**Ground Truth Data Set Compilation and Name Disambiguation**
**Katherine M. Wisser**
**October 2017**
**SNAC: Social Networks and Archival Context**

## Executive Summary

This report documents the design, methodology and descriptive analysis of two data sets generated for the SNAC initiative. These data sets are intended to provide ground truth data for the project. Included for each data set is a key to the columns established in the data set, as well as summary tables for the characteristics found in individual SNAC records. Additionally, there is an appendix that compares the number of records for each known entity in the first data set based on previous research and this recent effort. A second appendix provides insight into a single corporate body. The report concludes with some ideas for suggested areas of research that is underway or could be pursued.

## Introduction

The goal of this research initiative was to generate a data set that may be used as ground truth data for human and automated identity disambiguation between records in the SNAC database. Despite a number of false starts, the research produced two data sets. Each data set presents its own perspective on the problem of identity disambiguation.

Traditional authority structures rely on the identification of an identity and the construction of a character string (heading) for the representation of that unique identity. Complications can arise when determining individual distinct but related identities as well as distinct and unrelated identities with similar looking character string representations.  In traditional authority structures, minimal contextual information was provided to assist in the disambiguation function. That contextual information was stored with the heading and variant forms of name in a separate record, allowing the distinct character string to act as the sole representation of the entity in other metadata expressions. The SAA *Glossary of Archival and Records Terminology* defines authority control as: "The process of establishing the preferred form of a heading, such as proper name or subject, for use in a catalog, and ensuring that all catalog records use such headings." It is further noted in the *Glossary* that "the preferred form of a heading is typically defined by a standard. Once established, the form is usually recorded in an authority file for future reference, along with cross-references from other forms of the heading to ensure consistency."[1] This system has worked effectively in both the library catalog environment and through traditional archival descriptive output.

Recent developments have presented expanded opportunities for the generation and aggregation of contextual information specifically around entities that create and are the subject of archival materials and other resources. Efforts such as VIAF and ISNI take the cue from the long-standing Library of Congress Authority File in achieving large-scale identity aggregation. SNAC has also taken on this kind of aggregation work. The advantages and challenges, however, stem from the basis of SNAC records; that is, original archival description created not for the intention of identity disambiguation but instead for the description of archival materials. Additionally, SNAC, based on the EAC-CPF standard,

[1] *https://www2.archivists.org/glossary/terms/a/authority-control*

documents its sources and goes further in providing space for formal descriptive components and relationships that can help with disambiguation.

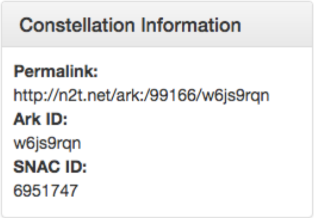
This research explores the data in existing SNAC records that can be used to resolve identities. It is based on the generation of two data sets, created through different strategies, that reveal issues surrounding identity disambiguation, human assessments versus machine matching possibilities, and some insight into existing descriptive practices. This report outlines the creation of the data sets, decisions made in the format and assessment keys established and some observations from the data sets created. It concludes with a discussion of potential research avenues that these data sets provide.

## Data Sets

The main objective of this research initiative was to produce a "ground truth" data set that can be used to inform both human and automated matching processes. The process to develop these data sets was one of the most challenging aspects of this work. While several attempts were made to generate a single ground truth data set, eventually two strategies were adopted. This resulted in two data sets (identified below and through file names as Data Set I and Data Set II), which present the different, although minor, weaknesses for the proposed use. For the purposes of this description and analysis, these data sets are treated separately.

With the exception of Column A and Column F, which differ between the two data sets, discussed below, the data sets follow the same structure:

- Column B contains the headings from the record.
- Columns C and J contain identifiers. These identifiers are both unique to the SNAC record. It is not clear what the difference is between the two identifier structures, therefore both are included in the data sets (see Figure 1).



