# Collaborative Research Assistant

John Finlay, Instructor Neumont University
Christopher Stolworthy, Student Neumont University
Daniel Parker, Student Neumont University

## 1. Introduction

From a genealogy researcher's perspective, there are still several technological problems that are yet to be solved. Some of these problems center on sharing and collaboration. Others come from a researcher's overall user experience with genealogy software. These problems often discourage and frustrate genealogy researchers. The PhpGedView Research Assistant Module is designed to solve these problems, aid genealogists in their research, and provide a better overall user experience for them. It accomplishes this in many ways including helping to manage and track research, guiding the researcher with artificial intelligence, and simplifying online searching. The Research Assistant also greatly simplifies the data entry process by redesigning the user interface and the workflow for entering research results.

## 2. Identifying the Problems

One of the first research oriented problems that needs to be solved is a way for researchers to preserve and share their research with each other. It is important that they be able to share both research that yielded results and research that did *not* yield results. They also need to be able to track who is working on research tasks so that duplication of work is minimized.

Imagine that multiple family members, living in different cities, are researching the same family line. Without a way to track and manage the research that they are individually working on, they could easily duplicate each other's work. Also imagine that your children or grandchildren come back to look at someone you researched, how do you tell them what sources you have already looked in? Genealogists need a tool to help them track, manage, and log research.

There is also a large gap between the way someone researches information and the way they enter it into their genealogical data management system. The data entry process can become laborious. As an example, if a researcher finds a census record which enumerates 6 individuals, they now have to navigate to each of those 6 individuals and enter in multiple fact records for each person. At the same time, they must ensure that source citations are properly and consistently entered. Genealogy researchers need a way to simplify their translation of research results into genealogical data. They also need a way to simplify the entry of that data while including proper source citations.

## 3. Managing Distributed Research

For the genealogy researcher, the typical research workflow is an iterative process with four stages: 1. Analyze the data, 2. Determine possible sources, 3. Research, and 4. Enter results.

This workflow is shown below in Figure 1. After the results are entered, the researcher cycles back to the beginning, and analyzes the data again to hunt for the next clue.

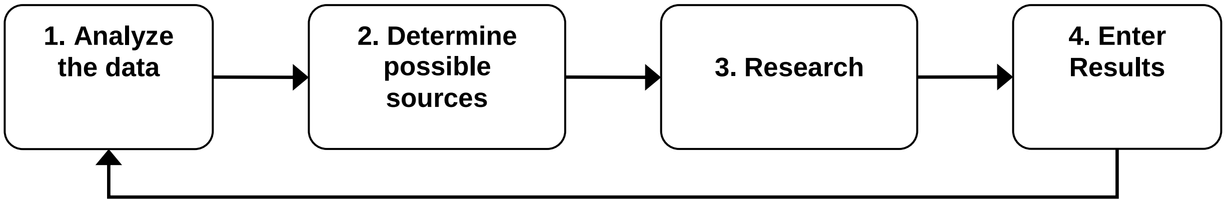
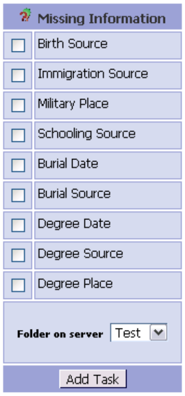


Figure 1 – Genealogy Research Workflow

While managing this workflow we need a way to track the flow of information through each stage. The Research Assistant allows users to create "research tasks" where they enter descriptions of the research to perform. These research tasks can be associated with individuals or families from the genealogy data. This allows them to keep track of all the research that has ever been done or is planned to be done for an individual. Research tasks can also be organized into virtual folders. This allows researchers the option to group research tasks according to their particular preference. For example, they may want to organize research tasks into folders that correspond to their hard-copy filing system. These research tasks can also be assigned to a specific researcher who will do the research. All of this information is available online so that any interested researchers can access the information anytime and anywhere.

## 4. Analyze the data

A researcher begins by analyzing the data they already have. During the analysis phase the researcher identifies the information they would like to research next. They can use the research assistant to create research tasks to track and manage what they plan to do.



