Achieving Complete Coverage for the 1910 U.S. Census on the Family Tree

Joseph Price
Kevin Richins

BYU Record Linking Lab

Abstract

The 1910 United States Census includes 92 million people of which 68 million were part of a family with children. In this paper, we describe an approach we are using to help ensure that each of these families has a profile on the Family Tree. Our approach includes an automated tree add tool combined with several mechanisms for allowing volunteers to access record hints for the 1910 census. We describe how this approach can dramatically reduce the amount of duplication on the Family Tree, create a platform for evaluating various record linking methods, and facilitate some amazing discovery experiences, especially when combine with the 1950 census or other record collections.

The Family Tree at *familysearch.org* is a public, shared wiki-style genealogy platform that anyone can use to gather together information and records about deceased individuals. The most common use of this platform is for people to compile research about their own ancestors and then to collaborate with other people who might share common ancestors. Many people have been able to use the Family Tree to trace their ancestors back many generations and across several centuries. Other users focus on collecting information on their "cousins" or the descendants of one of their ancestors. There have also been a number of community projects in which profiles are created on the Family Tree for individuals that lived in a particular location or have a particular surname. The combined result of these different approaches is a Family Tree that now includes profiles for over 1.2 billion deceased individuals from all over the world.

The wiki-style format of the Family Tree makes it possible for anyone in the world to add new people to the Family Tree, edit information, attach sources, and upload photos and stories. This shared tree for the human family has the potential of becoming one of the most important datasets for social science research. For example, the Family Tree could eventually include information about everyone living in various countries right before the 1918 pandemic as a way to better understand the long-run impact of this event and examine the interventions that were successful at ameliorating the impact of that pandemic. Many countries in the world have full-count census records or vital registration collections in the 1910s with high coverage that could make it possible to have some information about everyone living in the country right before the 1918 pandemic. We use the 1910 U.S. Census in this paper as a demonstration of what is possible by combining a large rich record collection with the tools and resources that are available when using the Family Tree.

The 1910 United States Census includes 92.2 million people. The census had 32 columns and included information about name, gender, age, birthplace, mother and father's birthplaces, marital status, and relationship to the household head. Individuals doing research on the Family Tree often draw on information from this census to help find more information about their family and add in family members that might have been missed. In order to validate the information they add, they often attach the 1910 census as a source to the profile. In fall of 2021, 54.5 million people from the 1910 census had been attached to a profile on the Family Tree (58.9%). In addition, FamilySearch uses machine learning to find possible matches between the census and profiles on the Family Tree, and there were an additional 7.1 million people from the 1910 census that had a possible match on the Family Tree. The purpose of this paper is to develop ways to help ensure that the other 30.7 million that do not presently appear to be on the Family Tree can have a profile on the Family Tree as well.

First, we have developed an automated tool that can identify families from the 1910 census where no one in the family appears to be on the Family Tree. The automated tool then uses the information from the census to create a profile of each individual on the Family Tree, connect them together as a family, and attach the census record as a source. In the spring of 2021, we used this tool to create profiles on the Family Tree for 13.1 million individuals. We also enlisted volunteers to help us attach additional sources to these families and to use the information from these sources to help expand the connections for each family so that they become connected with other families. We have continued to monitor these families to see how often people connect with them when doing their own research.

Second, we have helped develop several tools that allow people to attach the 1910 census records to profiles already on the Family Tree. We have shared these hints with volunteers using

our Customized Hints App, an email campaign, and by partnering with Goldie May, a genealogy company that created a great volunteer user experience using the hints from the 1910 census. Many of these hints are for families where some members of the family already have profiles on the Family Tree while others do not. These hints make it possible to add new people to the Family Tree using information from the census record.

Third, the 1910 census includes about 8 million individuals that were not living with any relatives at the time of the census. This group includes roomers, boarders, and servants. This group is much less likely to have a profile on the Family Tree and are also less likely to have record hints if we add them to the Family Tree. This makes it much more likely that these individuals will remain as unconnected nodes in the network structure of the Family Tree. We have chosen not include this group as part of the main focus of this paper, although we provide a discussion about how this group might be better included on the Family Tree.

Achieving full coverage on the Family Tree will have at least three immediate benefits. First, ensuring that each person alive in the U.S. in 1910 has a profile on the Family Tree can dramatically reduce the number of duplicate profiles. The census record provides a one-to-one mapping between the population and the Family Tree. Thus, each person in the census should be on the Family Tree only once. In addition, every profile on the Family Tree of someone living in the U.S. in 1910 should be attached to the 1910 census. Ensuring that all of these matches are created will result in an identification of most of the duplicates on the Family Tree for this group since two profiles that are connected to the same census record can be merged together (or one of the profiles should be detached from the census record).