Constellation Information

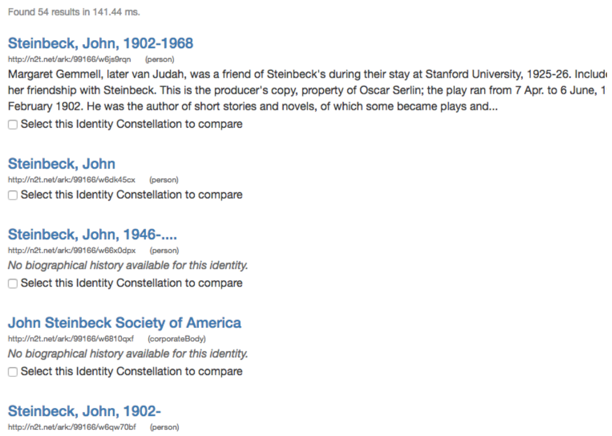
**Permalink:**
http://n2t.net/ark:/99166/w6js9rqn
**Ark ID:**
w6js9rqn
**SNAC ID:**
6951747

*Figure 1: Two identifiers for each record*

- Columns D & E contain evidence from the original source records and notes from the SNAC records that were used in the human assessment of the records' matching.
- The assessment is reflected in Column F (but according to different scales as discussed below).
- Columns G, H and I include identifiers for related LCNAF, VIAF and ISNI records (connected in External Resources in the SNAC record).
- Columns K-S include data derived from the SNAC record directly, including date information, nationality, language, subjects, occupations, and places.
- Column T includes any links to records that are deemed by SNAC, "May be the same as."
- Columns U, V and W includes numbers of links to archival collections, resources, and CPF connections.
- Column X consists of a space to include notes, for example, when a personal name record is identified as an entity type, "corporateBody".

### *Data Set I*

The first data set in the ground truth data project is derived from a list of known entities. Those entities were used in previous research for SNAC looking at the undermatching of records. The original source was a research initiative associated with the Small World project.[2] A selection of 54 entities were explored and the number of potential records for that entity were selected. This selection was based only on the heading retrieved in the results list (see Figure 2) as well as "may be same as" links followed within SNAC records.



Found 54 results in 141.44 ms.

**Steinbeck, John, 1902-1968**
http://n2t.net/ark:/99166/w6js9rqn (person)
Margaret Gemmell, later van Judah, was a friend of Steinbeck's during their stay at Stanford University, 1925-26. Includ
her friendship with Steinbeck. This is the producer's copy, property of Oscar Serlin; the play ran from 7 Apr. to 6 June, 1
February 1902. He was the author of short stories and novels, of which some became plays and...
□ Select this Identity Constellation to compare

**Steinbeck, John**
http://n2t.net/ark:/99166/w6dk45cx (person)
□ Select this Identity Constellation to compare

**Steinbeck, John, 1946-....**
http://n2t.net/ark:/99166/w66x0dpx (person)
*No biographical history available for this identity.*
□ Select this Identity Constellation to compare

**John Steinbeck Society of America**
http://n2t.net/ark:/99166/w6810qxf (corporateBody)
*No biographical history available for this identity.*
□ Select this Identity Constellation to compare

**Steinbeck, John, 1902-**
http://n2t.net/ark:/99166/w6qw70bf (person)

*Figure 2: Results list*

This approach resulted in 560 individual records total for the 54 entities searched.

Based on a review of the records, an "anchor record" was chosen. This record consists of the most complete record representing the literary figure being sought. This anchor record then served as the point of comparison for other records in an entity record cluster. Therefore, the sureness ratings in this data set are based on whether the other records in the cluster match the anchor record. Anchor records tend to be very full records with numerous connections to archival collections, resources, corporate bodies, persons and families. They also can include subjects, occupations and places as well as segmented demographic information such as birth and death dates, nationality and languages. Finally, they tend to have extensive biographical notes embedded within them, helping to fully establish direct and indirect evidence to be used in the comparison activities.

**Table 1: Key to fields for Data Set I**

| Column | Field | Input guidelines |
|---|---|---|
| A | Entity | Grouping mechanism for records being tested: identified by heading of literary figure being search. |
| B | Heading | Heading from SNAC record |
| C | SNAC ID | Alphanumeric string assigned through SNAC. Each SNAC ID begins "w6" |

