Creation of an Evaluation Paradigm for “RecordMatch” and its Application to GenMergeDB Clustering Results

Patrick Schone (patrickjohn.schone@ldschurch.org)

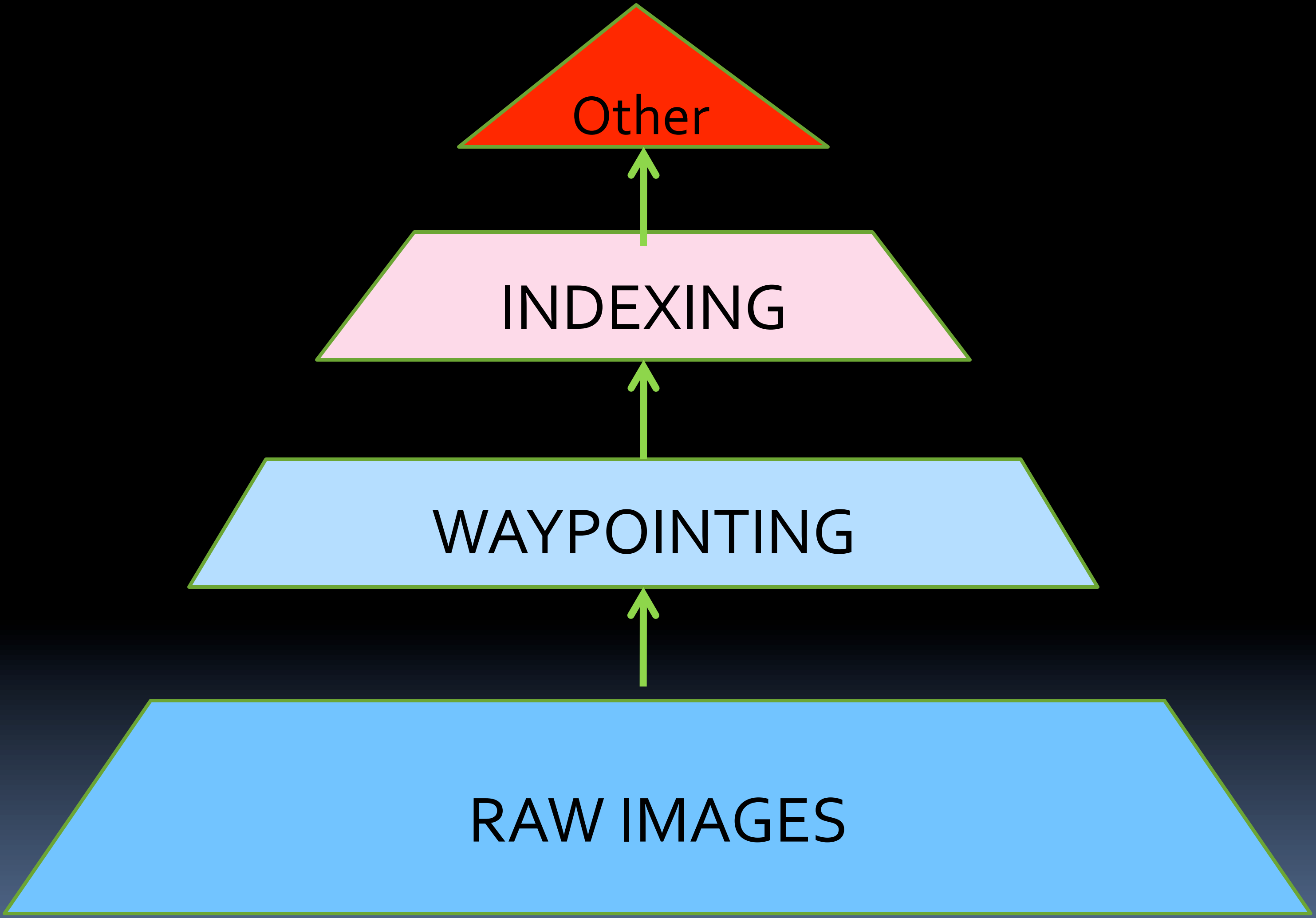
11 February 2011

OUTLINE

- BACKGROUND ON “RECORD MATCH”
- STRATEGY FOR EVALUATION
- SCORING & RESULTS
- CONCLUSIONS

BACKGROUND ON “RECORD MATCH”

Data:Raw+Annotation



Other

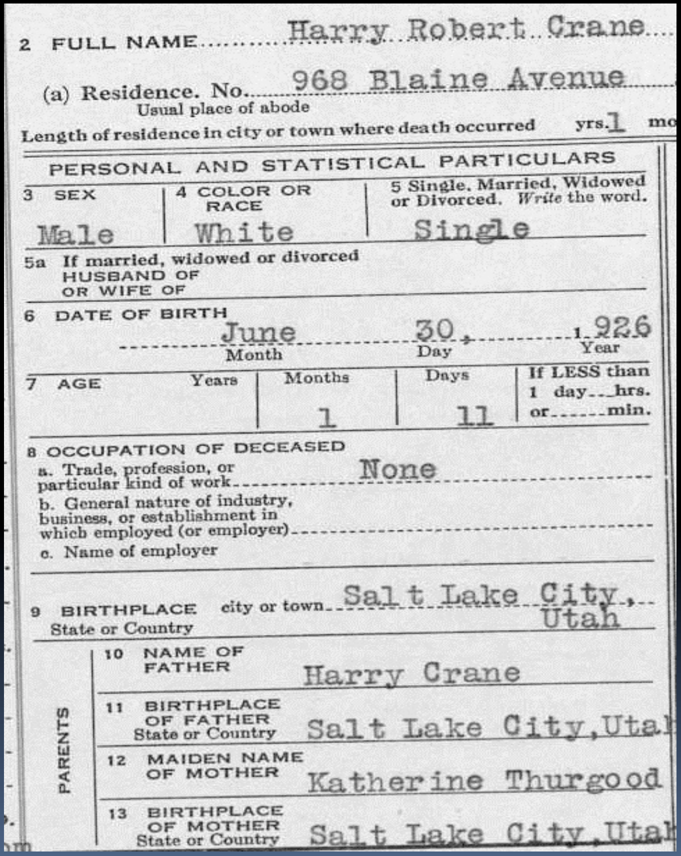
INDEXING

WAYPOINTING

RAW IMAGES

An “Indexed” Record

“Indexing” or “Extracting” renders the information in a record as text.