Second, the Family Tree can provide a powerful way to test the predictions of different record linking models. There is considerable academic interest in developing better ways to link

individuals across records. The linkages on the Family Tree can provide important training data from developing these methods but can also be used as a platform to allow researchers to validate their predictions. The end result of our project will be that everyone in the 1910 census will have a profile on the Family Tree. Therefore, anyone that creates predicted links between the 1910 census and any other record could use the SourceLinker tool on FamilySearch to check the predictions of their model. Each additional record that is attached to a profile increases the accuracy of the information about that individual and makes it more likely that he or she can be linked to additional records.

Third, the Family Tree coverage for the 1910 census could be combined with the 1950 census to create an amazing new discovery experience for individuals. A young adult in the US could have up to 8 great-grandparents present in the 1950 census (and some of their grandparents). Each of these great-grandparents will have nearly all of their 4 grandparents and many of their 8 great-grandparents in the 1910 census. By linking together the 1950 and 1910 census (and the census records in between), it would be possible to help an individual with ancestors from the United States to be able to see a full 6-generation fan chart of their family. It would also facilitate discovery experiences that allow individuals to connect with a large number of cousins. There are already some discussions about discovery experiences built around claiming an ancestor to allow living relatives to connect to him or her. The same functionality is now possible when people attach sources or upload photos and allow other users to view their relationship.

Measuring Current Coverage of the Family Tree

There were 92.2 million people included in the 1910 United States Census. The first way that we can identify which of these people are on the Family Tree is to check for what fraction of the people in the 1910 census have been attached to a profile on the Family Tree. Each profile on the Family Tree has a section for sources, and FamilySearch provides a tool called SourceLinker that makes it possible to compare information from a record to the information for a profile and determine if it is a match. SourceLinker is set up so that records that include multiple family members can all be decided and resolved at the same time. This is particularly helpful for census records where the record can include many family members and the information from each person can help to know if the entire family is a match. This allows matches for specific individuals to be made even when their information is incorrectly transcribed in the census record.

In Table 1, we provide statistics about the coverage rate of the Family Tree in November 2021. We split the coverage into the number of individuals from the 1910 census who have been attached to a profile on the Family Tree and also the number of people that have a possible record hint. The sum of these two numbers provides an estimate for the fraction of the US population that has a profile on the Family Tree. We start by reporting the coverage rate for the full population in the 1910 US census. Among these individuals, 62.8% are attached to a profile on the Family Tree and there are an additional 9.2% who have a record hint to a profile on the Family Tree. These results in a combined coverage rate of 72.0%.

We then narrow our sample into specific groups using information about each person's relationship to the household. First, we focus just on people who are part of a nuclear family relationship which we define to be people who report being the head, spouse, son, or daughter.

This reduces our sample to 78.0 million people and increases the coverage rate to 81.6%. Second, we focus on just households that have at least two household members who are part of the nuclear family. This limits the sample to 75.7 million people and increases the coverage rate to 83.5%. Finally, we focus on just families that have at least one son or daughter in the household. This limits the sample to 67.9 million people and increases the coverage rate to 88.8%.

The results in Table 1 highlight one of the decisions that we make for the analysis of this paper. We chose to focus just on individuals who have a direct nuclear-family relationship to the household head and exclude individuals like servants and borders or other relatives. Most of the methods that we describe in this paper are based on used family connections to help extend the coverage on the Family Tree. Later in this paper, we describe some approaches that could be used to include these individuals on the Family Tree.

We also give our main focus to families with children in the 1910 census since the children in these families will be the parents in the 1950 census. Thus the 1910 census becomes a key connector that could amplify the discovery experiences that are possible with the public release of the 1950 census. Focusing on children is also advantageous since we will have information for at least one of their parents which makes it easier to find them in birth, marriage, or death records which often the parents' names. This can be particularly useful for linking women across records. Finally, if the approach we use in this paper were used across each census year, then it is possible to ensure that all adults who lived in the US are on the Family Tree by using the census record that occurred when they were a child.

In Figure 1, we provide some insight about the change over time in the growth rate of the Family Tree. From among the sample of 67.9 million people who were part of a family with at

least one child, we drew a random sample of 100,000 white individuals and 100,000 Black individuals. We then checked to see which of the individuals in this random sample had their 1910 census record attached to a profile on the Family Tree. We then used the contributor information for each of these profiles to identify the date that the profile was first added to the Family Tree.

