Toward a Metadata Generation Framework

A Case Study at Johns Hopkins University

Mark Patton, <mpatton@jhu.edu>
David Reynolds, <davidr@jhu.edu>
G. Sayeed Choudhury, <sayeed@jhu.edu>
Tim DiLauro, <timmo@jhu.edu>
The Johns Hopkins University

Introduction

In the June 2003 issue of D-Lib Magazine, Kenney et al. (2003) discuss a comparative study between Cornell's email reference staff and Google's Answers service. This interesting study provided insights on the potential impact of "computing and simple algorithms combined with human intelligence" for library reference services. As mentioned in the Kenney et al. article, Bill Arms (2000) had discussed the possibilities of automated digital libraries in an even earlier D-Lib article. Arms discusses not only automating reference services, but also another library function that seems to inspire lively debates about automation—metadata creation. While intended to illuminate, these debates sometimes generate more heat than light.

In an effort to explore the potential for automating metadata generation, the Digital Knowledge Center (DKC) of the Sheridan Libraries at The Johns Hopkins University developed and tested an automated name authority control (ANAC) tool. ANAC represents a component of a digital workflow management system developed in connection with the digital Lester S. Levy Collection of Sheet Music.

The evaluation of ANAC followed the spirit of the Kenney et al. study that was, as they stated, "more exploratory than scientific." These ANAC evaluation results are shared with the hope of fostering constructive dialogue and discussions about the potential for semi-automated techniques or frameworks for library functions and services such as metadata creation. The DKC's research agenda emphasizes the development of tools that combine automated processes and human intervention, with the overall goal of involving humans at higher levels of analysis and decision-making.

Others have looked at issues regarding the automated generation of metadata. A session at the 2003 Joint Conference on Digital Libraries was devoted to automatic metadata creation, and a session at the 2004 conference addressed automated name disambiguation. Commercial vendors such as OCLC, Marcive, and LTI have long used automated techniques for matching names to Library of Congress authority records. We began developing ANAC as a component of a larger suite of open source tools to support workflow management for digital projects.

This article describes the goals for the ANAC tool, provides an overview of the metadata
records used for testing, describes the architecture for ANAC, and concludes with discussions of the methodology and evaluation of the experiment comparing human cataloging and ANAC-generated results.

Automated Name Authority Control (ANAC)

ANAC, described in DiLauro et al. (2001) and Warner and Brown (2001), is a tool specifically developed to identify the Library of Congress (LC) authorized name for each name in the descriptive metadata of the Lester S. Levy Collection of Sheet Music. As an example, consider that the same individual wrote the lyrics for both Levy Collection titles My Idea of Something to Go Home to and Pretty as a Picture, but the lyricists are listed as "Robert Smith" and "Robert B. Smith", respectively. Name authority control enhances access by clustering the works of a named creator, independently of the name used on a given work.

The main reason for undertaking ANAC was to develop a tool that would reduce the costs associated with introducing name authority control to the Levy metadata. Relying exclusively on human catalogers would be substantially more expensive and time consuming because of the size of the collection. There are about 29,000 descriptive metadata records in the Levy Collection with roughly 39,000 names. The number is not precise because the names must be extracted from a free form statement of responsibility. The LC authority file contains about 3.5 million entries.

Levy Metadata and Architecture of ANAC

The Levy metadata records are stored as individual files in XML. An example record is provided below:

```xml
<?xml version="1.0"?>
<lsm:record
xmlns:lsm="http://levysheetmusic.mse.jhu.edu/documentation/LevySchema.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://levysheetmusic.mse.jhu.edu/documentation/LevySchema
http://levysheetmusic.mse.jhu.edu/documentation/LevySchema.xsd">
  <lsm:title>I Would Not Die In Spring Time. Ballad.</lsm:title>
  <lsm:composerlyricistarranger>Composed and Arranged for the Piano Forte By
Milton Moore.</lsm:composerlyricistarranger>
  <lsm:names>
    <lsm:name role="composer,arranger">Milton Moore</lsm:name>
  </lsm:names>
  <lsm:publication>
    <lsm:location>Baltimore</lsm:location>
    <lsm:publisher>F.D. Benteen</lsm:publisher>
    <lsm:date>1850</lsm:date>
  </lsm:publication>
  <lsm:formofcomposition>strophic</lsm:formofcomposition>
  <lsm:instrumentation>piano and voice</lsm:instrumentation>
  <lsm:firstline>I would not die in Spring time when all is bright
around</lsm:firstline>
  <lsm:performer>Sung With the Most Unbounded Success by Mr. Turner, the
American Ballad Singer.</lsm:performer>
  <lsm:plateno>1754</lsm:plateno>
  <lsm:subjects>
```
The name fields are ignored because they are often wrong. They were added by an offline script that extracted the names from the statement of responsibility. As a preprocessing step, ANAC reads in the Levy metadata records, runs its own rule-based name extraction algorithm on the statement of responsibility, and stores the metadata in an internal format to speed processing.