2 FULL NAME Harry Robert Crane

(a) Residence. No. 968 Blaine Avenue
Usual place of abode

Length of residence in city or town where death occurred yrs. 1 mo

PERSONAL AND STATISTICAL PARTICULARS

3 SEX	4 COLOR OR RACE	5 Single, Married, Widowed or Divorced. *Write* the word.
Male	White	Single

5a If married, widowed or divorced HUSBAND OF OR WIFE OF

6 DATE OF BIRTH June 30, 1926
Month Day Year

7 AGE	Years	Months	Days	If LESS than 1 day hrs. or min.
		1	11	

8 OCCUPATION OF DECEASED

a. Trade, profession, or particular kind of work None

b. General nature of industry, business, or establishment in which employed (or employer)

c. Name of employer

9 BIRTHPLACE city or town Salt Lake City, Utah
State or Country

PARENTS

10 NAME OF FATHER Harry Crane

11 BIRTHPLACE OF FATHER State or Country Salt Lake City,Uta

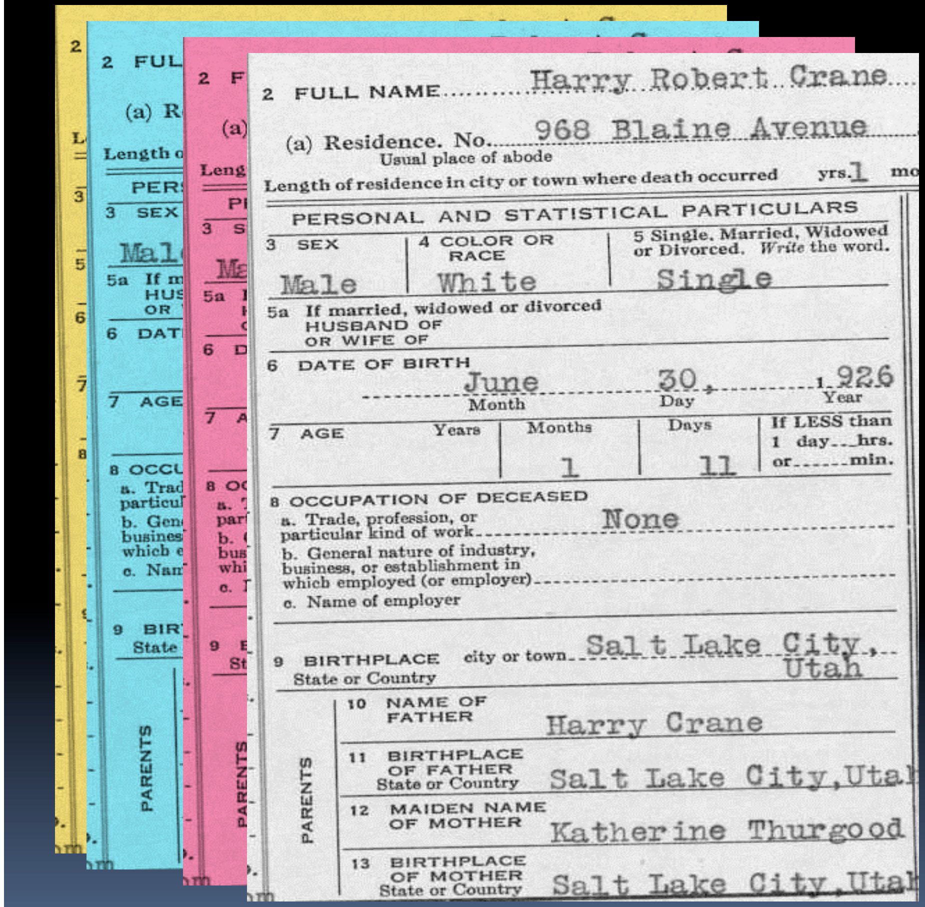
12 MAIDEN NAME OF MOTHER Katherine Thurgood

13 BIRTHPLACE OF MOTHER State or Country Salt Lake City,Uta

Full Name: Harry Robert Crane
(a) Residence No. 968 Blaine Avenue
Sex: Male, …
Date of Birth: June 30, 1926
Age: 1m 11d
Birthplace: Salt Lake City, Utah
Name of Father: Harry Crane
Bplace of Father: Salt Lake City, Utah
Mother’s Maiden: Katherine Thurgood
Bplace of Mother: Salt Lake City, Utah

Searchable Collection

Once a file is indexed/extracted, the information becomes searchable



2 FULL NAME Harry Robert Crane

(a) Residence. No. 968 Blaine Avenue
Usual place of abode

Length of residence in city or town where death occurred yrs. 1 mo

PERSONAL AND STATISTICAL PARTICULARS

3 SEX	4 COLOR OR RACE	5 Single, Married, Widowed or Divorced. *Write* the word.
Male	White	Single

5a If married, widowed or divorced HUSBAND OF OR WIFE OF

6 DATE OF BIRTH June 30, 1926
Month Day Year

7 AGE	Years	Months	Days	If LESS than 1 day...hrs. or......min.
		1	11	

8 OCCUPATION OF DECEASED

a. Trade, profession, or particular kind of work None

b. General nature of industry, business, or establishment in which employed (or employer)

c. Name of employer

9 BIRTHPLACE city or town Salt Lake City,
State or Country Utah

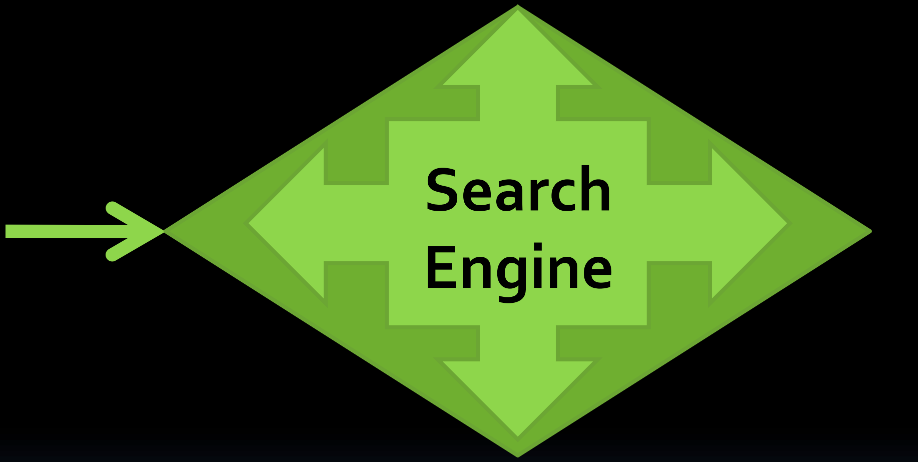
PARENTS

10 NAME OF FATHER Harry Crane

11 BIRTHPLACE OF FATHER State or Country Salt Lake City,Uta

12 MAIDEN NAME OF MOTHER Katherine Thurgood

13 BIRTHPLACE OF MOTHER State or Country Salt Lake City,Uta

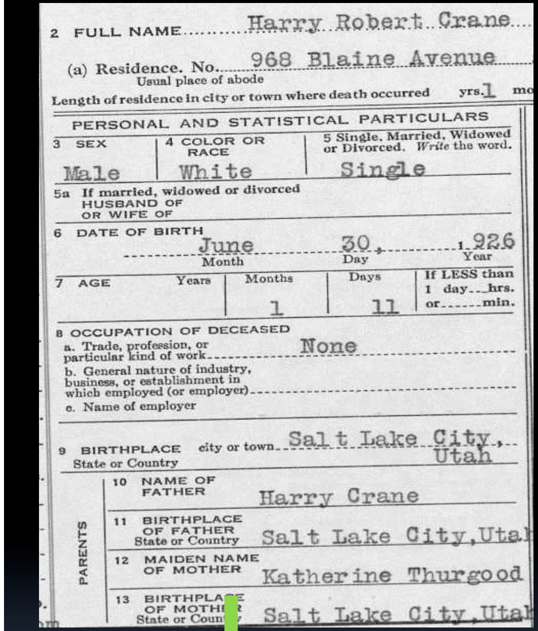


Search Engine

Searching for Ancestors

FAMILYSEARCH
WHERE GENERATIONS MEET

Patrons can current use www.familysearch.org to search for their ancestors in historical records.