FamilySearch.org was created in 2012 and so we don't have any information about the date people were added prior to that but we do know that they were added to the databases that were compiled together into the original Family Tree prior to 2012. We find that in 2012, about 30% of White individuals and 2% of Black individuals who were part of a family with children in the 1910 census had a profile on the Family Tree. Figure 1 shows the growth rate for both of these groups across the next 10 years. By 2020, the coverage rate for White individuals has reached about 65% and the coverage rate for Black individuals was about 10%. Within the next two years, the coverage rate for white individuals had reached 90% and Black individuals had reached 75%.

All of the information provided in Figure 1 is just about profiles on the Family Tree where the 1910 census record is attached. In addition to these profiles that are already attached, there are an additional 6.5 million individuals in the 1910 census that have a very likely match on the Family Tree. FamilySearch has a machine learning algorithm that looks for possible matches between records and profiles, and it uses these matches to provide users on the website with record hints. These record hints appear in the upper right corner of the profile page as a blue icon, and users that click on them can decide whether to attach these records using SourceLinker.

The Match Team at FamilySearch provided us with a match file that they created in November 2021. Table 2 includes the statistics for these matches by gender, relationship to

household head, and match quality. We exclude from this table any records that are already attached, and we also only use the record hint that has the highest match score (for records that have multiple possible matches to the Family Tree). As such, the different match score groups in Table 2 are mutually exclusive, and the denominator is based on the set of records that are not attached to a profile already. From this table, we find that about 50.5% of male heads of household and 45.1% of female heads of household have a 5-start hint, compared to 50.1% for people listed as the spouse, 52.3% for sons and 5.18% for daughters. We find that about 7.4% of the full sample has a 4-star hint and 4.6%% have a 3-star hint. Thus, altogether, among the people who are not already attached to a profile on the Family Tree, there are only 36.7% who have no possible match on the Family Tree.

Achieving Complete Coverage

The previous section highlights that about 87.8% of the 67.7 million people in the 1910 US census currently have a profile on the Family Tree. In this section, we describe three approaches that we will use to help the remaining 8.2 million people also have a profile on the Family Tree. In this section, we first describe the automated approach that we have developed to add families to the Family Tree. Second, we describe our crowd-sourced method of allowing volunteers to engage with record hints as a way to add additional family members from the census to the Family Tree. Third, we discuss how we address this issue of adding individuals who are not related to the core family in the home and some of the issues about information quality for these individuals.

Before describing these three methods, we will discuss the benefits of achieving full coverage on the Family Tree. Many of our methods involve using information from a record to

create a profile on the Family Tree. At that initial point, the information on the Family Tree will be an exact replica of the information in the record. This raises a question of the value of creating a profile on the Family Tree when individuals can search for that same information in the record. We argue that a profile on the Family Tree provides several features that are not currently possible for all records.

First, any information on a profile can be edited very easily. It is currently possible for some of the fields on census records to be corrected by users, and this will eventually be possible for most fields on most records. However, this is already very easy to do on a profile, and these user corrections on the Family Tree could potentially be used to help improve the quality of the indexed records. Many individuals have incorrectly spelled names on indexed records or nicknames that can make searching for these records more difficult. Profiles on the Family Tree provide the ability to immediately edit the information and also provide comments on why the information is correct.

Second, profiles make it possible to upload photos, stories, and notes. The more contributors that we can attract to work on the Family Tree, the more likely it is that we will gather the complete set of information about each person who ever lived. Photos and stories provide the type of discovery that can draw new types of contributors to work on the Family Tree. In addition, these photos and stories often provide contextual information and additional clues that can help identify other record sources for the individuals. Family photos can help identify the need to search for additional family members. Photos can often include clues that point to a birthday, wedding day, or date of a funeral. Notes can provide important ways for contributors to collaborate and share their research with each other. Currently, none of these

features are available for records (but some of these could be added as features to record collections).

Third, profiles on the Family Tree provide several ways for living relatives to find each other. Each time that a person attaches a source or uploads a picture or story, the username of the contributor is recorded. For individuals who would like to identify connections with living relatives, they can allow others to click on their user name and access email information or enable messaging through FamilySeach. They can also activate a feature that allows other contributors to see how they are related based on connections in the Family Tree. The Family Tree only makes public information about deceased individuals for privacy reasons, but the connection tools on Family Tree allow for individuals to opt in to ways for others to contact them or see their relationship.