Missing Information

□ Birth Source
□ Immigration Source
□ Military Place
□ Schooling Source
□ Burial Date
□ Burial Source
□ Degree Date
□ Degree Source
□ Degree Place

Folder on server: Test

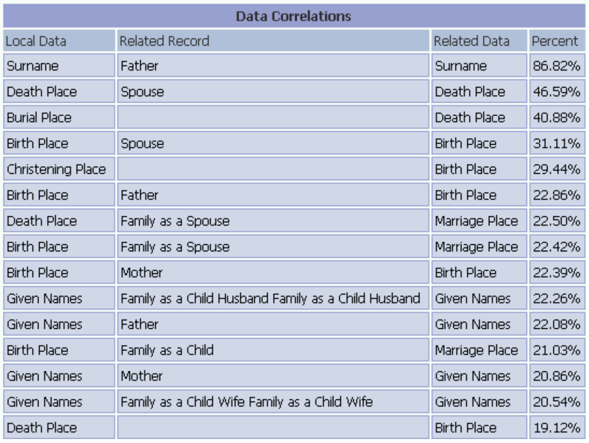
Add Task

Figure 2 – Missing information

The Research Assistant aids users in this step by looking through an individual's record and identifying missing information. Figure 2 on the right shows an example of the type of information the software will identify as missing information.

Beyond providing missing information, we have expanded the program to further assist the researcher through the use of Bayesian data mining techniques. Bayesian data mining begins by analyzing existing data to determine probabilities and establish rules. These probabilities and rules are then used to predict trends or highlight anomalies[1]. With genealogy we can apply these techniques to predict what the missing information most likely is.

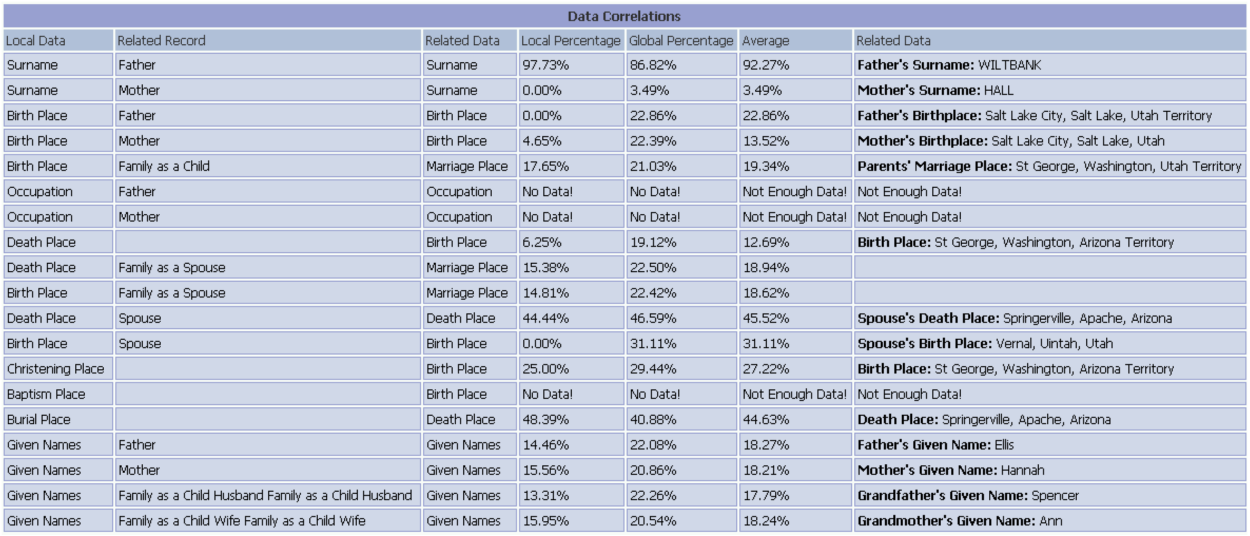
The Research Assistant currently has the tools to analyze a genealogy data set and discover the probabilities for several preset rules. The results are shown below in Figure 3. For example this figure shows that for the given data set there is an 87% probability that you will share the same surname as your father. There is no surprise there, given that this data set is from a traditional American family. But these results would be very different for cultures that follow different patronymics. It is interesting to note that for this data set it is more likely for you to be born in the same place your spouse was born (31%), than for you to be born in the same place your father was born (23%), or even the same place where your parents were married (22%). Calculating these probabilities for a specific data set allows us to give more accurate predictions.



| Data Correlations | | | |
|---|---|---|---|
| Local Data | Related Record | Related Data | Percent |
| Surname | Father | Surname | 86.82% |
| Death Place | Spouse | Death Place | 46.59% |
| Burial Place | | Death Place | 40.88% |
| Birth Place | Spouse | Birth Place | 31.11% |
| Christening Place | | Birth Place | 29.44% |
| Birth Place | Father | Birth Place | 22.86% |
| Death Place | Family as a Spouse | Marriage Place | 22.50% |
| Birth Place | Family as a Spouse | Marriage Place | 22.42% |
| Birth Place | Mother | Birth Place | 22.39% |
| Given Names | Family as a Child Husband Family as a Child Husband | Given Names | 22.26% |
| Given Names | Father | Given Names | 22.08% |
| Birth Place | Family as a Child | Marriage Place | 21.03% |
| Given Names | Mother | Given Names | 20.86% |
| Given Names | Family as a Child Wife Family as a Child Wife | Given Names | 20.54% |
| Death Place | | Birth Place | 19.12% |

