The questionnaire comprises three sections: **Section A** includes the questions related to the most general and important information identifying the content, scope and provenance of the databases and the information about their creators. **Section B** contains more specific and detailed questions about databases, such as the period(s) of observation, sampling design and procedures, data collection, linkage process and others. **Section C** contains detailed questions about sources used for the databases: their type, scope, content, state of preservation, etc.

## Section A

### 1. Title of the database

| **Utah Population Database** |

#### 1.a. Subtitle, which brings meaning to the title (scope, place, time period):


#### 2. Abbreviation

| **UPDB** |

### 3. Links to website(s):

| 3.a. Homepage | **http://healthcare.utah.edu/huntsmancancerinstitute/research/updb** |

| 3.b. Get to data | **https://hci-updblapp.hci.utah.edu/updbl** |

- this is a query tool for the UPDB

### 4. Abstract: describes content of the database. Max. length: 300 words

Please indicate:

- Scope and main goal
- Time and territory covered by data
- Sample strategy
- Main sources

| The Utah Population Database (UPDB) at the University of Utah is one of the world’s richest sources of in-depth information that supports research on genetics, epidemiology, demography, and public health. For more than 40 years, researchers at the University of Utah and other institutions have used UPDB to identify and study individuals and families that have higher than normal incidence of cancer or other diseases, to analyse... |
patterns of genetic inheritance, and to identify specific genetic mutations. In addition, demographic studies have shown trends in fertility transition and changes in mortality patterns for both infants and adults. The UPDB is the only database of its kind in the United States and one of few such resources in the world. The central component of the UPDB is an extensive set of Utah family histories, in which family members are linked to demographic and medical information. **Original goal**

The Utah Population Database (UPDB) is a rich source of information for genetic, epidemiological, demographic and public health studies. For over 40 years, researchers have used this resource to identify and study families that have higher than normal incidence of cancer or other diseases, to analyze patterns of genetic inheritance and to identify specific genetic mutations. Demographic studies have observed and analyzed the trends in the fertility transition and changes in mortality patterns for both infants and adults. The UPDB is the only such database of its kind in the US and one of few such resources in the world. The central component of UPDB is an extensive set of Utah family histories, in which family members are linked to demographic and medical information. The UPDB also includes diagnostic records on cancer, cause of death, and medical details associated with births. The UPDB provides access to almost 8 million records and supports over 160 research projects. These data can only be used for biomedical and health related research.

**Sample strategy**

Genealogies: Over 185,000 Utah families were identified on "Family Group Sheets" from the archives at the Utah Family History Library, which is maintained by the LDS Church. These sheets contain demographic and kinship information on three generations. For a family to be initially selected for UPDB from the genealogies, at least one member had to have a vital event (birth, marriage, death) on the Mormon Pioneer Trail or in Utah. Some family members have birth dates as early as 1740; most occur after 1780; the genealogy set ends with birth dates about 1970. Other records: complete statewide records with no sampling; the fields that are available vary over time and fields that are computerized vary over time.

**Time and territory covered by data**

Utah for the most part. Genealogies include migrants into Utah; early migrants originated in New England, Missouri, Illinois and Northern Europe. The families tended to settle in Utah and southern Idaho. 1790 - present

**Main sources**

The UPDB includes diagnostic records about cancer,
cause of death, and medical details associated with births. It also includes claims data from state-wide inpatient hospital discharge records as well as ambulatory surgery records from hospital outpatient departments and ambulatory surgery centers. The UPDB provides access to information on more than 8 million individuals and supports over 160 research projects. This information can only be used for biomedical and health-related research; the privacy of individuals represented in these records and confidentiality of the data is strictly protected.

5. **Keywords:**
Please use the recommended keywords if they are applicable: demography, life course, census, church register, civil certificates, population register, history, social science, genetics, migration, occupations.

Please add your own keywords, if you have data not covered by the recommended terms.