Fourth, the information from profiles on the Family Tree is searchable on the web and can be accessed without having a FamilySearch account. Neither of these is currently available for record collections. FamilySearch has created public-access versions of the profiles on the Family Tree that can be easily shared on social media or other websites. FamilySearch also provides a search portal for these public-access profiles at *ancestors.familysearch.org*. These public-access profiles provide a life summary, photos, timeline, and information on all immediate family members. Visitors to these public-access profiles cannot view the actual records that are attached to the individuals unless they log into FamilySearch. However, this becomes a natural place for people to set up a FamilySearch account so that they can access the records and are able to make edits or additions to the profile.

Note that all of these benefits of creating profiles on the Family Tree are valid even in the case in which the person added to the Family Tree does not have any family members. A

singleton node on the Family Tree can still provide a starting point for people to learn more about their family and contribute what they know or photos and stories that they have. However, once a profile on the Family Tree becomes more interconnected with other profiles, then the Family Tree provides a host of discovery experiences. These include allowing people to see how they are related to each other, how they are related to family people, from what countries their ancestors are, which of their ancestors has a similar face, and the ability to access some of the millions of record hints that are created each year. It is this ability to draw more individuals into the Family Tree that motivates the approach that we describe in the next three sections.

Automated Tree Growth

In fall 2020, we created a computer algorithm that could use information from a U.S. Census record, create a profile on the Family Tree for each individual, link them together with the immediate family members identified in the census record, and attach the census record as a source to the profile. This automated process can also be done by hand. There is a feature on FamilySearch that allows users to add an unconnected person to the Family Tree and then do the same steps the automated process would do. However, our automated add family tool allows us to free up volunteer time to work on the aspects of family history that involve more judgment and discernment (such as attaching additional sources, fixing names, deciding between conflicting dates, and merging duplicates).

We collaborated with FamilySearch to determine the inclusion criteria for this sample. The sample that we added to the tree had to meet the following criteria. First, the individual had to be listed as head, spouse, son, or daughter on the census record. Second, the household had to have at least one parent and one child. Third, we excluded mothers who would have been less

than age 17 or older than 50 when the child was born. Fourth, we excluded families where the head and spouse differed in age by more than 12 years. Fifth, we dropped any families where anyone in the family were on a list of inappropriate names that FamilySearch provided us and where any member of the family had a "*" or "?" in their name. Altogether, there were 14 million individuals who met these criteria.

One of the reasons FamilySearch chose some of the restrictions was to ensure that children are linked to their biological parents on the Family Tree (this includes the age restriction for mothers and the age gap restriction for couples). However, the Family Tree allows for multiple types of parent-child relationship (biological, step, and adopted). An individual can be linked to multiple parents and there is an option to define the type of parent-child relationship. Step-families are an important relationship to document on the Family Tree, and many children will spend more of their childhood being raised by a step-father than by their biological father. In many cases, they might not even know who their biological father is.

There were also 600,000 individuals who are part of a family where someone's name in the family includes a special character. There are a few ways to address this group. First, the BYU Record Linking is currently auto-indexing the name field from the 1910 census, and it is possible that the machine learning algorithm couples with post-processing to match incomplete names to name dictionaries. This could help to identify the correct spelling for these names that include "*" and "?" (both of which are used to indicate uncertainty about certain characters in a name). Second, we can link these families to other census records and use the name spelling in the other record to make a correction to the 1910 census. Record linking algorithms can match families across census records even when certain names have been misindexed or include special characters. Third, even without the ability to fix these names, these families could still be added

to the Family Tree. It is obviously less than ideal to include a name with special characters on the Family Tree, but it is likely to be better than that family not being on the tree at all. It is likely that as family members connect with the 1910 family, they will be able to use other records and their own knowledge to fix the name.

Record Hints

The most important criteria that we used before using our add-family tool was to make sure that no person in the family had a likely match on the Family Tree. We used match files provided by FamilySearch to identify which individuals in the 1910 census are attached to a person on the Family Tree or have a possible match. For families in the 1910 census where any individual in the family was attached to the Family Tree or had a possible match, we did not use the add-family tool for anyone in the family. However, we found that among these families where at least one person was already on the Family Tree, there were often other family members who were not on the Family Tree.