Figure 3 – Data correlations from a sample data set

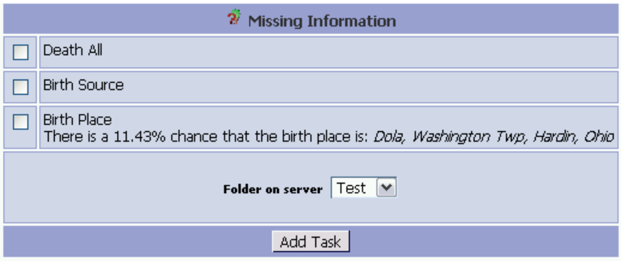
We can also run these same calculations on a particular individual and his close relatives. This provides even more localized probabilities for a person. Then combining the local probabilities with the remote probabilities we can determine the most likely values for information. An example of a table showing the combined probabilities including the related data is shown below in figure 4.



| Data Correlations | | | | | | |
|---|---|---|---|---|---|---|
| Local Data | Related Record | Related Data | Local Percentage | Global Percentage | Average | Related Data |
| Surname | Father | Surname | 97.73% | 86.82% | 92.27% | **Father's Surname:** WILTBANK |
| Surname | Mother | Surname | 0.00% | 3.49% | 3.49% | **Mother's Surname:** HALL |
| Birth Place | Father | Birth Place | 0.00% | 22.86% | 22.86% | **Father's Birthplace:** Salt Lake City, Salt Lake, Utah Territory |
| Birth Place | Mother | Birth Place | 4.65% | 22.39% | 13.52% | **Mother's Birthplace:** Salt Lake City, Salt Lake, Utah |
| Birth Place | Family as a Child | Marriage Place | 17.65% | 21.03% | 19.34% | **Parents' Marriage Place:** St George, Washington, Utah Territory |
| Occupation | Father | Occupation | No Data! | No Data! | Not Enough Data! | Not Enough Data! |
| Occupation | Mother | Occupation | No Data! | No Data! | Not Enough Data! | Not Enough Data! |
| Death Place | | Birth Place | 6.25% | 19.12% | 12.69% | **Birth Place:** St George, Washington, Arizona Territory |
| Death Place | Family as a Spouse | Marriage Place | 15.38% | 22.50% | 18.94% | |
| Birth Place | Family as a Spouse | Marriage Place | 14.81% | 22.42% | 18.62% | |
| Death Place | Spouse | Death Place | 44.44% | 46.59% | 45.52% | **Spouse's Death Place:** Springerville, Apache, Arizona |
| Birth Place | Spouse | Birth Place | 0.00% | 31.11% | 31.11% | **Spouse's Birth Place:** Vernal, Uintah, Utah |
| Christening Place | | Birth Place | 25.00% | 29.44% | 27.22% | **Birth Place:** St George, Washington, Arizona Territory |
| Baptism Place | | Birth Place | No Data! | No Data! | Not Enough Data! | Not Enough Data! |
| Burial Place | | Death Place | 48.39% | 40.88% | 44.63% | **Death Place:** Springerville, Apache, Arizona |
| Given Names | Father | Given Names | 14.46% | 22.08% | 18.27% | **Father's Given Name:** Ellis |
| Given Names | Mother | Given Names | 15.56% | 20.86% | 18.21% | **Mother's Given Name:** Hannah |
| Given Names | Family as a Child Husband Family as a Child Husband | Given Names | 13.31% | 22.26% | 17.79% | **Grandfather's Given Name:** Spencer |
| Given Names | Family as a Child Wife Family as a Child Wife | Given Names | 15.95% | 20.54% | 18.24% | **Grandmother's Given Name:** Ann |

Figure 4 – Combined data correlations with related data

Once armed with this data for a person, we can enhance the missing information to suggest the most probable data for a missing birth place or the most probable source the researcher should look in. Figure 5 to the right shows an example of this.