Demography, life course, census, genetics, vital records

genealogy

6. **Citation:** Indicate how you want others to cite your database.

Utah Population Database

7. **IDS compatible:** Indicate with Yes or No whether the database is IDC compatible, if Yes, please specify.

No

8. **Has the database already been completed or it is still under construction?**

8.a. If completed, please indicate the years of its construction?

1750 - 2012

8.b. If under construction, please indicate, when it is planned to complete it?


8.c. Please add a brief description of future plans for the database.

Annual updates from data contributors

II. **Contact information**

1. **Name of institute or organisation**

   University of Utah, Salt Lake City, Utah

   1.a. Website
   http://healthcare.utah.edu/huntsmancancerinstitute/research/updb

   1.b. Location: city, country
   Salt Lake City, U.S.

   1.c. Postal address
   University of Utah, 2000 Circle of Hope, Huntsman Cancer Institute, Salt Lake City, UT 84112

   1.d. Phone

2. **Name of primary responsible person**

   Ken Smith or Alison Fraser

   2.a. His/her email address
   Ken.Smith@fsc.utah.edu

   2.b. Postal address

   2.c. Phone
   801-585-5135 or 801-581-4468

3. **Administrative information**

   3.a. When this form was filled?
   February 2015

   3.b. Who did it?
   Mineau and Fraser

4. **Main economic funding** (Name of organization(s) who made the grants /sustain it)

   Support for the UPDB is provided by the Pedigree and Population Resource (PPR), which is directed by Ken R. Smith, PhD. The Utah Resource for Genetic
III. Sources: core characteristics

1. Type of the sources.
Indicate how many sources were used for the database and what kind (register, census, certificates ...). Please enter Yes or No and the time period for the main sources. In case of other sources, not listed below, please add their type and specify their main characteristics.
Detailed questions about the characteristics of all core sources are in section C.

<table>
<thead>
<tr>
<th>Type of source</th>
<th>Yes/No</th>
<th>Start year</th>
<th>End year</th>
<th>Explanations:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baptisms</td>
<td>Y</td>
<td>1830</td>
<td>1975</td>
<td>The Church of Jesus Christ of Latter-day Saints (LDS) was organized in New York in 1830. Baptism occurs at age 8 or later. Reported on LDS genealogies.</td>
</tr>
<tr>
<td>2. Marriages from church registers</td>
<td>Y</td>
<td>1790</td>
<td>1975</td>
<td>Reported on LDS genealogies</td>
</tr>
<tr>
<td>3. Burials</td>
<td>Y</td>
<td>1900</td>
<td>1975</td>
<td>Reported on LDS genealogies</td>
</tr>
<tr>
<td>4. Population registers, maintained by church or state</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Civil marriage/divorce certificates</td>
<td>Y</td>
<td>1978</td>
<td>2010</td>
<td>Utah vital records</td>
</tr>
<tr>
<td>9. Nominative lists</td>
<td>Y</td>
<td>2012</td>
<td></td>
<td>Utah Voter Registration</td>
</tr>
<tr>
<td>10 Military draft records</td>
<td>Y</td>
<td></td>
<td></td>
<td>Utah World War II Draft Registrations</td>
</tr>
<tr>
<td>13. Other: Social Security Death Index</td>
<td>Y</td>
<td>1960</td>
<td>2011</td>
<td>For SSNs issued in Utah and Idaho; plus selected others.</td>
</tr>
<tr>
<td>16. Other: Hospital Discharge Summaries</td>
<td>Y</td>
<td>1996</td>
<td>2012</td>
<td>Annual electronic updates</td>
</tr>
</tbody>
</table>

IV. The database: core characteristics

1. Period covered by data: give first and last year of date, if possible
   1790 - present

2. Territory covered by data
   Utah for the most part. Genealogies include migrants into Utah; early migrants originated in New England, Missouri, Illinois and Northern Europe. The families
3. Geographical characteristic: local, regional, national, cross-national
Census records provide towns and precincts for cities. Genealogy records provide national and cross-national origins for migrants. 1940 Census is for all US.