We used these census records where part of the family was on the Family Tree but other family members were not to create tree-extending record hints. For example, a family in the 1910 census record might include a head and spouse with their four children. We can use the match files for these records to check which of these six individuals already has a profile on the Family Tree. A tree-extending hint would be created if the head and spouse in this family had a profile on the Family Tree and some of the children but where there is at least one of the four children who does not have a profile on the Family Tree. Someone using SourceLinker would then be able to add in the missing children for this family using the information from the census record.

We have helped develop several tools that allow volunteers to help attach these tree-extending and normal record hints for the 1910 census to the Family Tree. First, we developed a customized hints web app that allows volunteers to access hints by surname or place of residence. Once a volunteer selects their search query, they will see a set of dots on a map, each of which represents a distinct record hint. The location of the dot on the map is based on the geo-coordinates of the town of residence, which we randomly slightly jigger so that the dots can each be seen when a volunteer zooms in rather than all being exactly on top of each other. When a volunteer clicks on a dot in the app, it takes them to SourceLinker on FamilySearch. Once a volunteer has clicked on a dot, that dot will disappear.

Second, we were able to partner with the genealogy company Goldie May who created a volunteer experience that allows people to search for hints by surname or place of residence. Goldie May will automatically open the hint in SourceLinker and allow the user to determine if the record hint is a match or not. If the user decides it is a match, they can click a button in Goldie May which will then automatically attach the census record for each person in the household. This computer-assisted approach can dramatically reduce the time required to attach record hints and allows volunteers to spend their time on the key moment of deciding whether the record is a match and less time actually clicking (a census family with 6 people requires 18 mouse clicks on Sourclinker to complete but one two mouse clicks on Goldie May).

Third, we created an email campaign in which over 250 volunteers signed up to receive a daily email that provided links to 5 record hints. The email campaign was designed to work on a phone or computer. This approach allows volunteers to contribute in a small way each day. The five hints take about 5 minutes to complete and can be accessed in one click anytime directly from the email that is shared. The email campaign has the potential of being very customized to

the ability and interests of the individual helping (including providing hints for their surname or where they live).

Fourth, we created a set of Google Sheets for each state that provided some statistics for each city about the coverage rate on the Family Tree for the 1910 census. These statistics could be sorted by county or city and provided information on the number and fraction of people from that city in the 1910 census that have a profile on the Family Tree and whether they have the source attached. We then provided a tab in which the urls of the record hints for the state were sorted by city and county. Clicking on the url in the Google Sheet would take volunteers to SourceLinker and then they could mark this on the Google Sheet.

The key thing about all four of these approaches is that they interact directly with the SourceLinker tool on FamilySearch. All of the decisions about record hints and the attaching of records happen on SourceLinker and the tools we helped developed are simply ways of allowing volunteers to access record hints to work on. While most work on the Family Tree has historically been done by individuals working on people to whom they are directly related, the tools that we have developed facilitate community projects in which individuals are working for record hints for a specific surname (similar to one-name studies) or a specific location (similar to community reconstitution projects).

All of the approaches described in this section involve human volunteers making decisions and attaching the sources. We can use some features about the record hints to sort them based on their difficulty level. The match score that FamilySearch creates can be one of these features, and we can combine this feature across multiple family members. For example, if we have a family with several members that all have really high match scores to a census record, and all of those matches point to the same person in the census record, then we can be very confident

that this is a correct match and this would be an easy hint. In contrast, if there is a situation where there is conflicting information between the family on the Family Tree and the family in the record, or where the family is smaller and the match scores are lower, then we would be less confident about the match. This would be a more difficult hint.

This ability to identify the difficulty level of hints makes it possible to include a much larger group of volunteers to help with the project. People with less experience with family history can help with the easy hints while those with more experience can focus their time on the hints that are best suited to their level of expertise.

Individuals not related to the household head

The most challenging group from the 1910 census to add to the Family Tree are individuals who are not related to the household. This group includes roomers, boarders, and servants. While 79.4% of individuals in the 1910 census with any type of familial relationship with the head of household have a profile on the Family Tree, the coverage rate for these unrelated household members is only 10.8%. In addition, when these individuals have profiles on the Family Tree it is harder to connect them to their extended family. A further complication is that the census was usually reported by one individual in the household (with a mark in the census to indicate who that person was), so the information for an unrelated individual in the household is likely to be less accurate than someone who is more closely related to the person who reported the information for the census (especially if the roomer or boarder wasn't home when the census was enumerated).

