Automating the Construction of Authority Files in Digital Libraries: A Case Study

James C. French[*] Allison L. Powell
Department of Computer Science
University of Virginia
Charlottesville, Virginia 22903
{french|alp4g}@cs.virginia.edu

Eric Schulman
National Radio Astronomy Observatory[†]
520 Edgemont Road
Charlottesville, VA 22903-2475
eschulma@nrao.edu

Technical Report No. CS-97-02
January 14, 1997

Abstract

The issue of quality control has become increasingly important as more online databases are integrated into digital libraries. This can have a dramatic effect on the search effectiveness of an online system. Authority work, the need to discover and reconcile variant forms of strings in bibliographic entries, will become more difficult. Spelling variants, misspellings, translation and transliteration differences all increase the difficulty of retrieving information. This paper is a case study of our efforts to create an authority file for authors' institutional affiliations in the Astrophysics Data System. The techniques surveyed here for the detection and categorization of variant forms have broader applicability and may be used in authority work for other bibliographic fields.

1 Introduction

As the pace of electronic publication accelerates, there will be increasingly many online databases. The challenge is to integrate them into coherent digital libraries that let users have unimpeded access to accurate information. There will be increasing reliance on automated techniques to aid information providers as they seek to reach this goal.

In recent years there has also been an increasing emphasis on data quality in online databases [10]. One aspect of improving data quality is detecting variant names for unique entities in the database. This is called authority work [3] and results in the creation of authority files that maintain the correspondence between all the allowable forms for strings in a particular bibliographic field, say, author or journal name. In this paper we look at techniques to aid in detecting variant forms of strings in bibliographic databases.

Taylor [14] elucidates two principles of authority control. The first is that all variants of a name will be brought together under a single form so that once users find that form, they will be confident

[*] This work supported in part by Dept. of Energy Grant no. DE-FG05-95ER25254 and NSF grant CDA-9529253.

[†] The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.

that they have located everything relating to the name. The second ensures that a user will find a name if the catalog has it. Although Taylor's study casts doubt on the utility of the second principle, the first is declared to be the "absolutely indespensible part of authority control."[14, p. 15] It is this aspect of authority control that we are trying to support with the work described in this paper. More concretely, we are trying to automate this authority control mechanism to achieve a transparent facility having the characteristics described by Auld[3].

> A bibliographic record, together with all variant forms of each associated heading, would be entered into the system. The computer would establish linkages between the preferred forms of headings and the bibliographic record. When a user keyed in a known form of a heading, the system would follow the internal linkages and display the requested item even though the preferred form of the heading might be quite different from the form entered. [...] to the user it would appear that a direct linkage existed between the form of heading entered and the bibliographic record displayed. The authority control mechanism would be invisible so far as the user was concerned.[3, p. 327].

Problems arise when bibliographic databases are integrated. The different component databases might use different authority conventions. Users familiar with one set of conventions will expect their usual forms to retrieve relevant information from the entire collection when searching. Therefore, a necessary part of the integration will be the creation of a joint authority file in which classes of equivalent strings are maintained. These equivalence classes can be assigned a canonical form which, in principle, could be substituted for the original strings in the combined database. In practice, this will generally be impractical or impossible because of intellectual property constraints. So a mapping will have to be maintained between searchers and systems that hides this heterogeneity from users. Tools are needed to automate this process. Techniques that underlie effective tools are the topic of this paper.

In the remainder of this paper we describe a particular instance of authority work, the creation of affiliations for authors in the Astrophysics Data System (ADS). Although we describe the techniques within the framework of a particular application, it is clear that they are more broadly applicable to authority work for other bibliographic fields.

In the next section we describe the particular problem and outline our approach in Section 3. In Section 4 we present a series of experiments and results characterizing the effectiveness of the various techniques. We conclude with some recommendations and a discussion of future work.

2 The ADS Database — A Case Study

We are collaborating with astronomers at the Smithsonian Astrophysical Observatory who have provided us with bibliographic records from the Astrophysics Data System (ADS)[2]. The ADS is an extensive collection of bibliographic data, abstracts, and full text from astronomy and astrophysics journals and conference proceedings. It contains approximately 240,000 entries for articles from over 1000 journals and conference proceedings. Some of these sources are indexed beginning before the year 1900. Our experiments are performed on a subset of this database containing approximately 146,000 refereed articles.