[2] http://slis.simmons.edu/smallworld/

| **Column** | **Field** | **Input guidelines** |
|---|---|---|
| D | Notes from original source(s) | Any notes from original source used in human judgment and feeding into sureness rating. |
| E | Notes on connected resources | Titles of any works associated with the entity, whether from the SNAC record or the original source data |
| F | Sureness rating | Human judgment placed on whether or not the entity in the record matches the entity in the anchor record. Values include:<br>1 Matched based on same titles of resources associated with the entity, e.g., *Forever Amber*, for Kathleen Winsor<br>2 Direct evidence in description and original source data<br>3 Indirect evidence in description and original source data<br>4 Does not match<br>5 Cannot determine because of a lack access to original source |
| G | LCNAF | LCCN from related Library of Congress Name Authority File |
| H | VIAF | Record number in associated VIAF record |
| I | ISNI | ISNI identifier in associated VIAF record |
| J | Constellation ID | Constellation ID assigned to entity record |
| K | Birth date | Dates identified as birth date for the entity in the SNAC record |
| L | Death date | Dates identified as death date for the entity in the SNAC record |
| M | Active date 1 | First listed active date for the entity in the SNAC record |
| N | Active date 2 | Second listed active date for the entity in the SNAC record |
| O | Nationality | Nationality (or abbreviation) of the entity in the SNAC record |
| P | Language | Language of the entity in the SNAC record |
| Q | Subjects | Subjects associated with the entity in the SNAC record |
| R | Occupations | Occupations associated with the entity in the SNAC record |
| S | Places | Places associated with the entity in the SNAC record |
| T | MayBeSameAs | SNAC IDs for records identified by SNAC as "May Be the Same As" or "Similarity assertions" |
| U | # of archival collections | Number of archival collections related to the entity in the SNAC record |
| V | # of resources | Number of resources related to the entity in the SNAC record |
| W | # of CPF | Number of corporate bodies, persons, and families related to the entity in the SNAC record |

### *Data Set II*

Data Set II was created by exploring a browsable list of headings for an alphabetic range (D). "Pairs" were identified as potential matching entities. While they are initially referred to as "pairs," that does not imply that the number of records identified in this browsing stage was limited to two. "Pairs" consist of two or more headings near each other in the alphabetical browsing list. After the identification of "pairs" through browsing, the headings were searched in the search interface to find additional records that would not appear close in an alphabetical list but could represent the same entity. This would include records where the heading was in direct order, rather than indirect order, for example. In some cases, this expanded the number of records in the original pair. The result of these stages created entity record clusters similar to that in Data Set I. While in Data Set I, these clusters are identified by the name of the original entity sought (Data Set I, Column A), in Data Set II, they are identified by a number (Data Set II, Column A).

Because the pairs and resulting entity record clusters are randomly selected, the idea of an anchor record cannot apply in this data set. These clusters consist of individuals who may only appear within the archival collection from which their name (and thus the generation of the SNAC record) was derived. This presents problems on several levels, based on a lack of familiarity with the entities by the researcher and a lack of accompanying information to help establish and distinguish the identity from but also with others.

**Table 2: Key to fields for Data Set II**

| Column | Field | Input guidelines |
|---|---|---|
| A | Entity Example No. | Grouping mechanism for records being tested: identified by a number. |
| B | Heading | Heading from SNAC record |
| C | SNAC ID | Alphanumeric string assigned through SNAC. Each SNAC ID begins "w6" |
| D | Notes from original source(s) | Any notes from original source used in human judgment and feeding into sureness rating. |
| E | Notes on connected resources | Titles of any works associated with the entity, whether from the SNAC record or the original source data |
| F | Sureness rating | Human judgment placed on whether or not the entities in the records match each other. Values include:<br>1 Match between records<br>2 May or may not match between records<br>3 Does not match<br>4 Cannot determine |
| G | LCNAF | LCCN from related Library of Congress Name Authority File |
| H | VIAF | Record number in associated VIAF record |
| I | ISNI | ISNI identifier in associated VIAF record |
| J | Constellation ID | Constellation ID assigned to entity record |
| K | Birth date | Dates identified as birth date for the entity in the SNAC record |

| Column | Field | Input guidelines |
|---|---|---|
| L | Death date | Dates identified as death date for the entity in the SNAC record |
| M | Active date 1 | First listed active date for the entity in the SNAC record |
| N | Active date 2 | Second listed active date for the entity in the SNAC record |
| O | Nationality | Nationality (or abbreviation) of the entity in the SNAC record |
| P | Language | Language of the entity in the SNAC record |
| Q | Subjects | Subjects associated with the entity in the SNAC record |
| R | Occupations | Occupations associated with the entity in the SNAC record |
| S | Places | Places associated with the entity in the SNAC record |
| T | MayBeSameAs | SNAC IDs for records identified by SNAC as "May Be the Same As" or "Similarity assertions" |
| U | # of archival collections | Number of archival collections related to the entity in the SNAC record |
| V | # of resources | Number of resources related to the entity in the SNAC record |
| W | # of CPF | Number of corporate bodies, persons, and families related to the entity in the SNAC record |