This leaves an open question about what to do with 9.6 million people in the 1910 census who do not have a familial relationship with the household head. An interconnected tree is the

most useful in terms of creating discovery experiences for people who use FamilySearch. However, there is a reasonable argument to be made to, at the very least, create a placeholder on the Family Tree for every individual in the 1910 census. Creating a placeholder profile would allow other records to be attached for these people and it is possible that these additional records could provide information about their parents or other family members. As advances are made in record linking techniques, it might become possible to better link individuals who have sparse or incorrect information. In addition, as we link a larger fraction of the population from the 1910 census to other census records, there will be fewer possible matches left for each roomer and boarder in the 1910 census. As the set of possible matches shrinks, the ability to identify a unique and correct link will also increase.

It is also important to acknowledge for some of the most marginalized groups, it is possible that the 1910 census might be the only opportunity to identify that person in a record collection. This is particularly true for specific immigrants from countries where records might not be available during this time period. If one of the goals of the Family Tree is to provide a profile for everyone who ever lived on earth, it will become important to develop approaches that are inclusive as possible for groups that have low baseline probabilities of being identified in other record collections.

Community Reconstitution Projects

One of the most important reasons to ensure that everyone living in the United States in 1910 has a profile on the Family Tree is that it can provide the base layer needed to create a community reconstitution project. Having complete coverage on the Family Tree for a specific year can provide a large motivation for individuals to contribute what they know to the Family

Tree in hopes of connecting to this base layer of individuals on the Family Tree. If everyone in a community were to pursue this strategy, they would be able to see how they are related to each other and also how they are related to other people that played a prominent role in the community. These types of discovery experiences can be an important motivator to encourage more people to contribute to the Family Tree.

The year 1910 is far enough back that everyone on the tree can be safely added to the Family Tree without concern of individuals still living today being included. It is also far enough back that individuals in the 1910 census are likely to have a few generations of descendants to provide large interconnected networks of related individuals to use for family history discovery experiences and scientific studies. However, the year 1910 is also close enough that it would not require much effort for most people to connect back to an ancestor who was alive in 1910. For most adults today, they would simply need to know something about one of their great-grandparents (or maybe a grandparent).

In this section, we describe a community reconstitution project from Tacoma, Washington that illustrates the way in which the combination of the automated add-family tool and record hints can be used to ensure that everyone has a profile on the Family Tree. We also show how attaching additional record hints can help connect these families to additional family members and allow the profiles to become more interconnected.

There were 77,923 individuals who were part of a family with at least one child living in Tacoma, Washington, in the 1910 census. Of these individuals, 96.6% now have a profile on the Family Tree. Figure 2 provides the timing of when these profiles were first created. Note that prior to the start of the 1910 Census Project, about 65% of individuals from this census were

already on the Family Tree. The 1910 Census Project in March 2020 provided a notable increase with the coverage rate reaching about 90%

This growth in the coverage made it possible for us to help with a community project to connect these families together by attaching additional sources, finding death dates and maiden names, and merging duplicates. This community project was led by Mary Anderson who enlisted the efforts of hundreds of volunteers to do this additional work. One of the goals of the project was to help the African American community be able to see their family and learn more about their ancestors. This community effort has organized an African American family history event and organized lots of events to help people see their family on the Family Tree.

The Tacoma Community Project provides a good example of how the computer-assisted tools described in this paper can be used as a catalyst to invite and engage more individuals from a community to contribute to the Family Tree. This community engagement is crucial for tapping into the unique data sources and family memories that are held by individuals in the community.

Conclusion

The purpose of this paper is to describe an approach that can be used to ensure that everyone who was living in the United States in 1910 has a profile on the Family Tree. We document that the coverage rate is already at 87.8%. Ensuring that these profiles are all connected to their extended families will require a considerable amount of additional effort but would provide the type of interconnected network that will provide family history discovery experiences and linked family datasets for scientific studies.

This project provides a template for what is possible in other countries with public full-count census records near the year 1910. These include the censuses in Canada (1911), England

and Wales (1911), Denmark (1911), and France (1906). The census records for each of these countries are already on FamilySearch, and all of the analytics and tree growth approaches that we use in this paper can be applied to these other countries. This coverage of the full-country populations could provide a base layer to create longitudinal (and cross-country) datasets that could be used to study some of the major events in the world that occurred shortly after 1910, including the 1918 pandemic and World War I.