2 FULL NAME Harry Robert Crane

(a) Residence. No. 968 Blaine Avenue
Usual place of abode

Length of residence in city or town where death occurred yrs. 1 mo

PERSONAL AND STATISTICAL PARTICULARS

3 SEX Male

4 COLOR OR RACE White

5 Single, Married, Widowed or Divorced. *Write* the word. Single

5a If married, widowed or divorced HUSBAND OF OR WIFE OF

6 DATE OF BIRTH June (Month) 30, (Day) 1926 (Year)

7 AGE Years Months 1 Days 11 If LESS than 1 day ___ hrs. or ______ min.

8 OCCUPATION OF DECEASED
a. Trade, profession, or particular kind of work None
b. General nature of industry, business, or establishment in which employed (or employer)
c. Name of employer

9 BIRTHPLACE city or town Salt Lake City, State or Country Utah

PARENTS

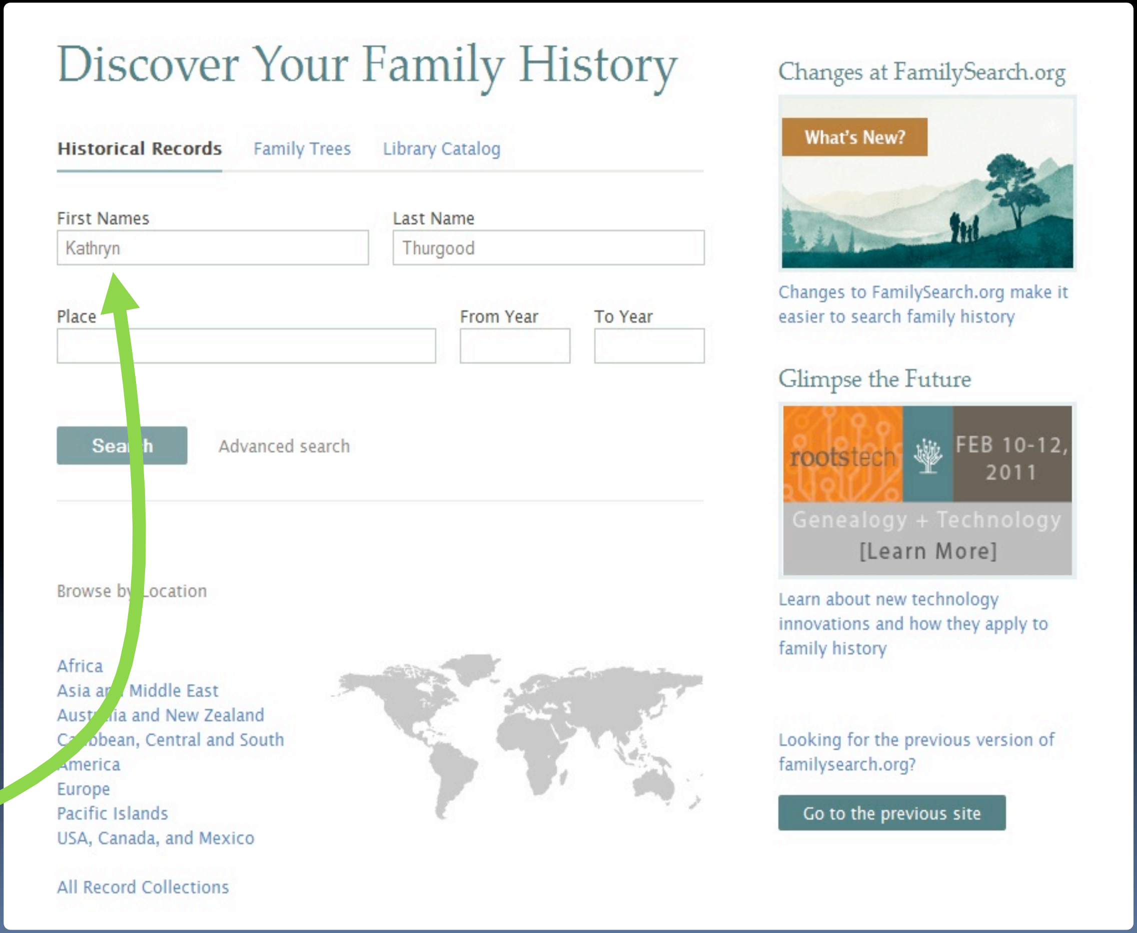
10 NAME OF FATHER Harry Crane

11 BIRTHPLACE OF FATHER State or Country Salt Lake City,Utah

12 MAIDEN NAME OF MOTHER Katherine Thurgood

13 BIRTHPLACE OF MOTHER State or Country Salt Lake City,Utah

Mother's Maiden:
Katherine
Thurgood



Discover Your Family History

Historical Records | Family Trees | Library Catalog

First Names: Kathryn

Last Name: Thurgood

Place

From Year

To Year

Search   Advanced search

Browse by Location

- Africa
- Asia and Middle East
- Australia and New Zealand
- Caribbean, Central and South America
- Europe
- Pacific Islands
- USA, Canada, and Mexico

All Record Collections

Changes at FamilySearch.org

What's New?

Changes to FamilySearch.org make it easier to search family history

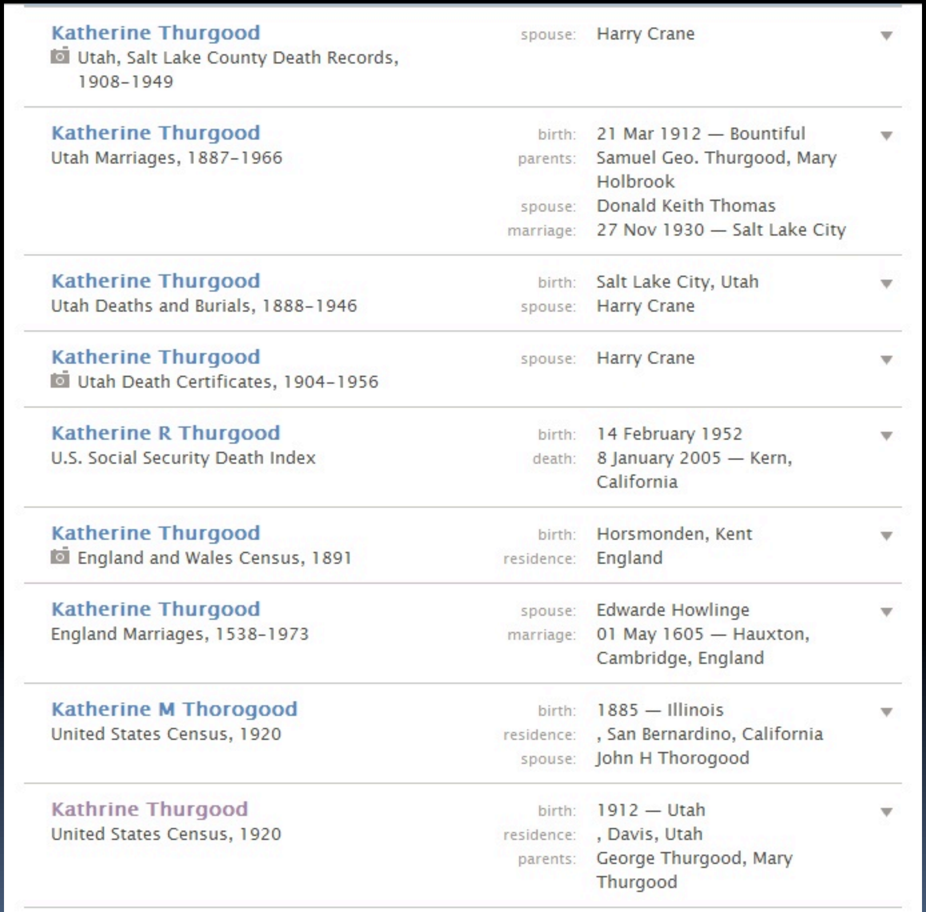
Glimpse the Future

Learn about new technology innovations and how they apply to family history

Looking for the previous version of familysearch.org?