The 3.5 million authority records from the LC name authority file were stored in a MySQL relational database to enhance performance. The MARC record data model was preserved. The benefit of this approach was the simplicity of database creation. The drawback was that accessing the records required knowledge of the peculiarities of the MARC record format.

The most important operations performed on the database are retrieving the set of records with a given surname and retrieving a record given an LC identifier. The latter operation is only important when processing the training data for the first time. After that, the internal database identifier for a MARC record is kept with its corresponding LC identifier.

**Methodology**

ANAC uses a naïve Bayes classifier (Domingos and Pazzani 1997) to assign a name in a Levy metadata record to an LC record. The algorithm works as follows. Given a name N in a Levy record, for each LC record R with the same surname as N, calculate the probability that the correct assignment of N is R. Choose the most probable assignment. If the most probable assignment is below a certain threshold, report that an LC record for N could not be found. We limit ourselves to LC records with the same surname because checking 3.5 million LC records for each name is prohibitively time consuming.

The probability calculation uses Bayes rule. Let H be the hypothesis that N is correctly assigned to R. Let E be evidence used to evaluate H. Then \( P(H|E) = \frac{P(E|H)P(H)}{P(E)} \). The evidence E consists of multiple parts, each of which is the result of a Boolean test involving N and R. The naïve Bayes conditional independence assumption simplifies the calculation of P(H|E). Let \( E_i \) be the \( i^{th} \) piece of \( n \) pieces of evidence. In this case, \( P(H|E) \) is the product for \( i = 1 \) to \( n \) of \( P(E_i|H)P(H) / P(E_i) \). Missing pieces of evidence are skipped.

\( P(E_i|H) \) is estimated from the training data. \( P(E_i) \) and \( P(H) \) are prior probabilities that \( E_i \) will be observed and that \( H \) will be true, respectively. The prior probabilities were determined by sampling 10,000 LC records and Levy metadata records chosen uniformly at random to produce frequency distributions over each piece of evidence. The threshold was determined by manually evaluating the training data. Based on these results, we chose the threshold as the greatest probability that would not misclassify an N that did have an assignment as one that did not.

**Evidence**

The evidence used to determine the probability of a match between a name to an LC record
is a set of Boolean tests involving the name, the Levy metadata associated with that name, and the LC record.

The following fields were used by ANAC:

**Levy record:**

- **Given name**: often abbreviated
- **Middle names**: often abbreviated
- **Family name**
- **Modifiers**: titles and suffixes
- **Date**: publication year
- **Location**: publication location (city)

**LC record:**

- **Given name**: includes abbreviations
- **Middle names**: includes abbreviations
- **Family name**
- **Modifiers**: titles and suffixes
- **Birth**: year of birth
- **Death**: year of death
- **Context**: miscellaneous data

The tests used are: first name equality and consistency, middle name equality and consistency, music terms present in LC record context, name modifier consistency, Levy sheet music publication consistent with LC author birth and death, and Levy record publication location in LC record context. Two name fragments were considered equal if they featured the same string after being converted to lower case and having punctuation removed. Two name fragments were considered consistent if equal, or if their first letters were equal, or if one was an abbreviation of the other. The equality and consistency of a name fragments were treated as one piece of evidence because they are conditionally dependent. One advantage to the naïve Bayes classifier was that new tests could easily be added. It should be noted that "family name" is always equal because we use it to select the set to potential matches.

**Results**

In order to train the system, the Cataloging Department at the Sheridan Libraries generated ground truth data. For each name in 2,000 randomly selected Levy metadata records, catalogers recorded the authorized form of the name when a matching authority record was available. The entire process required 311 hours (approximately seven minutes per name). The human catalogers used much the same type of evidence as ANAC in establishing matches. Catalogers examined name similarity; compared publication dates from the Levy records to birth and death dates in the authority records; and examined authority record note fields for musical terms. In addition, the catalogers often searched for bibliographic records of other editions of a particular title to determine the authoritative name assigned to the subject.

From these 2,000 records there were 2,841 names, of which 1,878 had matching LC records. There were 795 names, or 28% of the total, that did not have an LC record. We excluded 168 names for which an LC record did not exist in ANAC's authority database.