### *Comparison between Data Set I and Data Set II*

The primary difference between the two data sets centers on an ability to determine certainty of matches when little or nothing previously known about the entities is considered. With the first data set, the entities considered are American literary figures. This approach has two distinct advantages. The first centers on the fact that these figures have associated works or resources that can be leveraged. This mirrors more traditional authority work, where part of the task is to manage the relationships between names and works. The second centers on the fact that there is general knowledge about literary figures that can contribute to assessments of matching with direct but also indirect evidence. Therefore, in considering the matching assessments, a more granular rating system was required for the first data set.

The second data set is comprised of randomly identified entities, the groupings which were created because of a proximity of their headings in a browsable list. The challenges rest in the fact that these entities are random, and, as discussed above, lack the advantage of previous knowledge and/or accompanying documentation.

## Data observations

These data sets provide some insight into the differences between known entities and those randomly selected. The data sets also demonstrate the frequency and pattern of formal descriptive elements provided in the SNAC records.

**Table 3: Assessment of Records of Data Set I**

| Assessment Rating | Number | Percentage |
|---|---|---|

| | | |
|---|---|---|
| 1: Matched based on same titles of resources associated with the entity, e.g., *Forever Amber*, for Kathleen Winsor | 85 | 15.2% |
| 2: Direct evidence in description and original source data | 219 | 39.1% |
| 3: Indirect evidence in description and original source data | 157 | 28.0% |
| 4: Does not match | 25 | 4.5% |
| 5: Cannot determine because of a lack access to original source | 20 | 3.6% |

* Note that this table does not include the 54 anchor records (9.6%).

**Table 4: Assessment of Records of Data Set II**

| Assessment Rating | Number | Percentage |
|---|---|---|
| 1: Match between records | 86 | 35.4% |
| 2: May or may not match between records | 141 | 58.0% |
| 3: Does not match | 4 | 1.6% |
| 4: Cannot determine because of a lack access to original source | 10 | 4.1% |

Tables 3 and 4 provide an overview of the occurrence of specific sureness ratings for the two data sets. Data Set 1 contains a more nuanced and granular assessment rating which ranges from sureness to indirect evidence. As discussed earlier, Data Set II does not benefit from a better understanding of the entities being examined, therefore the the sureness ratings are subsequently less detailed. In the first data set, the matches (rated 1, 2 and 3) constitute 82.3% of the entire data set. Less than 5% are determined to be different entities and just 3.6% were records that could not be determined because of a lack of access to the source data.

In the second data set, over one-third (35.4%) of records are determined to match, but a full 58.0% occupy an in-between space of may or may not. This rating is assigned when there is some indication that the entities may be the same but a lack of direct evidence or subtle indirect evidence that calls into some question the matching. In contrast to the first data set, the definitive "does not match" is a much smaller percentage. With the catch-all category of "may or may not," and the lack of direct or indirect evidence that would incline toward a contrary match, the "match" and "does not match" categories were harder to determine. Not surprisingly, the issues with access to original source data are similar between the two data sets.

**Table 5: Occurrences of Data Points in Data Set I**

| Data Point | Number | Percentage |
|---|---|---|
| LCNAF Records | 71 | 12.7% |
| VIAF Records | 94 | 16.8% |
| ISNI Records | 69 | 12.3% |
| Birth date | 130 | 23.2% |
| Death date | 100 | 17.9% |
| Active date 1 | 19 | 3.4% |
| Active date 2 | 19 | 3.4% |
| Nationality | 52 | 9.3% |
| Language | 52 | 9.3% |
| Subjects | 67 | 12.0% |
| Occupations | 39 | 7.0% |
| Places | 40 | 7.1% |
| May Be the Same As | 167 | 29.8% |
| Archival collections | 555 | 99.1% |
| Resources | 73 | 13.0% |
| CPF relationships | 437 | 78.0% |