4. Units of observation. Please enter Yes or No for each unit, which forms the sample, the number of units and write explanations/comments. Add other units if they are not listed below, for them explanations are especially important.

<table>
<thead>
<tr>
<th>Units of observation:</th>
<th>Yes /No</th>
<th>Number of units</th>
<th>Explanations:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individuals</td>
<td>Y</td>
<td></td>
<td>All records are individuals' records.</td>
</tr>
<tr>
<td>Married couples</td>
<td>Y</td>
<td></td>
<td>Individuals have been linked across generations to create multi-generational families.</td>
</tr>
<tr>
<td>Families</td>
<td>Y</td>
<td></td>
<td>Census records for 1880, 1900, 1910, 1920, 1930, 1940</td>
</tr>
<tr>
<td>Households</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Farms</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institutions</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Variables per unit included in the database

**On individuals:** Data of birth and dead, age, gender, marital status, religion, occupation, migration, relationship, etc.
Birth date, death date, age, gender, marital status, occupation, relationship, race, geographic location
Please add more variables, if they are not in the list

**On households:** Type of household, children present, age and number of children, etc.
Head of household, relationship to head, age
Please add more variables, if they are not in the list

6. Kinship relations:

6.a. How is kinship recorded in the database?

Genealogies: "Family Group Sheets" from the archives at the Utah Family History Library contain demographic and kinship information on three generations. Vital records: Births and deaths include information on two generations.

6.b. How deep (number of generations) is kinship information going?
The Utah family histories represent pedigrees that may span as many as ten generations. The majority of families living in Utah are represented in this database with a special emphasis on genealogy records of the pioneers of Utah and their Utah descendants. Merging data from genealogies and birth certificates allows information for new generations to be added and most families can be linked across five generations. For example, looking at all individuals born in Utah in 1950, 79 percent have grandparent information available in UPDB and 67 percent have five or more previous generations documented in the UPDB.

7. Completeness

7.a. Are all variables from the sources included in the database? Yes

7.b. Are all individuals who lived in the households of the sample recorded? Yes, for the Utah Census of 1880, 1900, 1910, 1920, 1930, 1940

8. Current data representation:

**Database Software (e.g. MySql, MsSql, Access, please**

**Kinship Analysis Tools:**

Kinclass: This program rapidly identifies arbitrary
classes of relatives for a set of individuals according to a set of criteria, such as first and second-degree relatives who are still alive.

Dynaped: This program takes the output from a control dataset and a kinship dataset, and performs various types of statistical analysis.

The functionality of this program includes:
- Calculation of familial disease incidence
- Calculation of familial average phenotypes, e.g. familial excess longevity
- Estimation of founder relative risks via pedigree-structured Poisson regression
- Extension of above methods to alternative inheritance models (e.g. mitochondrial, x-linked, y-linked, imprinting, social)

Descriptive Tools:
- Kinship Coefficients: Pairs of individuals are provided and their ancestors are compared to identify a common ancestor. If a common ancestor is identified, a kinship coefficient is calculated to indicate the degree of relatedness.
- Pedigree Drawing: Provides all descendants for import into Progen pedigree drawing program.

<table>
<thead>
<tr>
<th>9. Access conditions:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9.a. How does a user get access to the database?</td>
<td>(1) Application to the Utah Resource for Genetic and Epidemiologic Resource <a href="http://research.web.utah.edu/rge/">http://research.web.utah.edu/rge/</a> and (2) an IRB or equivalent from an institution of higher learning</td>
</tr>
<tr>
<td>9.b. What are the conditions and restrictions?</td>
<td><a href="http://healthcare.utah.edu/huntsmancancerinstitute/research/updb/access.php">http://healthcare.utah.edu/huntsmancancerinstitute/research/updb/access.php</a></td>
</tr>
</tbody>
</table>

V. Publications and reports

1. Main publications about the database itself (max. 5)

2. Main or exemplary publications on research based on the database (max. 5)


**Section B**
contains more specific and detailed questions about databases, such as the period(s) of observation, sampling design and procedures, data collection, linkage process and others.