Table 1. Coverage rates by sample inclusion criteria

Sample	N	Attached	Hints	Coverage
Full population	93,638,754	62.84%	9.18%	72.02%
Nuclear families	78,009,397	71.67%	9.95%	81.62%
Families with 2+ people	75,656,955	73.44%	10.10%	83.54%
Families with children	67,698,852	78.16%	9.62%	87.78%

Notes: Each row in this table provides a different sample of the population of the US in the 1910 census, which each row being a subsample of the row above it. The column "attached" means that the individual in the census has a profile on the Family Tree and the record is attached to that profile. The column "hints" means that the person in the record is not attached to a profile but there is a profile that FamilySearch has identified as a likely match. the column "coverage" is just the sum of attached and hints and represents the fraction of people in the census that likely have a profile on the Family Tree.

Table 2. Record by match quality and relationship to household head.

A. Number of Record Hints

Relationship	5-star hints	4-star hints	3-star hints	no hints
Male head	1,842,681	449,719	253,008	1,101,052
Female head	287,925	55,939	48,509	245,531
Spouse	1,833,070	413,797	264,503	1,145,507
Son	3,707,309	353,852	224,603	2,798,399
Daughter	3,605,719	348,482	230,071	2,775,899
Total	11,276,704	1,621,789	1,020,694	8,066,388

B. Fraction of Record Hints in Each Quality Bin

| Relationship	5-star hints	4-star hints	3-star hints	no hints
Male head	50.53%	12.33%	6.94%	30.20%
Female head	45.14%	8.77%	7.60%	38.49%
Spouse	50.13%	11.32%	7.23%	31.32%
Son	52.33%	4.99%	3.17%	39.50%
Daughter	51.81%	5.01%	3.31%	39.88%
Total	51.29%	7.38%	4.64%	36.69%

Notes: We split the sample by gender + relationship to head. For each record, we used the record hint that has the highest match score, so the columns in this table are mutually exclusive. For panel B, the rows sum to 100%.

Figure 1. Growth in records for 1910 census attached to profiles on the Family Tree (by race)
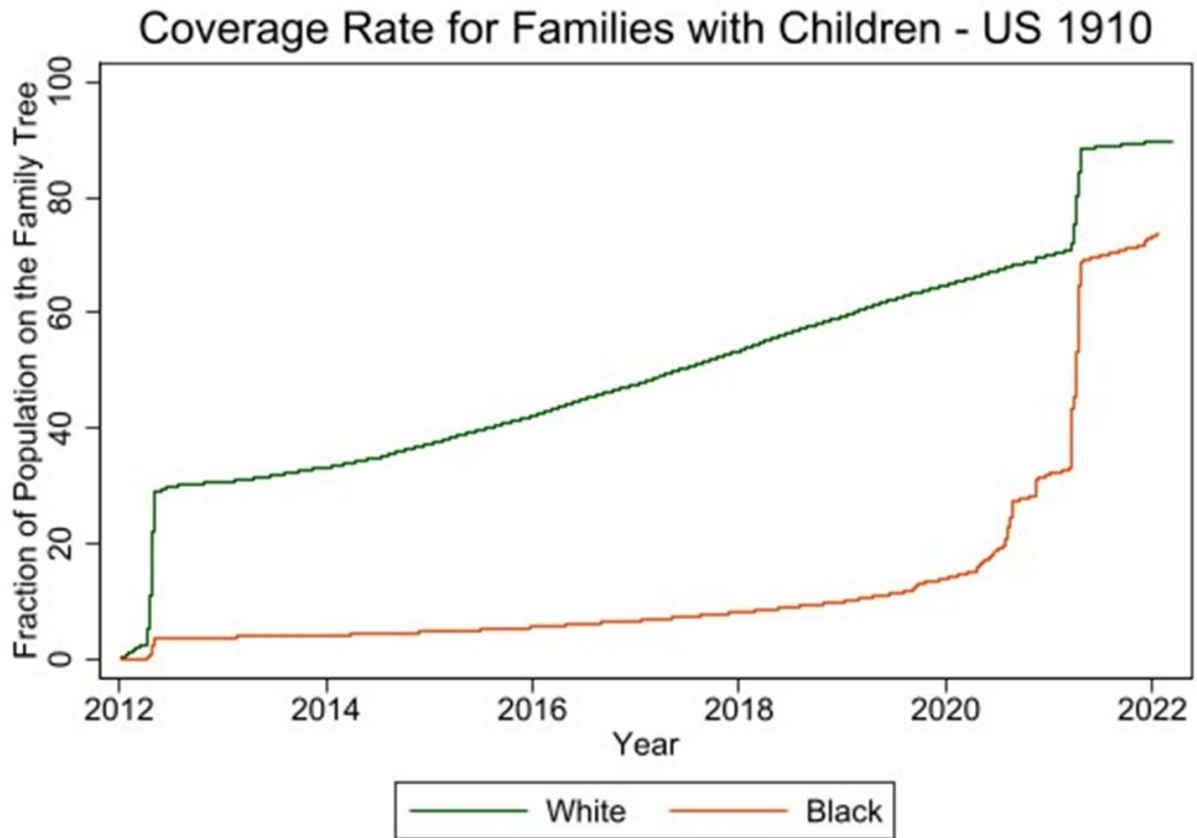
Notes: This figure is based on a random sample of 100,000 individuals from the 1910 census who were part of a family with children. We identified which of these individuals are on the Family Tree and attached to the census. We then used the contribution history to identify the year that the profile for these individuals was first created.