**Table 6: Occurrences of Data Points in Data Set II**

| Data Point | Number | Percentage |
|---|---|---|
| LCNAF Records | 27 | 11.1% |
| VIAF Records | 43 | 17.7% |
| ISNI Records | 31 | 12.8% |
| Birth date | 54 | 22.2% |
| Death date | 44 | 18.1% |
| Active date 1 | 48 | 19.8% |
| Active date 2 | 46 | 18.9% |
| Nationality | 17 | 7.0% |
| Language | 17 | 7.0% |
| Subjects | 24 | 9.9% |
| Occupations | 10 | 4.1% |
| Places | 49 | 20.2% |
| May Be the Same As | 34 | 14.0% |
| Archival collections | 241 | 99.2% |
| Resources | 31 | 12.8% |
| CPF relationships | 156 | 64.2% |

Tables 5 and 6 record the number and percentages of specific data elements in the SNAC records. The first three aspects examined include links to Library of Congress Name Authority Files, Virtual Archival Authority Files, and International Standard Number Identifiers bear similar numbers across the data sets (12.7% and 11.1%, 16.8% and 17.7%, and 12.3% and 12.8% respectively).

Date information is the most common formal descriptive element included in records throughout the two data sets. In the first data set, 141 records contain some kind of date information (birth and death dates, birth dates only, death dates only and various active dates). This constitutes 25.2% of the records in that data set. In the second data set, 144 records contain some kind of date information. This constitutes 59.3% of the records in that data set. The discrepancy of these percentages may be attributed to the size of the data sets themselves. A full one-third of the entity clusters in the first data set contain 10 or more records; another one-third contain 5-9 records. The final one-third contains 2-4 records. In contrast, in Data Set II, a full 86.7% (65) of the entity clusters contain 2-4 records; only 10.7% (8) contains 5-9 records, and only 2.7% (2) entity clusters contain 10 or more records (see Table 7).

**Table 7: Counts of records in entity clusters for Data Set I and Data Set II**

| Number of Records in an entity cluster | Data Set I | | Data Set II | |
|---|---|---|---|---|
| | Number | Percentage | Number | Percentage |
| 2-4 | 18 | 33.3% | 65 | 86.7% |
| 5-9 | 18 | 33.3% | 8 | 10.7% |
| 10 or more | 18 | 33.3% | 2 | 2.7% |
| Total | 54 | | 75 | |

Further, those 2 entity clusters contain 12 and 15 records respectively, whereas the highest numbers of records in the first data set include 40 (Steinbeck), 38 (Whitman), and 32 (Twain). For other record counts for Data Set I, see Appendix A.  The large number of records associated with entities in the first data set will naturally skew the percentages of date information. For example, in the entity cluster for John Steinbeck, of the 40 records, only 4 contain date information (10%).

Not surprisingly, nationality and language fields are less populated than date information. Date information is often derived from the heading itself, as dates have consistently been a part of heading formation. The differences observed in the date fields, therefore, is not as apparent in the two data sets, accounting for 9.3% of the records in Data Set 1 for both nationality and language as opposed to 7.0% of the records in Data Set II for both fields. With subjects, occupations and places, there are some differences (12.0% for Data Set I and 99% for Data Set II for subjects; 7.0% for Data Set I and 4.1% for Data Set II for occupations; and 7.1% for Data Set I and 20.2% for Data Set II for places). The biggest difference in the data sets is that of places, although there is not an explanation for why that would be.

A surface analysis of the connections to archival collections and resources are remarkably similar. Most records in both data sets have a connection to at least one archival collection. For connected resources, 13.0% of the records in Data Set 1 and 12. 8% of the records in Data Set II have at least one resource. In CPF relationships, the gap is slightly larger, 78.0% and 64.2% respectively. The counts of links to archival collections, resource and CPF entities will be discussed below in the context of future areas of research.

## Problems observed in records

The close examination of so many records brought a few things to light that impact the quality of the data in SNAC records.