Missing Information

☐ Death All

☐ Birth Source

☐ Birth Place
There is a 11.43% chance that the birth place is: *Dola, Washington Twp, Hardin, Ohio*

**Folder on server** Test

Figure 5 – Missing information with birth place suggestion

There is still a great deal of future work to do in the area of Bayesian data mining for genealogy and for the Research Assistant. Research needs to be done to see if expanding the code with the ability to infer its own rules as it analyzes the data will add any value. We also need to add adjust the probabilities where they have matching data. This round also focused largely on places, but similar work needs to be done for dates.

## 5. Determine possible sources

Once a researcher knows what they want to find, they then try to determine the possible sources that might lead them to this information. Sources are generally place oriented so it is important to determine a geographic location for the desired data. It is also important to have an estimated date range in order to further narrow the possible sources of information.

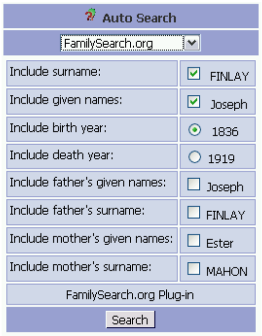
The Research Assistant can help in determining possible sources of information by taking the data from the analysis phase and querying a database to determine the most likely sources to contain that information. If the user has entered sources in their genealogy information then we can use their existing sources as part of our dataset. We can also expand this through web services to use a global database of source records. For example we could query the Family History Library catalogue or we could query an open community sponsored source repository.

Once a source has been identified, that source can be entered into the Research Assistant and associated with the research task. This allows us to track the research from a source perspective. This is useful if you plan to go and research in a source and want to easily find all of the research tasks that relate to that source.

## 6. Research

At the research stage, the researcher looks for the data in the source. This may involve looking in online indexes, or scrolling through rolls of microfilm, or driving to a cemetery. This is also the stage where hard-copies of data are created and textual results are extracted from the sources.

The Research Assistant aids the researcher in this stage by providing the ability to automatically search common online indexes with the click of a single button. Figure 6 shows the "Auto Search" tool for FamilySearch.org. You can see that much of the data for this individual is already filled in and all the user has to do is press the "Search" button. Based on the site to be searched, the user has the option of specifically selecting which data from the person's individual record they want to include in their search query.



Auto Search

FamilySearch.org

| | |
|---|---|
| Include surname: | ☑ FINLAY |
| Include given names: | ☑ Joseph |
| Include birth year: | ● 1836 |
| Include death year: | ○ 1919 |
| Include father's given names: | □ Joseph |
| Include father's surname: | □ FINLAY |
| Include mother's given names: | □ Ester |
| Include mother's surname: | □ MAHON |

FamilySearch.org Plug-in

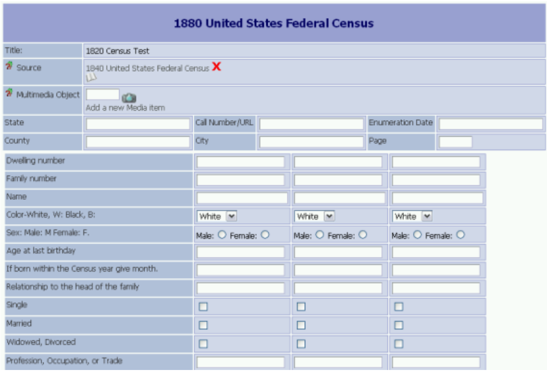
Figure 6 – Auto Search

This Auto Search tool uses a pluggable framework on the backend so that new sites can be easily plugged into the program. We currently have Auto Search plug-ins for Ancestry.com, Ancestry.co.uk, FamilySearch.org, Genealogy.com, EllisIslandRecords.org, GeneaNet.org, and WeRelate.org. In the future through web services, screen scraping, or ScreenCrayons[2], we hope to be able to pull information directly back from these sites and add it to the genealogy data. It has also been suggested that we provide a combined search of all of these sites and display the results in a combined format.

## 7. Enter results

Once the results have been accumulated, they now have to be translated into genealogical data and entered into your genealogical database. It is important at this stage to make sure that all of the data that is being entered is properly cited with the source where the data came from.

At this stage in the Research Assistant, the user chooses to complete the research task that they have been working on. When completing the task they will be prompted to choose a source citation form to enter their information. For example if the user was looking up information in a census record, they can choose to enter their results through a census citation form. Figure 7 shows an example of an 1880 Census Source Citation form. These forms are meant to more closely match the layout and format of a source's extracted data. Again the source citation forms use a pluggable architecture so that new forms can be created and added easily.