Figure 2. Growth in Family Tree coverage for Pierce County, WA 1910 census
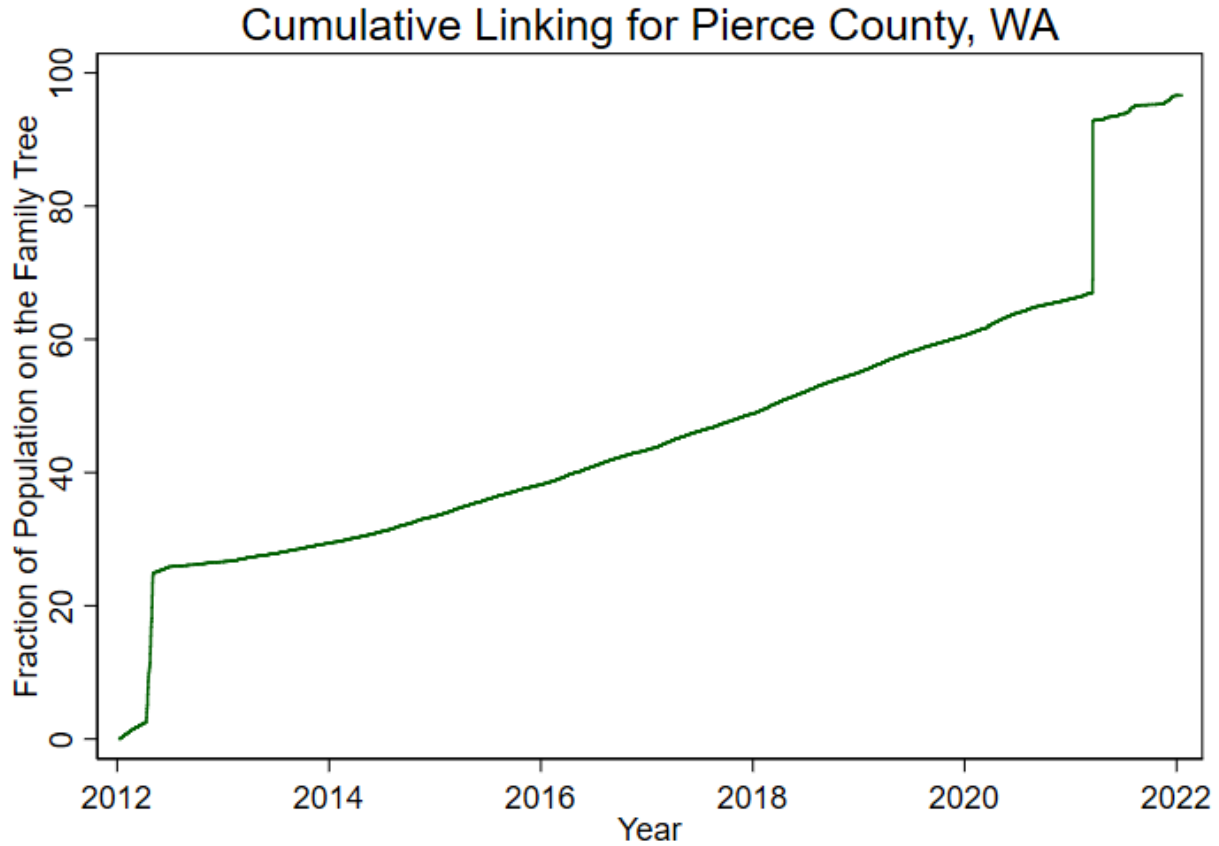
Notes: This figure is based on all individuals in the 1910 census who were living in Pierce County, Washington and are part of a family with children. We identified which of these individuals are on the Family Tree and attached to the census. We then used the contribution history to identify the year that the profile for these individuals was first created.