The astronomy community collects statistics about publication in that field, tracking changes in measures such as paper length, general productivity and institutional productivity. Traditionally, such statistics have been gathered by hand on a necessarily small subset of available documents (e.g. [1, 15]). We have been collaborating with astronomers interested in automatically gathering this information using the ADS database as the data source [11, 12]. Automatically gathering statistics about these electronic documents has allowed a much larger fraction of documents to be considered — it has also presented new challenges. For example, while it is relatively simple to compute statistics such as number of papers per journal for each calendar year, statistics involving authors and their affiliations are more difficult. Before we can determine how many papers were written by an individual or by someone at a given institution, we must first be able to reliably identify that individual or institution. Since the ADS does not currently have an author affiliation list, we will need to derive it from the data. Errors and inconsistencies in the data make this challenging.

The larger problem with which we are faced is a lack of authority control in the ADS database. This problem exists in every bibliographic field of the ADS records — authors' full names or initials may be used; journal titles may or may not be abbreviated; author affiliations are used very inconsistently. We do not mean to imply that the ADS is unique is this regard. This is a very common situation in many online databases and exists primarily because the data often comes from a variety of sources and merging it consistently is so labor intensive.

The specific challenge facing us is partially illustrated by Figure 1. While a preferred naming scheme for affiliations exists, it is not consistently used. Institution names are recorded in a variety of formats and include a range of information — from terse, abbreviated names to full names with complete postal addresses. Figure 1 shows the different names that appear in the database for the University of Virginia and a count of the number of times each variant appears. This list of variants will be used as a running example to illustrate the effects of each of the approaches to data cleanup that we employ.

The example in Figure 1 shows 21 variants for what is obviously a single affiliation. These variants illustrate a subset of the causes of the inconsistencies: misspellings; permuted word order; common terms (such as University) may or may not have been abbreviated; and full addresses of sites were sometimes used, when they were, state names were sometimes abbreviated. This is merely the tip of the iceberg. In addition to the causes listed above and illustrated in the example, there are multiple other potential causes, including: acronyms used inconsistently; affiliation names which change over time; transliteration conventions vary; and many variants in translation when non-English language affiliation names are keyed by native English speakers.

In this paper, we present the results of our efforts to identify the unique institutions listed as author affiliations in the ADS database. Many of these approaches are also applicable to other bibliographic fields. We must also provide a way to map variants of an institution's name to the canonical name for that institution. This is necessary so that corrections can be made to the database and so that new entries can be verified or corrected before being added to the database. Our goal is to resolve the variants, determine the canonical set of sites and produce a mapping from the variants of the site names found in the database to the canonical set using predominately automated methods. Others have considered similar problems with variant forms in bibliographic fields, for example, author names [13] and titles [17]. Borgman[4] surveys many other name-matching algorithms.

This data cleanup effort will allow us to expand the scope of the statistics-gathering effort. It will also make it possible for the ADS to provide services that are currently infeasible. Such services include an index of all papers with authors from a given institution and a definitive list of all papers

Affiliation string	Number of occurrences
Univ. of Virgina, Charlottesville, VA, US	1
Univ. of Virginia, Charlottesvill, VA, US	1
Univ. of Virginia, Charlottesville, VA, US	44
Univ. of Virginia, Charlottsville, VA, US	1
Univ. of Virginia, VA, US	1
University of VA., Charlottesville	1
University of Virginia, Charlottesville, VA, US	23
University of Virginia, Charlottesville, Virginia, US	1
University of Virginia, Virginia, US	1
Virgina Univ., Charlottesville, VA, US	1
Virgina, University, Charlottesville, VA	1
Virginia Univ.	2
Virginia Univ., Charlottesville	58
Virginia Univ., Charlottesville, VA	1
Virginia Univ., Charlottesville, VA, US	4
Virginia University, Charlottesville	1
Virginia University, Charlottesville, VA	1
Virginia, University	57
Virginia, University, Charlottesville	204
Virginia, University, Charlottesville, VA	77
Virginia, University, Charlottesville, Va.	83

Figure 1: Raw affiliation strings for the University of Virginia

by a particular author.

3 Approach

We have taken a multi-stage approach to resolving variant forms of affiliations in this database. We currently apply a combination of lexical steps, clustering based on edit distance and manipulation of the affiliation strings. Many of our later approaches were motivated by observations about results from earlier experiments. We present partial descriptions of each approach in this section; in some cases these descriptions will be elaborated in the Section 4 where they can be described in the context of the results which motivated them. We will use a running example to illustrate the results of these approaches.

3.1 Lexical Cleanup

The lexical cleanup steps contain database and/or domain dependent aspects. For the specific problem reported in this paper, these domain-dependent activities are based upon general knowledge about variants on place names and specific observations about the ADS database, including the observations listed earlier. The lexical cleanup steps include removing extraneous information and expanding abbreviations and acronyms. In general, the activity of expanding abbreviations and acronyms or removing extraneous information is not domain dependent. However, the activity of identifying extraneous information and providing the expanded form of acronyms is domain dependent and may require the input of a domain expert.