Go to the previous site

Search vs. Person-Matching



Katherine Thurgood Utah, Salt Lake County Death Records, 1908-1949	spouse:	Harry Crane
Katherine Thurgood Utah Marriages, 1887-1966	birth: parents: spouse: marriage:	21 Mar 1912 — Bountiful Samuel Geo. Thurgood, Mary Holbrook Donald Keith Thomas 27 Nov 1930 — Salt Lake City
Katherine Thurgood Utah Deaths and Burials, 1888-1946	birth: spouse:	Salt Lake City, Utah Harry Crane
Katherine Thurgood Utah Death Certificates, 1904-1956	spouse:	Harry Crane
Katherine R Thurgood U.S. Social Security Death Index	birth: death:	14 February 1952 8 January 2005 — Kern, California
Katherine Thurgood England and Wales Census, 1891	birth: residence:	Horsmonden, Kent England
Katherine Thurgood England Marriages, 1538-1973	spouse: marriage:	Edwarde Howlinge 01 May 1605 — Hauxton, Cambridge, England
Katherine M Thorogood United States Census, 1920	birth: residence: spouse:	1885 — Illinois , San Bernardino, California John H Thorogood
Kathrine Thurgood United States Census, 1920	birth: residence: parents:	1912 — Utah , Davis, Utah George Thurgood, Mary Thurgood

Issues of Straight Search:

A search tries to satisfy the elements of the user's query. It is not trying to find all instances of a particular person. So questions arise ....

[1] Are these all MY "Katherine Thurgood" or are there multiple people here?

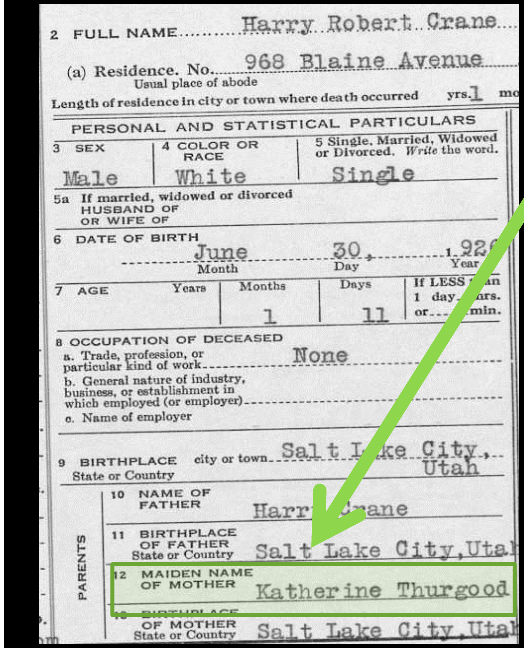
[2] Are there records that are about my "Katherine Thurgood" but this list hasn't provided me?

Defining “Records Match”

“Records Match” is an experimental process whose hypothesis is:

Supposing a person seeks records for an individual X,
it <u>should</u> be possible to automatically provide all records that refer to X.

Example: “Records Match”



2 FULL NAME Harry Robert Crane

(a) Residence. No. 968 Blaine Avenue
Usual place of abode

Length of residence in city or town where death occurred yrs. 1 mo

PERSONAL AND STATISTICAL PARTICULARS

3 SEX Male | 4 COLOR OR RACE White | 5 Single, Married, Widowed or Divorced. *Write* the word. Single

5a If married, widowed or divorced HUSBAND OF OR WIFE OF

6 DATE OF BIRTH June (Month) 30, (Day) 1926 (Year)

7 AGE Years | Months 1 | Days 11 | If LESS than 1 day hrs. or min.

8 OCCUPATION OF DECEASED
a. Trade, profession, or particular kind of work None
b. General nature of industry, business, or establishment in which employed (or employer)
c. Name of employer

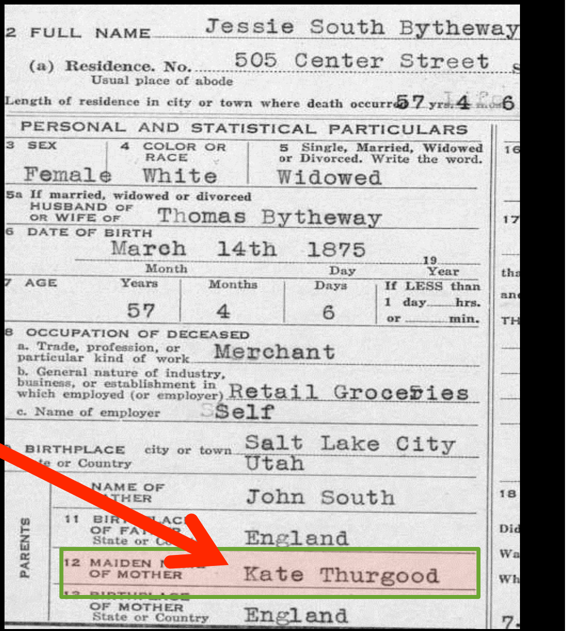
9 BIRTHPLACE city or town State or Country Salt Lake City, Utah

PARENTS
10 NAME OF FATHER Harry Crane
11 BIRTHPLACE OF FATHER State or Country Salt Lake City, Utah
12 MAIDEN NAME OF MOTHER Katherine Thurgood
13 BIRTHPLACE OF MOTHER State or Country Salt Lake City, Utah

“Katherine Thurgood” of 1926 SLC Death Record of Child,

IS NOT
Kate Thurgood of 1932 SLC Death Record of Husband

BUT IT IS
Kittie Thurgood of 1900 Census.



2 FULL NAME Jessie South Bytheway

(a) Residence. No. 505 Center Street
Usual place of abode

Length of residence in city or town where death occurred 57 yrs. 4 mos. 6

PERSONAL AND STATISTICAL PARTICULARS

3 SEX Female | 4 COLOR OR RACE White | 5 Single, Married, Widowed or Divorced. Write the word. Widowed

5a If married, widowed or divorced HUSBAND OF OR WIFE OF Thomas Bytheway

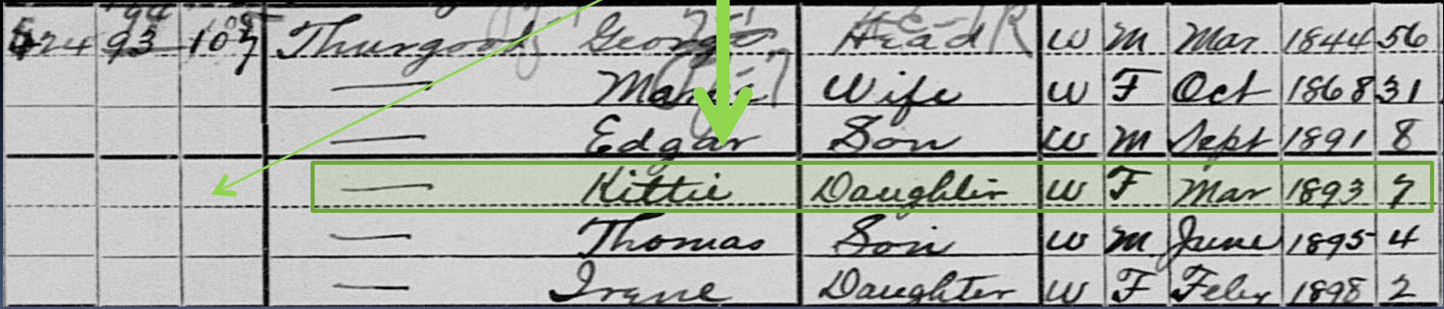
6 DATE OF BIRTH March 14th 1875 (Month, Day, 19 Year)