### VI. Observations

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How do individuals enter observation?</td>
<td>Genealogies: These provide information on full sibships regardless of place of birth so there is no censorship. Vital records: Utah birth records will not include children born in other states. Thus full sibships may not be available on all families. Census: Relationship between household members is enumerated.</td>
</tr>
<tr>
<td>2. How do individuals leave observation?</td>
<td>Genealogies: Families or family members that leave Utah may not be followed; we use the Social Security Index and Idaho cancer records to find some. Vital records: Individuals born in Utah who leave the state are not followed; we use the Social Security Index and Idaho cancer records to find some. Some research projects have included additional data, for example, the vital file from Centers for Medicare and Medicaid Services (CMS).</td>
</tr>
<tr>
<td>3. How do households enter observation?</td>
<td>They are observed every 10 years in the census enumeration.</td>
</tr>
<tr>
<td>4. How do households leave observation?</td>
<td>No longer observed in census due to migration or death.</td>
</tr>
<tr>
<td>5. Are some entry or exit dates unknown?</td>
<td>Yes</td>
</tr>
<tr>
<td>6. Are some entry or exit dates estimated?</td>
<td>Yes, if an individual is known to have died out of state then the date that they last had an event in Utah and the death date would provide some parameters.</td>
</tr>
<tr>
<td>7. Can observations be linked to geographic locations?</td>
<td>Yes</td>
</tr>
<tr>
<td>8. Are the dates and locations of movements within the observation area recorded?</td>
<td>No</td>
</tr>
<tr>
<td>9. Are all individuals who lived in selected households recorded? (Selection on basis of the sample or because sampled individuals are living in households)</td>
<td>Yes</td>
</tr>
<tr>
<td>10. Are there related observations that are not included in the database?</td>
<td>Genealogies and civil death certificates include the names of parents who may never have resided in Utah. For some years the birthplace of parents on death certificates is electronic.</td>
</tr>
</tbody>
</table>
### VII. Sampling design and procedures: how was sample(s) defined?

<table>
<thead>
<tr>
<th>1. Source(s):</th>
<th>Genealogies: Over 185,000 Utah families were identified on &quot;Family Group Sheets&quot; from the archives at the Utah Family History Library, which is maintained by the LDS Church. These sheets contain demographic and kinship information on three generations. For a family to be initially selected for UPDB from the genealogies, at least one member had to have a vital event (birth, marriage, death) on the Mormon Pioneer Trail or in Utah. Some family members have birth dates as early as 1740; most occur after 1780; the genealogy set ends with birth dates about 1970. Other records: complete state-wide records with no sampling; the fields that are available vary over time and fields that are computerized vary over time.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which source forms the basis for the sample</td>
<td>Vital records of births and deaths and decennial censuses represent all individuals in Utah.</td>
</tr>
</tbody>
</table>

