Kazai, G. and Milic-Frayling, N. (2009). On the evaluation of the quality of relevance assessments collected through crowdsourcing. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 21-22

MacEachren, A.M., (2000). Progress reports, *Cartography and GIS: facilitating collaboration*. *Progress in Human Geography*, 24(3), 445-456

MacEachren, A. M. (2001). *Cartography and GIS: extending collaborative tools to support virtual teams*. *Progress in Human Geography*, 25(3), 431-444

MacEachren, A. M., Cai, G., Sharma, R., Rauschert, I., Brewer, I., Bolelli, L., et al. (2005). Enabling collaborative geoinformation access and decision-making through a natural, multimodal interface. *International Journal of Geographical Information Science*, 19(3), 293-317

- McCulloch, I. M. (2008). "A blanket of inconsistencies..." The Battle of Ticonderoga, 2008 *The Journal of Military History*, 72, 889–900
- Mladenoff, D. J. et al. (2002). Narrowing historical uncertainty: probabilistic classification of ambiguously identified tree species in historical forest survey data. *Ecosystems*, 5, 539-553
- Ott, T. and Swiaczny, F. (1998). The analysis of cultural landscape change: A GIS approach for handling spatio-temporal data. *History and Computing*
- Otterbacher, J. 2009. 'Helpfulness' in online communities: a measure of message quality. In *Proceedings of the 27th international Conference on Human Factors in Computing Systems (Boston, MA, USA, April 04 - 09, 2009)*. CHI '09. ACM, New York, NY, 955-964.
- Paez, D., Bishop, I. D., Williamson, I. (2006). DISCUSS: A Soft Computing Approach to Spatial Disaggregation in Economic Evaluation of Public Policies. *Transactions in GIS*, 10(2), 265 – 278
- Powell, H. Jefferson (1987). Rules for Originalists. *Virginia Law Review*, 73(4), 659-699
- Rashid, A. M., Ling, K., Tassone, R. D., Resnick, P., Kraut, R., and Riedl, J. (2006). Motivating participation by displaying the value of contribution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 955-958
- Reinhold, S. (2006). WikiTrails: Augmenting Wiki Structure for Collaborative, Interdisciplinary Learning *International Symposium on Wikis: Proceedings of the 2006 international symposium on wikis*, 47-58
- Robertson, H. A. and McGee, T. (2003). Applying local knowledge: the contribution of oral history to wetland rehabilitation at Kanyapella Basin, Australia. *Journal of Environmental Management* 69(3), 275-287
- Sentilles, R. (2006). Identity, Speculation and History. *History & Memory*, 18(1), 120-151
- Shelton, S. and Weckerly, F. (2007). *California Fish and Game*, 93(4), 224-227
- Shiffer, M. J. (1995). Interactive multimedia planning support: moving from stand-alone systems to the World Wide Web *Environment and Planning B: Planning and Design*, 22, 649-664
- Showers, K.B., Malahleha, G.M., (1992). Oral evidence in historical environmental impact assessments: soil conservation in Lesotho in the 1930s and 1940s. *Journal of Southern African Studies* 18, 276–296

- Sieber, R. E. (2003). Public Participation: Geographic Information Systems across borders *Canadian Geographer*, 47(1), 50-61
- Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality - ICIQ 2005*, 442-454
- Stvilia, B., Twidale, M., Smith, L. C., Gasser, L. (2008). Information quality work organization in Wikipedia. *JASIST*, 59(6), 983–1001
- Talbert, R. (2006). New Classroom Maps for Ancient Geography. *Occasional Papers Committee on Ancient History*, 3, 81–92
- Warren-Kretzschmar, B., Von Haaren, C. (2006). The Interactive Landscape Plan: Use and Benefits of New Technologies in Landscape Planning and Discussion of the Interactive Landscape Plan in Koenigslutter am Elm, Germany *Landscape Research*, 31(1), 83-105
- Westerheijden, D. F., Stensaker, B. and Rosa , M. J. (2007). States and Europe and Quality of Higher Education Quality Assurance In *Higher Education*, 20, 73-95
- Wilkinson, D. M. and Huberman, B. A. (2007). Cooperation and quality in wikipedia. In *Proceedings of the 2007 international Symposium on Wikis. WikiSym '07. ACM*, 157-164
- Wood, J. (2005). 'How green is my valley?' Desktop geographic information systems as a community-based participatory mapping tool *Area*, 37(2), 159-170