7 AGE Years 57 | Months 4 | Days 6 | If LESS than 1 day hrs. or min.

8 OCCUPATION OF DECEASED
a. Trade, profession, or particular kind of work Merchant
b. General nature of industry, business, or establishment in which employed (or employer) Retail Groceries
c. Name of employer Self

9 BIRTHPLACE city or town State or Country Salt Lake City Utah

PARENTS
10 NAME OF FATHER John South
11 BIRTHPLACE OF FATHER State or Country England
12 MAIDEN NAME OF MOTHER Kate Thurgood
13 BIRTHPLACE OF MOTHER State or Country England



524	93	107	Thurgood	George	Head	W	M	Mar	1844	56
			—	Marie	Wife	W	F	Oct	1868	31
			—	Edgar	Son	W	M	Sept	1891	8
			—	Kittie	Daughter	W	F	Mar	1893	7
			—	Thomas	Son	W	M	June	1895	4
			—	Irene	Daughter	W	F	Febr	1898	2

Has "Records Match" Been Done Before?

"Records Matching" appears new, though it has some similarities to a number of efforts around the world :

[1] Matching patron-submitted ancestries in New Family Search
[2] Entity disambiguation
[3] Knowledge Base Population
[4] Web document clustering
[5] Cross-document coreference analysis
[6] Merging of personal ancestral files
[7] Technologies like Ancestry.com's "Shaky Leaf"
…

STRATEGY FOR EVALUATION

Collection Focus: Utah

OVERLAPPING COLLECTIONS

FamilySearch has access to a number of Historical Records Collections from Utah (1850-1956). Overlapping collections is important to creating interesting evaluation results.

COLLECTIONS FOR THIS STUDY

- Six census collections (1850, 1860, 1870, 1880, 1900, 1920)
- Utah marriage records
- Utah death records through 1956
- Some Utah birth records
- Indian Affidavits from Utah
- Military records of Utah Veterans

Other “Entity” Evaluations

There have been a number of evaluations for entity disambiguation/linking which can be useful for comparison:

[1] Web People (Artiles, et al, 2008)

Cluster web pages that are about the same individual even though the names may be the same.

[2] ACE Cross-document Co-ref Task (Przybocky, et al, 2008)

For every individual mentioned in each of a collection of documents, link all of the same individuals across documents.

[3] TAC 2009/2010: Knowledge Base Population Tasks

For some number of seed entities, identify which knowledge base (KB) entry is referred to by the seed, and then discover any new information about that seed which is missing from the KB

Proceed With Seeds

POLICY

Consistent cross-document evaluation is VERY hard for humans. So it is critical to find a set of potential candidate people whose names or variants appear frequently enough to be interesting, but not so frequently that evaluation becomes impossible.

SEEDS

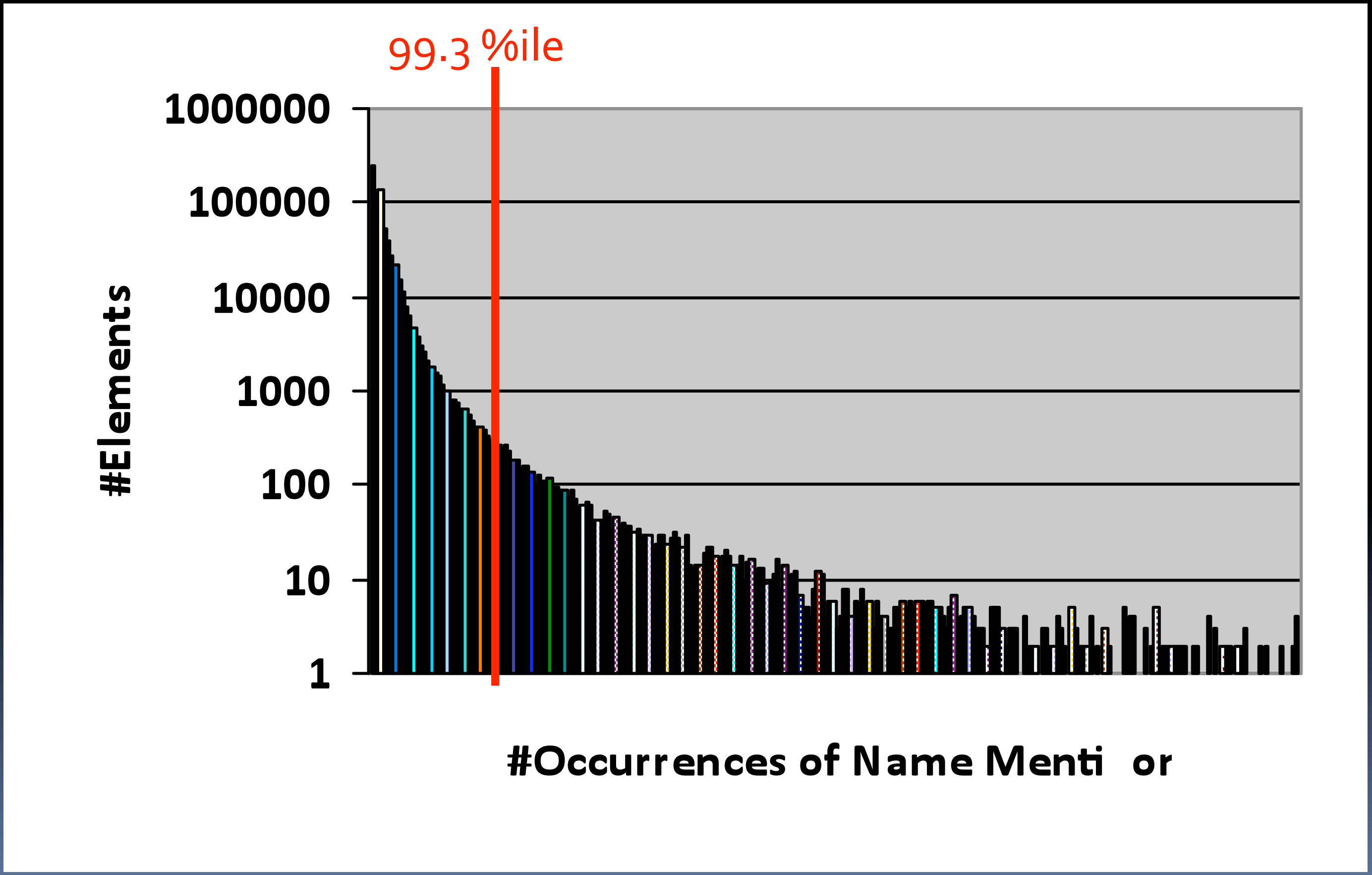
The semi-low-frequency names should be useful for identifying potential variants that may represent the same individual. This will be described in a moment.

ANALYSIS:

For each exact name string in the Utah collection, count the frequency of the name and use as seeds only names with frequencies of **20** or less.