The lexical cleanup steps were applied first and the results were used for most of the experiments that we describe in this paper. At a later point, we show a comparison of the results for lexically cleaned and non-lexically cleaned data to illustrate the usefulness of the lexical cleanup steps (see Table 2).

3.2 Edit Distance Clustering

We also noted that there were a significant number of unpredictable typographical errors in the collection. In addition, different transliteration conventions were used when translating Russian site names from the Cyrillic to the Latin alphabet. These inconsistencies could not be handled easily by lexical approaches. We chose to use an edit distance as a domain independent way to measure the difference between two strings. The edit distance[9, 16] is the number of insertions, deletions, transpositions and substitutions required to turn *string1* into *string2*. Edit distance has traditionally been used in approximate string matching [7], spelling error detection and correction [8], and more recently has been shown to be more effective than Soundex for phonetic string matching [18]. We used the edit distance algorithm presented by Hall and Dowling [7] as implemented by Zobel and Dart [18]. For each round of experiments, we computed an edit distance cost matrix which contained the distances between all affiliation strings in the current set. These distances were used to form affiliation clusters based on a fixed or variable edit distance threshold. The fixed edit distance threshold is some constant maximum value for the cost. The variable threshold caps the maximum cost at a fraction of the length of the shorter affiliation string. Hence, this cap may vary for each pair of strings compared. We will discuss the rationale for these values in the Section 4.

Our clustering procedure is currently order dependent. We work on the assumption that given a set of variants on an affiliation name, the ones which occur most frequently have the greatest chance of being correct (i.e. a specific typographical error will occur infrequently relative to the correct spelling). Therefore, affiliation strings are considered for clustering in the order of most to least frequently occurring. The most frequently occurring affiliation is considered first. It and all affiliation strings within the threshold limit distance from it are gathered into a cluster. These affiliations are then removed from further consideration and we progress to the next most frequently occurring affiliation string which has not yet become part of a cluster. The most frequently occurring affiliation string in a cluster is nominated as the cluster representative (i.e. the canonical form).

At all stages, we were intentionally conservative when forming clusters. Because some variants of affiliation names are extremely resistant to automatic resolution methods, human intervention will be necessary to form the final set of canonical affiliation names. We theorize that removing incorrectly placed items from a cluster will pose a greater burden to an individual than indicating that several smaller clusters belong together. In addition, cluster representatives are sometimes used as the affiliation set for later experiments. For accuracy, these representatives should be from extremely “clean” clusters — clusters which contain almost no erroneously placed members. Otherwise, the set of cluster representatives would be artificially small. Finally, our initial goal is to determine the canonical set of affiliation strings and a mapping from variants to the canonical set. For all of these reasons, conservative cluster formation was essential.

3.3 Affiliation String Manipulation and Word-based Matching

Finally, we noted that a general edit distance comparison of affiliation strings could not accurately capture the similarity of two strings if the words in one string were a permuted order of the words in the other. In addition, we also noted that some strings contained duplicate words. The duplicate words rarely added useful information. We therefore used an alternate internal representation of the affiliation string — the unique words of the string listed in lexically sorted order.

We will show that using the alternate representations of the affiliation string is helpful, but it does not allow fine control over the types of errors allowed. In a companion paper [6], we describe a new approach — approximate word matching — which finds a minimum distance matching between the words in *string1* and *string2*. We have performed experiments on a small subset of the collection using this new approach. The results are reported in [6]. Experiments using the full collection are forthcoming.

4 Results

4.1 Experiments

First, raw affiliations were extracted from the affiliation fields of the records in the ADS database. Lexical cleanup was then performed. These results are shown in Table 1. We show in Table 2 that it is effective to perform the lexical cleanup first, given the computationally expensive nature of many of the later steps. All further experiments are performed on the lexically cleaned affiliations.

An edit distance cost matrix was computed and clusters were formed using both fixed (see Table 2 for results) and variable thresholds (see Table 3 for results). Then the cluster representatives from the fixed threshold of one (EDT1) were used as a new affiliation set. A cost matrix was computed for the new affiliation set and these new affiliation strings were clustered using a variable threshold. These results are reported in Table 3 (both portions of Table 3 will be compared). Multiple sets of affiliation strings were then manipulated, producing the results shown in Tables 4 and 5. These results will also be compared below.