2. Sampling units:
Households, individuals, regions...

3. Variables used for selection:
Age, gender, marital status, other

4. Selection method:
Random, stratified random, total count, clustered, other

### VIII. Data collection

<table>
<thead>
<tr>
<th>1. Data collection period: When the data was collected and transcribed?</th>
<th>Varies with data set.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Data collection method: Public digital register, transcription, other</td>
<td>State vital records, government census records</td>
</tr>
<tr>
<td>2.a. If transcription, how was the transcription done:</td>
<td>By individuals and from LDS microfilm</td>
</tr>
<tr>
<td>◦ By individuals</td>
<td></td>
</tr>
<tr>
<td>◦ From scanned sources</td>
<td></td>
</tr>
<tr>
<td>◦ From LDS’s microfilms</td>
<td></td>
</tr>
<tr>
<td>◦ Automatic controls</td>
<td></td>
</tr>
<tr>
<td>2.b. How was the checking of the transcription done? For example, by proof reading?</td>
<td>Double keying</td>
</tr>
<tr>
<td>2.c. When was it done?</td>
<td>Varies with data set</td>
</tr>
<tr>
<td>2.d. Purpose of the transcription: please indicate</td>
<td>Research and genealogy</td>
</tr>
<tr>
<td>◦ LDS</td>
<td></td>
</tr>
<tr>
<td>◦ Research</td>
<td></td>
</tr>
<tr>
<td>◦ Genealogy</td>
<td></td>
</tr>
</tbody>
</table>

3. Control methods by researcher:
e.g. Internal consistencies such as a death cannot happen before a birth of the same person

4. Data collection staff:
Please indicate the number of people and their position (member of the project, free-lancer, other)

### IX. Linkage process

<table>
<thead>
<tr>
<th>1. Linkage: Which sources and units of observation have been linked: (e.g. birth/baptisms and death/burials,...)?</th>
<th>Births/Baptisms – Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marriages/ Divorces - Y</td>
<td></td>
</tr>
<tr>
<td>Deaths/Burials - Y</td>
<td></td>
</tr>
<tr>
<td>Census - Y</td>
<td></td>
</tr>
<tr>
<td>Utah fatal deaths -Y</td>
<td></td>
</tr>
<tr>
<td>SSDI - Y</td>
<td></td>
</tr>
</tbody>
</table>
### 2. Documentation of linking:

<table>
<thead>
<tr>
<th></th>
<th>UCR - Y</th>
<th>CDRI - Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.a.</td>
<td>Programme, manually, ...</td>
<td></td>
</tr>
<tr>
<td>2.b.</td>
<td>Name of software if used (and its parameters)</td>
<td>IBM web sphere Quality Stage</td>
</tr>
</tbody>
</table>

### 3. What are the rules for linking?

<table>
<thead>
<tr>
<th></th>
<th>Probabilistic record linking with programmatic validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flags definition (list them: age, name, extra knowledge, ...)</td>
<td></td>
</tr>
</tbody>
</table>

### 4. How each reconstructed person is traceable to the original sources /transcribed data?

<table>
<thead>
<tr>
<th></th>
<th>Keep all original source records and track with unique identifiers.</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each reconstructed person is traceable to the original sources /transcribed data?</td>
<td></td>
</tr>
<tr>
<td>Yes, there is a universal identification number. Because one person may have many records, it is imperative that we have one demographic control record (person record) to represent an individual. Otherwise, researchers would have to view many records for one individual and decide which information they wanted from each record. Thus we create a “person record” that is composite of information contained on all the records for a single person. As additional records are linked, the validity and quantity of the information on a given person increases. Rather than link different record sets to each other, a new record set is linked to the composite individuals with information from all sources.</td>
<td></td>
</tr>
</tbody>
</table>

### 5. How is linkage represented in the database?

<table>
<thead>
<tr>
<th></th>
<th>Yes, there is a universal identification number. Because one person may have many records, it is imperative that we have one demographic control record (person record) to represent an individual. Otherwise, researchers would have to view many records for one individual and decide which information they wanted from each record. Thus we create a “person record” that is composite of information contained on all the records for a single person. As additional records are linked, the validity and quantity of the information on a given person increases. Rather than link different record sets to each other, a new record set is linked to the composite individuals with information from all sources.</th>
</tr>
</thead>
<tbody>
<tr>
<td>For how each reconstructed person is traceable to the original sources /transcribed data?</td>
<td></td>
</tr>
<tr>
<td>For each reconstructed person is traceable to the original sources /transcribed data?</td>
<td></td>
</tr>
</tbody>
</table>