Proceed with Seeds (2)

If we consider the frequency of any name string in the collection, those name strings that occur with 20 or fewer occurrences represent **99.3%** of all entries.



Proceed with Seeds (3)

Given a particular seed, S, our goal is to find all potential variants for that seed. We did this by:

[1] Index overlapping n-grams of all the names into a search engine (eg., Apache's SOLR).

[2] Convert the seed into overlapping n-grams and query the IR engine

[3] Review, by hand, the top 100 results to identify any variants that look like they could be feasible variants for the seed, S.

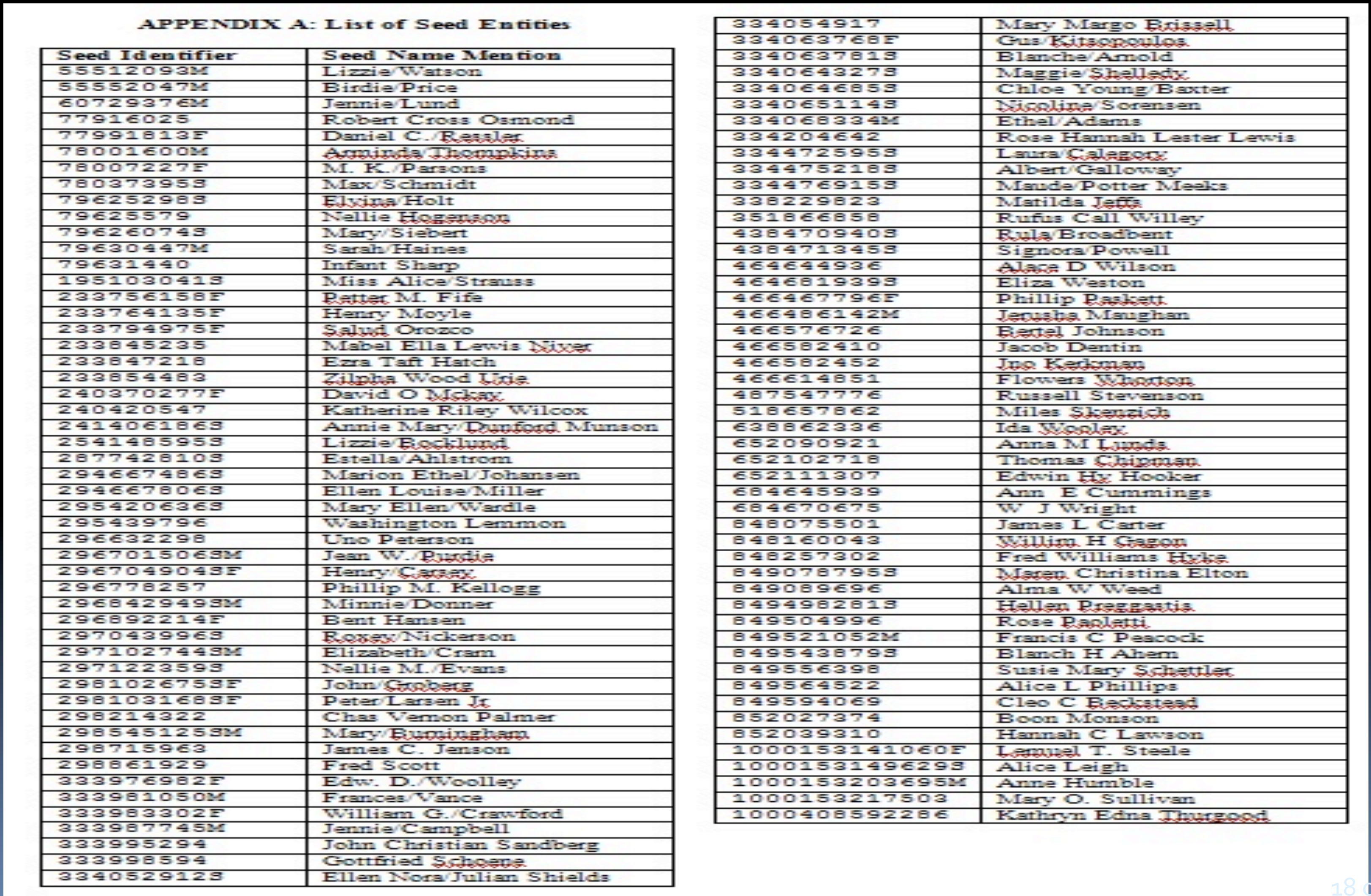
[4] Observe spouse names and, if it looks like S may have undergone a name change (usu. because of marriage or immigration), augment the results by search for each potential name change.

[5] Add, after the fact, any new names that are proposed by record match engines.

FAMILYSEARCH
WHERE GENERATIONS MEET

Proceed with Seeds (4)

We select 100 Seeds , generate the possibilities, and evaluated by hand. Approx 2200 mentions, and over 20000 pairs to consider.



APPENDIX A: List of Seed Entities

Seed Identifier	Seed Name Mention
55512093M	Lizzie/Watson
55552047M	Birdie/Price
60729376M	Jennie/Lund
77916025	Robert Cross Osmond
77991813F	Daniel C./Ressler
78001600M	Arminda/Thompkins
78007227F	M. K./Parsons
780373953	Max/Schmidt
796252983	Elvina/Holt
79625579	Nellie Hogenson
796260743	Mary/Siebert
79630447M	Sarah/Haines
79631440	Infant Sharp
1951030413	Miss Alice/Strauss
233756158F	Petter M. Fife
233764135F	Henry Moyle
233794975F	Salud Orozco
233845235	Mabel Ella Lewis Nixer
233847218	Ezra Taft Hatch
233854483	Zilpha Wood Urie
240370277F	David O Mckay
240420547	Katherine Riley Wilcox
2414061863	Annie Mary/Dunford Munson
2541485953	Lizzie/Rocklund
2877428103	Estella/Ahlstrom
2946674863	Marion Ethel/Johansen
2946678063	Ellen Louise/Miller
2954206363	Mary Ellen/Wardle
295439796	Washington Lemmon
296632298	Uno Peterson
2967015063M	Jean W./Purdie
2967049043F	Henry/Carsey
296778257	Phillip M. Kellogg
2968429493M	Minnie/Donner
296892214F	Bent Hansen
2970439963	Roxey/Nickerson
2971027443M	Elizabeth/Cram
2971223593	Nellie M./Evans
2981026753F	John/Groberg
2981031683F	Peter/Larsen Jr
298214322	Chas Vernon Palmer
2985451253M	Mary/Burningham
298715963	James C. Jenson
298861929	Fred Scott
333976982F	Edw. D./Woolley
333981050M	Frances/Vance
333983302F	William G./Crawford
333987745M	Jennie/Campbell
333995294	John Christian Sandberg
333998594	Gottfried Schoene
3340529123	Ellen Nora/Julian Shields
334054917	Mary Margo Brissell
334063768F	Gus/Kitsopoulos
3340637813	Blanche/Arnold
3340643273	Maggie/Shelledy
3340646853	Chloe Young/Baxter
3340651143	Nicoline/Sorensen
334068334M	Ethel/Adams
334204642	Rose Hannah Lester Lewis
3344725953	Laura/Calegory
3344752183	Albert/Galloway
3344769153	Maude/Potter Meeks
338229823	Matilda Jeffs
351866858	Rufus Call Willey
4384709403	Rula/Broadbent
4384713453	Signora/Powell
464644936	Alace D Wilson
4646819393	Eliza Weston
466467796F	Phillip Paskett
466486142M	Jerusha Maughan
466576726	Bertel Johnson
466582410	Jacob Dentin
466582452	Jno Kerkman
466614851	Flowers Whorton
487547776	Russell Stevenson
518657862	Miles Skenzich
638862336	Ida Wooley
652090921	Anna M Lunds
652102718	Thomas Chipman
652111307	Edwin Hy Hooker
684645939	Ann E Cummings
684670675	W J Wright
848075501	James L Carter
848160043	Willim H Gagon
848257302	Fred Williams Hyke
8490787953	Maren Christina Elton
849089696	Alma W Weed
8494982813	Hellen Preggastis
849504996	Rose Paoletti
849521052M	Francis C Peacock
8495438793	Blanch H Ahern
849556398	Susie Mary Schettler
849564522	Alice L Phillips
849594069	Cleo C Beckstead
852027374	Boon Monson
852039310	Hannah C Lawson
10001531410060F	Lamuel T. Steele
1000153149629 3	Alice Leigh
100015320369 5M	Anne Humble
1000153217503	Mary O. Sullivan
1000408592286	Kathryn Edna Thurgood