Found Errors:

- Mis-assignment of entity types: in Data Set 1, there are four instances of personal name records that are assigned "corporateBody" entity types. Given that entity type is one of the primary matching criteria, these records would automatically be discounted before an analysis of the heading or other facets. Examples of this include:

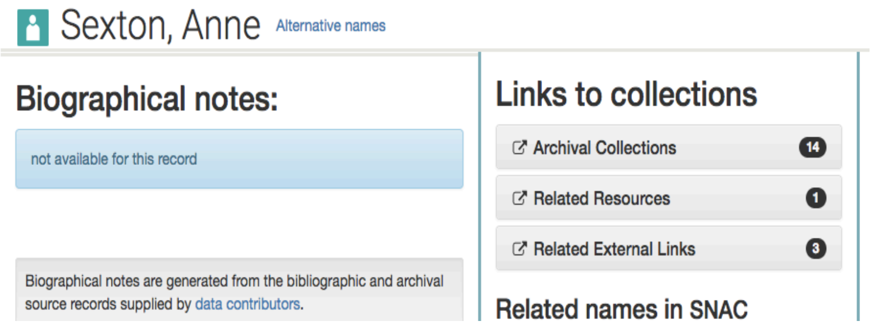
Alexander Woollcott, both rated at 2 for sureness rating:

- Alexander Woollcott, w6f32t4c
- Alexander Woollcott, 1887-1942, w609061v

Edith Wharton, both rated 2 or higher, with date information as well as additional name consistency

- Edith Wharton, w68479p1
- Wharton, Edith Newbold (Jones), 1862-1937

- Two different entities in the same SNAC record:
  - Anne Sexton (w6012ct7): In this record, the archival collections associated with the heading is the poet, Anne Sexton, who died in 1974. The LCNAF and VIAF records, however, are for an Anne Sexton who wrote *The Complete Book of Soft Toys* (1997).



Sexton, Anne Alternative names

Biographical notes:

not available for this record

Biographical notes are generated from the bibliographic and archival source records supplied by data contributors.

Links to collections

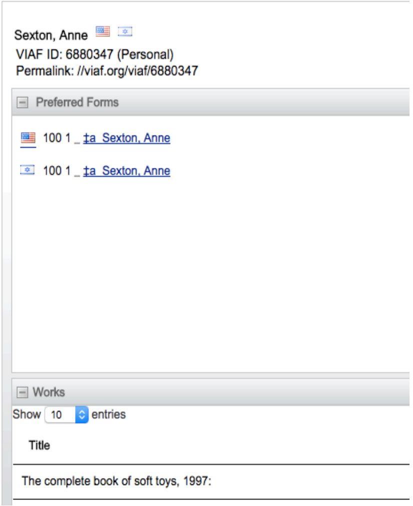
Archival Collections 14

Related Resources 1

Related External Links 3

Related names in SNAC

*Figure 2: Anne Sexton record with 14 archival collections referencing Anne Sexton, the poet*



Sexton, Anne

VIAF ID: 6880347 (Personal)

Permalink: //viaf.org/viaf/6880347

Preferred Forms

100 1 _ ‡a Sexton, Anne

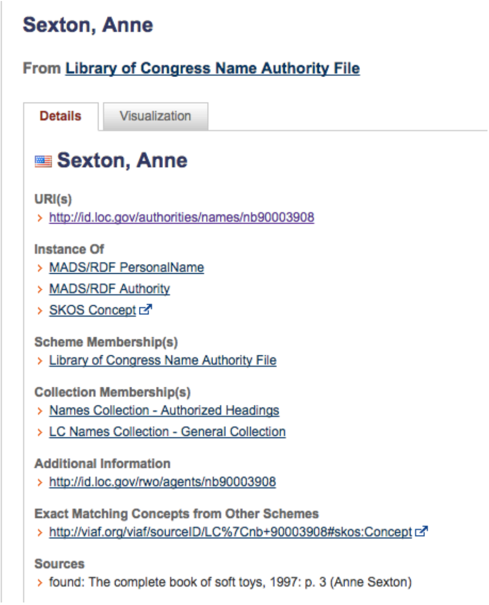
100 1 _ ‡a Sexton, Anne

Works

Show 10 entries

| Title |
|---|
| The complete book of soft toys, 1997: |

*Figure 3: Associated VIAF record for Anne Sexton record example*



Sexton, Anne

From Library of Congress Name Authority File

Sexton, Anne

URI(s)
- http://id.loc.gov/authorities/names/nb90003908

Instance Of
- MADS/RDF PersonalName
- MADS/RDF Authority
- SKOS Concept

Scheme Membership(s)
- Library of Congress Name Authority File

Collection Membership(s)
- Names Collection - Authorized Headings
- LC Names Collection - General Collection

Additional Information
- http://id.loc.gov/rwo/agents/nb90003908

Exact Matching Concepts from Other Schemes
- http://viaf.org/viaf/sourceID/LC%7Cnb+90003908#skos:Concept

Sources
- found: The complete book of soft toys, 1997: p. 3 (Anne Sexton)

*Figure 4: Associated LCNAF record for associated Anne Sexton Record*

The errors in source data reflect those found in earlier research. These errors for personal names include typographical errors such as punctuation differences, spelling differences, spacing and misspelling, date errors such as discrepancies or the presence of absence of date information, and names presented in different formats, such as inverted order versus direct order.[3] This latter error source may be from decisions to mark up and/or harvest from source data, where the construction of the data was not formed according to standardized name expressions.