ANAC was evaluated using ten-fold cross validation. The ground truth data was divided into ten equally sized sets, uniformly and randomly. For each set, ANAC was trained on that particular set and then evaluated against the remaining nine sets. For each trial we recorded four values:

1. the number of Levy names correctly matched to an authority record,
2. the number of names incorrectly matched to an authority record,
3. the number of names correctly identified as not having a matching authority record, and
4. the number incorrectly identified as not having a matching authority record when one was available.

The results from these trials were as follows:

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.58</td>
<td>0.00</td>
</tr>
<tr>
<td>Precision (LC record exists)</td>
<td>0.77</td>
<td>0.00</td>
</tr>
<tr>
<td>Precision (LC record does not exist)</td>
<td>0.12</td>
<td>0.00</td>
</tr>
</tbody>
</table>

These results were extremely consistent across all trials. Overall, ANAC was successful 58% of the time. When a name had an LC record, ANAC was successful 77% of the time, but when an LC record did not exist for a name ANAC was successful only 12% of the time. The reason for this discrepancy is that ANAC cannot learn whether or not a name has been added to the LC authority file.

It took ANAC five hours and forty-five minutes to classify the 2,673 (2,841 minus 168) names, or about eight seconds per name. The database-bound process of retrieving the candidate set of MARC records given a family name consumed most of this time.

**Observations and Conclusions**

From the outset of the project, we worried that our metadata was not clean enough to do name authority work. For example, we did not have names in separate metadata elements. As mentioned earlier, we needed to extract them from a free form statement of responsibility. Additionally, many of the fields, such as publication date and publication location, contained malformed data. Fixing these problems before completing the ANAC work was not practical given the size of the collection and our resource constraints. In spite of these problems, ANAC performed reasonably well.

Also from the beginning, it was never anticipated that ANAC would entirely replace the human effort, but it could be a valuable complement. ANAC is fast enough to be used while the cataloger works on other parts of the record. For example we could easily create a graphical interface that, given a name and Levy record, displays the best matches along with ANAC's level of confidence. Alternately, ANAC might be run ahead of time to annotate the Levy metadata with possible matches. A follow-up study to examine this type of integration would be a valuable point of comparison.

Using a smaller training set created earlier in the project (not the 2,000 record evaluation set), we noticed several instances of names that were classified as not having an LC record. Upon further examination, however, the best match seemed quite likely to be correct. As a follow-up effort to characterize human cataloging error, we could identify names to which ANAC assigns a high probability of matching a given LC record, but that were assigned to the "without an LC record" class. A cataloger could then investigate these records more carefully.

Our development of ANAC was motivated by efficiency concerns. But in our analysis we have not included the effort needed to develop ANAC. Not all institutions have the resources available to us. This is an important consideration since ANAC, as currently written, is not a generalizable tool. ANAC took advantage of music-specific information from the notes field in the LC authority file. This useful, but specific, provision works well for the Levy Collection given its musical content. Undoubtedly, such specification information would have less utility for collections without musical content. There may or may not be
domain-specific terminology available, depending on the collection.

With factors that would both enhance and inhibit ANAC's performance with other types of collections, rather than considering the specific aspects of each factor, it is worth proposing the idea that a framework for automated metadata generation could be viable. A precedent for this type of approach is the Gamera document analysis framework (Droettboom et al., 2002) developed by the Digital Knowledge Center.

This type of framework could be developed such that different classifiers or techniques, and different metadata fields from both the collection and LC authority file (or other canonical source), could be chosen to tune the automated metadata generation. The combination of customizable, automated metadata tools and strategic human intervention could increase both efficiency and accuracy of metadata generation.

On occasion, the metadata debates focus on the question of whether metadata is even necessary. Our experience with the Library of Congress' Archive Ingest Handling Test (AIHT) provides evidence that metadata, at least a core set of fields, serves to facilitate large-scale ingestion and subsequent access. The principles behind metadata remain valid and useful. However, it is worth questioning the practices associated with generating metadata and considering the creation of frameworks that combine both automated and human resources.

Acknowledgements

This research was supported through generous grants from the National Science Foundation (DLI-2 IIS9817430) and the Institute of Museum and Library Services (NLG LL90167). We would like to thank Karl MacMillan for developing the cataloging web application, and Marius Stans for performing most of the initial LC database checking. We also thank Jacquelyn Gourley for her assistance in preparing this article.

References


Copyright © 2004 Mark Patton, David Reynolds, G. Sayeed Choudhury, and Tim DiLauro