Proceed with Seeds (5)

Example: From Seed to Selection Set

SEED:
Kathryn Edna
Thurgood,
1000408592286

Kathryn Edna/Thurgood	295640551-S
Kathryn Edna Thurgood	1000408592286
KATHRYN EDNA THURGOOD	1000408587481-S
KATHRYN EDNA THURGOOD	1000408592284-M
KATHRYN EDNA THURGOOD	1000408592285-M
Kathryn Edna Thurgood	1000408592286
KATHRYN EDNA THURGOOD	1000408592290-M
KATHRYN EDNA THURGOOD	1000408587481-S
Kathryn/Thurgood	77986494-M
Kathryn/Thurgood	455454993-M
Kathryn/Thurgood	455454993-M
Kathryn Thurgood	233750764-M
Kathryn Thurgood	233799133-M
Katheryne Edna/Thurgood	117156205-S
Edna/Thurgood	296654466-S
Edna Thurgood	1000408595047
EDNA THURGOOD	1000408585337-S
Edna Thurgood	848167146
Edna Thurgood	849569781
Edna Thurgood	1000408595047
EDNA THURGOOD	1000408585337-S

Kathrine Thurgood	848166351
Kate/Thurgood	79625714-M
Kate/Thurgood	420048297-M
Kate/Thurgood	420048297-M
Kate Thurgood	233775363-M
Kate Thurgood	351869678-M
Katherine/Thurgood	77985422-M
Katherine/Thurgood	296997578-S
Katherine/Thurgood	453630315-M
Katherine/Thurgood	453630315-M
Katherine Thurgood	233747307-M
Catherine Thurgood	241204791
Kittie Thurgood	466606627
Kathryn Crane	849042896-S
Kathryn Crane	849042897
Kathryn Crane	849042898-M
Kathryn Crane	849042898
Kathryn Crane	849042899-M
Kathryn Inez Crane	77986494
Kathryn Inez Crane	455454993
Kathryn Inez Crane	455454993
Kathryn Inez Crane	233750764
Kathryn Inez Crane	1000408592285
Kathryn Craner	233840170
Kathryn E./Crehan	287378443-S

Annotate Truth Set

Each entry of the expanded table constitutes an instance of a record that contains information. We morph the records data to center around the seed individual & we align the information from each instance in order to vet the results.

Name	ID	Bdate	Bplace	Mdate	Mplace	Ddate	Sname	Fname	Mname	Cname	CBD	CBP	CDD
Kathryn Edna/ Thurgood	295640551-S			22-Jul-14	Salt Lake Co., Utah	--	Harry Niles/ Crane			--	--	--	--
Kathryn Edna Thurgood	1000408592 286	08 MAR 1893	SALT LAKE CITY,SALT LAKE,UTAH	--	--	7-May-81	--	GEORGE THURGOOD	MARIA STECK	--	--	--	--
KATHRYN EDNA THURGOOD	1000408587 481-S			22-Jul-14	Salt Lake City,Salt Lake,Utah	--	HARRY NILES CRANE			--	--	--	--
KATHRYN EDNA THURGOOD	1000408592 284-M	04 APR 1888+/-10		--	--	--	HARRY NILES CRANE	--	--	BETTY LOUISE CRANE	4-Apr-20	SALT LAKE CITY,SALT LAKE,UTAH	6-Jul-40
KATHRYN EDNA THURGOOD	1000408592 285-M	31 OCT 1883+/-10		--	--	--	HARRY NILES CRANE	--	--	KATHRYN INEZ CRANE	31-Oct-15	SALT LAKE CITY,SALT LAKE,UTAH	11-Jan-27
Kathryn Edna Thurgood	1000408592 286	08 MAR 1893	SALT LAKE CITY,SALT LAKE,UTAH	--	--	7-May-81	--	GEORGE THURGOOD	MARIA STECK	--	--	--	--
KATHRYN EDNA THURGOOD	1000408592 290-M	30 JUN 1894+/-10		--	--	--	HARRY NILES CRANE	--	--	ROBERT HARRY CRANE	30-Jun-26	SALT LAKE CITY,SALT LAKE,UTAH	11-Aug-26
KATHRYN EDNA THURGOOD	1000408587 481-S			22-Jul-14	Salt Lake City,Salt Lake,Utah	--	HARRY NILES CRANE			--	--	--	--
..	..	..	..	..	..	..	..	..					
Kittie Thurgood	466606627	Mar 1893	Utah	--	--	--		George Thurgood	Marie Thurgood	--	--	--	--

Annotate Truth Set (2)

These sets then get broken up into *clusters* based on the dates/places/family relations generated from each line of data.

Here, 46 entries (1035 decisions) cluster into 10 GROUPS

SEED:
Kathryn Edna Thurgood,
1000408592286

Kathryn Edna/Thurgood	295640551-S
Kathryn Edna Thurgood	1000408592286
KATHRYN EDNA THURGOOD	1000408587481-S
KATHRYN EDNA THURGOOD	1000408592284-M
KATHRYN EDNA THURGOOD	1000408592285-M
Kathryn Edna Thurgood	1000408592286
KATHRYN EDNA THURGOOD	1000408592290-M
KATHRYN EDNA THURGOOD	1000408587481-S
Kathryn/Thurgood	77986494-M
Kathryn/Thurgood	455454993-M
Kathryn/Thurgood	455454993-M
Kathryn Thurgood	233750764-M
Kathryn Thurgood	233799133-M
Katheryne Edna/Thurgood	117156205-S
Edna/Thurgood	296654466-S
Edna Thurgood	1000408595047
EDNA THURGOOD	1000408585337-S
Edna Thurgood	848167146
Edna Thurgood	849569781
Edna Thurgood	1000408595047
EDNA THURGOOD	1000408585337-S

Kathrine Thurgood	848166351
Kate/Thurgood	79625714-M
Kate/Thurgood	420048297-M
Kate/Thurgood	420048297-M
Kate Thurgood	233775363-M
Kate Thurgood	351869678-M
Katherine/Thurgood	77985422-M
Katherine/Thurgood	296997578-S
Katherine/Thurgood	453630315-M
Katherine/Thurgood	453630315-M
Katherine Thurgood	233747307-M
Catherine Thurgood	241204791
Kittie Thurgood	466606627
Kathryn Crane	849042896-S
Kathryn Crane	849042897
Kathryn Crane	849042898-M
Kathryn Crane	849042898
Kathryn Crane	849042899-M
Kathryn Inez Crane	77986494
Kathryn Inez Crane	455454993
Kathryn Inez Crane	455454993
Kathryn Inez Crane	233750764
Kathryn Inez Crane	1000408592285
Kathryn Craner	233840170
Kathryn E./Crehan	287378443-S