## Conclusion

The intensive construction of data sets such as these provides a foundation for considering identity resolution for headings that appear to be representative of the same entity. This research, however, has also illuminated more fundamental questions about the interpretive and subjective nature of archival description, and how strategies may evolve with on-going developments that come both from the technology and social and cultural forces. Who we choose to include in narrative and formal descriptive passages, where we choose to harvest data, and how we contextualize those entities all seem to be relevant questions that could be debated in the coming years. Certainly data sets like these help to reveal past description behavior that illustrates what has been naturally occurring before we considered data-centric models for our representations.

One research initiative already underway is the analysis of connected corporate bodies, persons and families (CPF) among the records in Data Set I. That data set was chosen because of the increased degree of sureness in the rating. That along with the increased number of CPF connections associated with the entities in Data Set I could yield more concrete evidence of its utility for matching assistance. This research direction combines automated data gathering and in-depth analysis based on the matching assessment and record clustering already established in the data set.

[3] Wisser, K.M., "The Error of Our Ways: Using Metadata Quality Research to Understand Common Error Patterns in the Application of Name Headings" S. Closs et al., eds. *MTSR* 2014, CCIS 478, pp. 83-94, 2014.

Additional points of analysis could include related archival collections, resources, subjects, occupations and places. It is anticipated that they will not be equally useful. For instance, the low occurrence of values for occupations (7.0% in Data Set I and 4.1% in Data Set II) may eliminate its utility for this kind of data leveraging. The same research strategy could be employed on a larger scale with entity record clusters in Data Set II. This would require the examination of additional entity record clusters beyond the 75 examined for this research. A larger record base may serve as a balance against the randomness of the entities in the data set. Along with these specific data points, individual clusters should be explored to see where the SNAC data elements share or do not share commonalities. That kind of research has not been explored here, but it underlies the utility of these data sets. Individual cluster analysis and then comparison across clusters could reveal patterns that could also be leveraged.

Overall, the SNAC data as a whole reveals many aspects to both the challenges of data harvesting and aggregation, but also to the opportunities for discovery that are embedded in data at such a magnitude.

## Appendix A: Record cluster numbers for Data Set I, 2014 and 2017

This appendix examines the number of records associated with the entity in 2014 research completed and this research in 2017 (Data Set I).

**Table 7: Comparison of the number of records for each known entity in Data Set I**

| Entity | 2014 | 2017 |
|---|---|---|
| Zukofsky, Louis, 1904-1978 | 1 | 7 |
| Young, Stanley | 1 | 6 |
| Yoder, Don | 1 | 7 |
| Woollcott, Alexander | 4 | 17 |
| Wolfe, Thomas | 2 | 9 |
| Winters, Yvor | 1 | 4 |
| Winsor, Kathleen | 1 | 3 |
| Williams, William Carlos | 1 | 7 |
| Williams, Tennessee | 1 | 29 |
| Wicker, Tom | 1 | 4 |
| Whitman, Walt | 4 | 38 |
| Wharton, Edith | 5 | 18 |
| Weidman, Jerome | 1 | 3 |
| Washington, Booker T. | 4 | 4 |
| Walter, Eugene | 1 | 5 |
| Walser, Richard | 1 | 5 |
| Vidal, Gore | 3 | 14 |
| Updike, John | 1 | 17 |
| Twain, Mark | 1 | 32 |
| Thoreau, Henry David | 3 | 7 |
| Thaxter, Celia | 3 | 18 |
| Tate, James | 1 | 2 |