1880 United States Federal Census

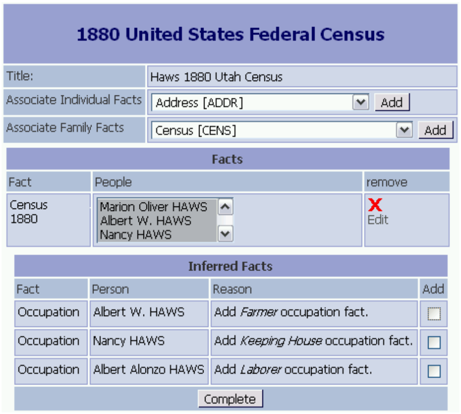
| | | | | | |
|---|---|---|---|---|---|
| Title: | 1820 Census Test | | | | |
| Source | 1840 United States Federal Census | | | | |
| Multimedia Object | Add a new Media item | | | | |
| State | | Call Number/URL | | Enumeration Date | |
| County | | City | | Page | |

| | | | |
|---|---|---|---|
| Dwelling number | | | |
| Family number | | | |
| Name | | | |
| Color-White, W: Black, B: | White | White | White |
| Sex: Male: M Female: F. | Male: ○ Female: ○ | Male: ○ Female: ○ | Male: ○ Female: ○ |
| Age at last birthday | | | |
| If born within the Census year give month. | | | |
| Relationship to the head of the family | | | |
| Single | □ | □ | □ |
| Married | □ | □ | □ |
| Widowed, Divorced | □ | □ | □ |
| Profession, Occupation, or Trade | | | |

Figure 7 – Example 1880 Census Citation Form

After entering the source citation and filling out the appropriate data extraction form, the Research Assistant uses this new data to automatically infer new facts. There is a wealth of genealogy information in a census record for example. We can learn such information as birth dates and places, marital status, occupation, and family relationships. Because the user entered the data into a structured form, we can automatically compare it with existing data and infer new data from it. The user then has the option of choosing what inferred data they want automatically added. Refer to Figure 8 for an example of some inferred facts that were generated from a census citation.

In most programs, the user must first navigate to the person they are interested in, and then enter the fact data for that person, along with the source citation. They then navigate to the next person and repeat. This is an especially time consuming and tedious data entry process. Especially when using a program that does not allow linking to source citations, or at least the ability to copy citations from one fact to another. Through the research assistant you can enter all of the data gleaned from a source on a single screen and choose the people it should be associated with. An example of this can be seen in Figure 8 to the left. Once this data is submitted it will be added to the people's individual records with all of the proper citations.



1880 United States Federal Census

| | |
|---|---|
| Title: | Haws 1880 Utah Census |
| Associate Individual Facts | Address [ADDR] Add |
| Associate Family Facts | Census [CENS] Add |

Facts

| Fact | People | remove |
|---|---|---|
| Census 1880 | Marion Oliver HAWS<br>Albert W. HAWS<br>Nancy HAWS | Edit |

Inferred Facts

| Fact | Person | Reason | Add |
|---|---|---|---|
| Occupation | Albert W. HAWS | Add *Farmer* occupation fact. | □ |
| Occupation | Nancy HAWS | Add *Keeping House* occupation fact. | □ |
| Occupation | Albert Alonzo HAWS | Add *Laborer* occupation fact. | □ |

Complete

Figure 8 – Citation Facts and Inferred Facts

Any factual data entered through the Research Assistant will not be editable through the normal editing interfaces. You wouldn't want fact information gleaned from a source extraction to be changed somewhere else. Also if the information were updated you would want that same information updated everywhere else it is used.  So if any user chooses to edit a fact that was entered through the research assistant, they will be taken back to the research assistant to edit the data.

## 8. Conclusion

The Research Assistant helps the genealogy researcher at each stage of their research workflow.  Using artificial intelligence techniques it can aid the user in analyzing their data.  It can help the researcher find the information they are looking for and it can help them enter that information into their genealogy database.  It integrates closely with the PhpGedView genealogy application so that research is connected to the data.  Because it is an internet application, the data is available anytime and anywhere to anyone who needs it.

## References

1. Lauria, E. J. and Tayi, G. K. 2003. Bayesian data mining and knowledge discovery. In *Data Mining: Opportunities and Challenges*, J. Wang, Ed. Idea Group Publishing, Hershey, PA, 260-277.

2. Olsen, Dan R. 2005. "Screen Annotation of Family History Sources", Family History and Technology Workshop 2005.

## Credits

Most of the implementation work on the Research Assistant presented here was performed by students from Neumont University under the instruction of John Finlay. Special thanks are due to those students and to the school for their contribution to the genealogical and open source community.