SCORING & RESULTS

Scoring

PRECISION:

If the system reports S items and R of them are correct,

Precision=R/S

RECALL:

If the system reports R correct items, but it should have found F of them,

Recall=R/F

Scoring (2)

We evaluate using B-Cubed Method

Ta is the truth set for element *a*, that ***Ha*** is the hypothesis for element *a*

P= |***Ta*** ∩ ***Ha*** |/| ***Ha***|

R= |***Ta*** ∩ ***Ha*** |/| ***Ta*** |

T: {A1,A2,A3,A4}, {B1,B2,B3,B4}, {C1,C2,C3}

H: {A1,A2,B1}, {A3,C1,C2}, {A4}, {B2,B3,B4}, {C3,UV1,UV2}

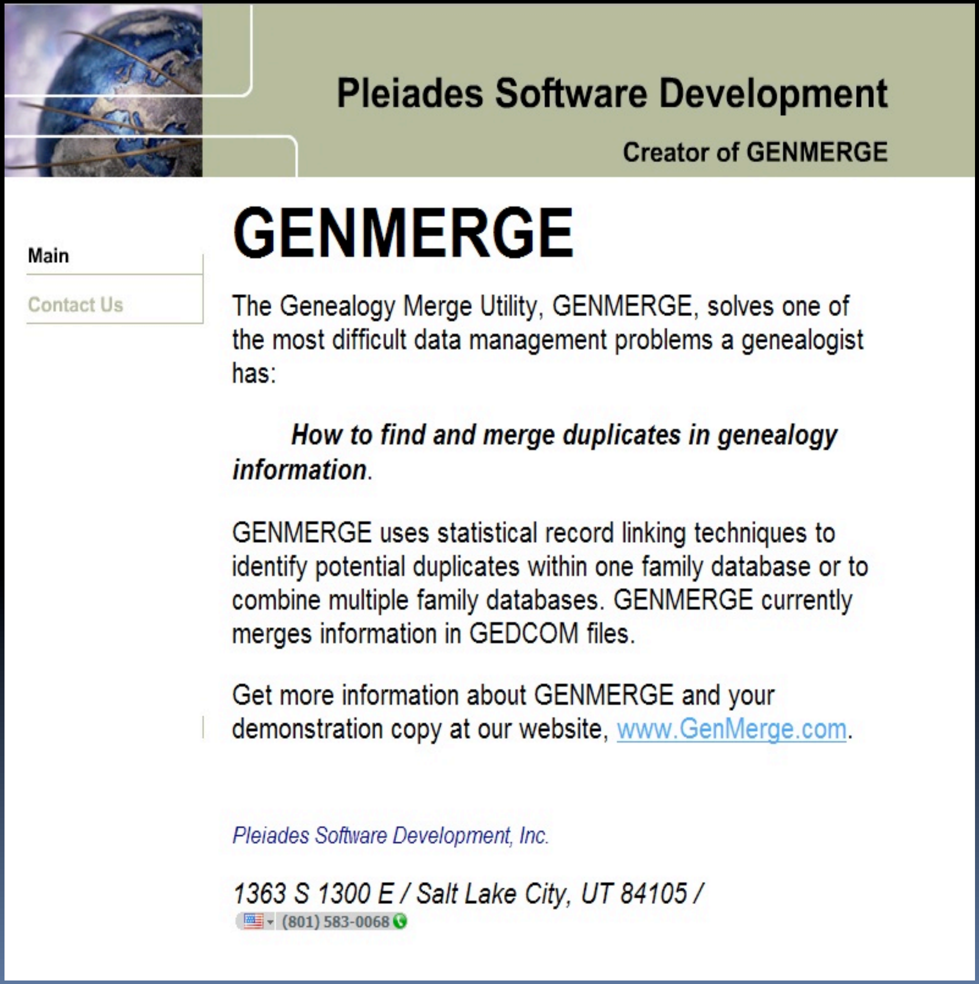
Entry	Prec.	Recall	Entry	Prec.	Recall
A1	2/3	2/4	B3	3/3	3/4
A2	2/3	2/4	B4	3/3	3/4
A3	1/3	1/4	C1	2/3	2/3
A4	1/1	1/4	C2	2/3	2/3
B1	1/3	1/4	C3	1/1	1/3
B2	3/3	3/4	(Unvetted/UV: deleted)		

25/33 and the average recall would have been 17/33

New Tools & Need For Evaluations

GenMergeDB is a product of Pleiades Software Development, Inc. which suggests that is can be applied to Record Match.

Question: Can we create an evaluation and see how well this, and other technologies, work on Records Matching?



Pleiades Software Development

Creator of GENMERGE

GENMERGE

The Genealogy Merge Utility, GENMERGE, solves one of the most difficult data management problems a genealogist has:

How to find and merge duplicates in genealogy information.

GENMERGE uses statistical record linking techniques to identify potential duplicates within one family database or to combine multiple family databases. GENMERGE currently merges information in GEDCOM files.

Get more information about GENMERGE and your demonstration copy at our website, www.GenMerge.com.

Pleiades Software Development, Inc.

1363 S 1300 E / Salt Lake City, UT 84105 /
(801) 583-0068

On Scoring GenMergeDB

GenMergeDB attempts to link together any reasonable entries from within the whole collection – for the purpose of family reconstitution.

For this eval, we only consider those names that they propose that intersect with the evaluation collection (though we add to the eval any names that they propose to be associated with the seed individual).

Comparison Scores

* We need to compare to a baseline. How well does GenMerge work versus doing NOTHING…. No Merging.

We call this the SHATTER ALL BASELINE.

* We also need to include in the scoring process a comparison as to whether we evaluate SEED entities only, or all SEED-Derived entities.

GenMerge: Seed Clusters Only

GEN MERGE				SHATTER-ALL			
Ave P	Ave R	#Clusters	#Mentions	Ave P	Ave R	#Clusters	#Mentions
0.989	**0.690**	201	712	1.000	0.143	712	712

If we only look at performance of seed-related clusters, we see that GenMerge has very high precision and reasonable recall (which is about five times better than Shatter-All).

GenMerge: All Clusters

GEN MERGE				SHATTER-ALL			
Ave P	Ave R	#Clusters	#Mentions	Ave P	Ave R	#Clusters	#Mentions
0.980	0.806	1049	2197	1.000	*0.382	2197	2197

If we only look at performance of all clusters, we see that GenMerge still has very high precision and great recall.

**The fact that Shatter All has better recall here than with Seed-only clusters suggests that there are associates of non-seed entries which we have not included in scoring but are probably present in the data.*

POTENTIAL FOLLOW-ON & CONCLUSIONS

Conclusions

* Clear that GenMerge software is doing well. The LDS Church's FHD/Family Reconstitution team has taken these results and disseminated GenMergeDB's entire Utah output.

* There are still records that could be automatically found, so that will require some additional study

* Right now, testing FamilySearch's Common Pedigree.

* The study was specific to Utah because of strong overlapping collections. It would be beneficial to revisit this task and look at all US records or all records.

* We are undertaking an extensive evaluation that might help with this task. Evaluation may also extend the number of results available for this Utah collection.