| Entity | 2014 | 2017 |
|---|---|---|
| Tarn, Nathaniel | 1 | 2 |
| Swansea, Charleen | 1 | 4 |
| Sukenick, Ron | 1 | 2 |
| Stone, Alma | 1 | 3 |
| Stephenson, Shelby | 1 | 13 |
| Steinbeck, John | 1 | 40 |
| Stein, Gertrude | 3 | 19 |
| Stegner, Wallace | 1 | 8 |
| Stafford, Clayton | 1 | 4 |
| Spencer, Elizabeth | 2 | 4 |
| Sorrentino, Gilbert | 1 | 8 |
| Smith, David Stanley | 1 | 2 |
| Smith, Betty | 1 | 8 |
| Simpson, Bland | 1 | 2 |
| Shapiro, Karl | 1 | 5 |
| Sexton, Anne | 2 | 7 |
| Scott, Evelyn | 1 | 7 |
| Saroyan, William | 1 | 13 |
| Hardwick, Elizabeth | 1 | 9 |
| Harmon, William | 1 | 6 |
| Hawthorne, Nathaniel | 1 | 24 |
| Hearn, Lafcadio | 1 | 7 |
| Hearon, Shelby | 1 | 4 |
| Hellmann, Lillian | 1 | 17 |
| Hemingway, Ernest | 2 | 18 |
| Hergescheimer, Joseph | 1 | 2 |
| Higginson, Thomas Wentworth | 3 | 26 |
| Hobson, Linda Whitney | 1 | 2 |
| Howe, Fanny | 2 | 10 |
| Howe, S.G. (Samuel Gridley) | 3 | 12 |
| Humphrey, William | 1 | 5 |
| Hurst, Fannie | 3 | 7 |
| Hwang, David | 1 | 4 |

## Appendix B: A Corporate Body Record Cluster

This research initiative looked exclusively at personal names. Corporate bodies present more complex issues because of more frequent name changes, merges, and so on. One such corporate body was explored to illustrate the complexities that would be present in a similar research strategy on corporate bodies was explored. Despite the challenges, some information may be very fruitful from a corporate body focused initiative.

**Table 8: Sample Corporate Body**

| Heading | Notes | Identifiers | Archival / Resources / CPFs |
|---|---|---|---|
| Goulding, D'Almaine & Co. | Hodges family collection, sheet music collection | SNAC ID: w68175qs<br>Const. ID: 3264807 | 2/0/1 |
| Goulding, Phipps, D'Almaine & Co. | Newland/Zeuner collection, Hodges family collection | SNAC ID: w6rs421c<br>Const. ID: 5984190 | 2/0/2 |
| Goulding & D'Almaine | 19th century letter, to publisher | SNAC ID: w69969cn<br>Const. ID: 26933628 | 1/0/1 |
| Goulding & D'Almaine | Hodges family collection | SNAC ID: w6zb1gm6<br>Const. ID: 52399699 | 1/0/1 |
| D'Almaine & Co. | Sheet music collection, London | SNAC ID: w6166p16<br>Const. ID: 47776162 | 1/0/1 |
| Goulding & D'Almaine | Newland Zeuner collection, printed music, 19th century | SNAC ID: w68f20h8<br>Const. ID: 43148375 | 1/0/1 |
| D'Almaine & Co. | Sheet music, London | SNAC ID: 67530069<br>Const. ID: 67530069 | 1/0/1 |
| D'Almaine & Co. | Sheet music, 19th century | SNAC ID: w6t85psh<br>Const. ID: 978287 | 1/0/0 |
| D'Almaine & Co. | Hodges family collection | SNAC ID: w66z1f6s<br>Const. ID: 14875705 | 1/0/1 |
| D'Almaine & Co. | References being sent copies of works, 1839 | SNAC ID: w6hb41zd<br>Const. ID: 52188513 | 1/0/1 |
| Goulding, D'Almaine, Potter & Co. | Hodges family collection | SNAC ID: w6x18b<br>Const. ID: 32016403 | 1/0/1 |
| Goulding, Phipps & D'Almaine | Compositions, listed as publisher, 19th century;<br>Hodges family collection, Music Division, Library of Congress | SNAC ID: w6fc75t4<br>Const. ID: 46278454 | 1/0/1 |
| D;Almaine and Co (music publishers: active 1501-1801) | Music publishers, 19th century France<br>Active date 1: 1501<br>Active date 2: 1801<br>Places: Schortau, Saxony | SNAC ID: w6bh29w7<br>Const. ID: 35773782 | 2/0/0 |