### 6. Linkage percentage

<table>
<thead>
<tr>
<th></th>
<th>Varies with dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each reconstructed person is traceable to the original sources /transcribed data?</td>
<td></td>
</tr>
</tbody>
</table>

### 7. Quality of linkage (own evaluation)

<table>
<thead>
<tr>
<th></th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each reconstructed person is traceable to the original sources /transcribed data?</td>
<td></td>
</tr>
</tbody>
</table>

### 8. What reference/coding systems have been linked to the data?

<table>
<thead>
<tr>
<th></th>
<th>Occupational Coding (Occupations and Industry on Births and Death certificates are coded, but incomplete); 1980 Census Occupation and Industry Classification; 1990 Census Occupation and Industry Classification. These codes apply to occupations for all years, not just 1980 and 1990. An occupation code and industry code were developed for the 1880 census data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference system</td>
<td></td>
</tr>
<tr>
<td>Occupational titles:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Address on Utah driver license has been geocoded. The system used was Universal Transverse Mercator (UTM). We have eastings and northings; as long as you stay in the same UTM zone, these maps to longitude and latitude. Also have census tract and census block related to 2000 census. Vital records contain place codes for city, county and state.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference system</td>
<td></td>
</tr>
<tr>
<td>Locations (including geo-referenced systems):</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Risks, complications, malformations on Birth Certificates coded to ICD 9 for 1978-1988. Cause of Death on Death Certificates coded to ICD6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference system</td>
<td></td>
</tr>
<tr>
<td>Other: Diagnoses:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Y</th>
<th>Other: Race:</th>
<th>Vital records and cancer records have National Center for Health Statistics (NCHS) race codes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Other:</td>
<td>Hispanic classification on vital records and cancer records.</td>
</tr>
</tbody>
</table>
Section C

contains detailed questions about sources used for the databases: their type, scope, content, state of preservation, etc.

Please answer the questions about all the sources used for the database, but do it in a separate form for every type of the source.

<table>
<thead>
<tr>
<th>X. The main characteristics of the source (per every type of the source)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Official name of the source and its English translation</td>
</tr>
<tr>
<td>2. Purpose of the source:</td>
</tr>
<tr>
<td>2.a. Why was this source created?</td>
</tr>
<tr>
<td>2.b. Who created it?</td>
</tr>
<tr>
<td>3. Scope:</td>
</tr>
<tr>
<td>What group of the population was documented in this source?</td>
</tr>
<tr>
<td>4. Time period:</td>
</tr>
<tr>
<td>When the information of the sources was recorded?</td>
</tr>
<tr>
<td>Please indicate the start and the end date.</td>
</tr>
<tr>
<td>5. Geographical area:</td>
</tr>
<tr>
<td>What territory is covered by the source?</td>
</tr>
<tr>
<td>6. Content:</td>
</tr>
<tr>
<td>What was recorded?</td>
</tr>
<tr>
<td>7. Language of written material:</td>
</tr>
<tr>
<td>original sources and documentation</td>
</tr>
<tr>
<td>8. Preservation and storage:</td>
</tr>
<tr>
<td>8.a. Completely preserved</td>
</tr>
<tr>
<td>8.b. Partially destroyed by personnel according to systematic criteria</td>
</tr>
<tr>
<td>8.c. Partially destroyed or damaged for other reasons</td>
</tr>
<tr>
<td>8.d. Reorganized by producer of the source</td>
</tr>
<tr>
<td>8.e. Reorganized by record linkage procedures</td>
</tr>
<tr>
<td>8.f. Where the original records are stored (name of the archive or institution)?</td>
</tr>
<tr>
<td>9. Documentation:</td>
</tr>
<tr>
<td>9.a. Completely documented and accessible by:</td>
</tr>
<tr>
<td>9.b. Partially documented and accessible by:</td>
</tr>
<tr>
<td>9.c. No documentation, but accessible by:</td>
</tr>
</tbody>
</table>