Cleanup Method (applied sequentially in the order listed)	**Number of Distinct Affiliations**	**Δ**
None	20168	
Remove US, U.S.A., etc. if occurring at the end of an affiliation	19868	300
Remove US ZIP codes from end of affiliations	19850	18
Remove US state abbrevs. occurring at the end of an affiliation	18850	1000
Expand most obvious abbreviations (University, Institute, etc.)	17773	1077
Expand other selected abbreviations and acronyms	17598	175
Remove country names occurring at the end of an affiliation	16427	1171

Table 1: Reduction in the number of distinct affiliations using lexical cleanup approaches

4.2 Lexical Cleanup

The lexical cleanup steps were applied in the order listed in Table 1. Most steps could be combined into a single operation. They are presented here separately so that the impact of individual steps can be assessed. Other steps are simpler to perform if done sequentially, but not impossible if performed as a single step. For example, many affiliations include postal addresses at the end of the affiliation string. Knowing this, we can strip country names, ZIP codes, and state abbreviations from the end of the affiliation string. By working from the end of the string, we can be more confident that we are not erroneously removing important information. For example, given an affiliation `University of VA., Charlottesville, VA, US`, it should be useful to remove `US` and the second `VA`, but removing the first `VA` would eliminate useful information.

Lexical cleanup pays two dividends. When the cleanup step is performed, some differences in affiliation strings are removed, allowing those affiliation variants to collapse together. In Table 1, the Δ column shows the difference between the current number of distinct affiliations and the number in the previous step. In some steps, the apparent immediate payoff is small. However, the second payoff is seen when the edit distance clustering is performed. Consider the following affiliations.

```
NASA, Goddard Space Flight Center
National Aeronautics and Space Administration, Godard Space Flight Center
```

Expanding the NASA acronym would not cause the strings to collapse together immediately, because of the spelling error in "`Goddard`" in the second string. However, because the difference in the strings is small after the expansion, these affiliations would be clustered together using even a very conservative edit distance threshold.

Lexical approaches provided dramatic improvements in a few cases. However, the best immediate results were seen when common abbreviations and acronyms were tackled. This approach rapidly reaches a point of diminishing returns. It simply is not efficient to correct infrequently occurring variants in this way.

Figure 2 shows the results of lexical cleanup on our example affiliation set. We now have 14 variants instead of 21. However, the remaining strings illustrate the types of affiliation variants that do not respond well to lexical measures, including spelling errors, word permutation, and affiliations which include city names.

4.3 Fixed Threshold Edit Distance Clustering

Our first approach to handling the variants which did not respond well to lexical cleanup was edit distance clustering. We first investigated fixed threshold edit distance clustering. The results of this approach are listed in Table 2. Table 2 also illustrates the usefulness of lexical cleanup in reducing the number of items to be considered in later phases. Note that the values reported for the lexically cleaned-up affiliations have a different starting point than the values reported for the raw affiliations. The lexically cleaned affiliations have had all of the operations described in Table 1 performed, reducing their starting count from 20168 to 16427.

Note that an edit distance of one provides significant improvement. We hypothesize that all of these variants are simple typing errors. We use the cluster representatives from these edit distance one (EDT1) clusters in further experiments. The clusters formed using an edit distance of two appeared to be "clean" as well; however, we chose to use the more conservatively created set for

Affiliation string	**Number of occurrences**
University of VA., Charlottesville	1
University of Virgina, Charlottesville	1
University of Virginia	1
University of Virginia, Charlottesvill	1
University of Virginia, Charlottesville	67
University of Virginia, Charlottesville, Virginia	1
University of Virginia, Charlottsville	1
University of Virginia, Virginia	1
Virgina University, Charlottesville	1
Virgina, University, Charlottesville	1
Virginia University	2
Virginia University, Charlottesville	65
Virginia, University	57
Virginia, University, Charlottesville	364

Figure 2: Lexically corrected affiliation strings for the University of Virginia

future experiments. The upper portion of Figure 3 illustrates the clusters formed in our example using an edit distance of one.

	Lexically Cleaned-up		**Raw**	
Edit Distance Threshold	**Number of Clusters**	Δ	**Number of Clusters**	Δ
1	13527	2900	17226	2942
2	12786	741	16357	869
3	12160	608	15665	692
5	10924	1236	13554	2111

Table 2: Fixed threshold edit distance clustering using lexically cleaned up affiliation set (16427 affiliations) and raw affiliation set (20168 affiliations)

Increasing the edit distance threshold allows more affiliations to be clustered together and decreases the number of clusters. However, as illustrated by the lower portion of Figure 3, this approach breaks down for even moderate thresholds. For our example, using an edit distance of five causes clusters to form which contain variants of different affiliations. It is apparent that fixed threshold edit distance clustering alone will not be enough to group affiliation variants correctly. However, it is reasonable to assume that multiple typing errors or systematic transliteration variants could occur in a string. Our next approach is intended to handle this more effectively than fixed threshold clustering does.

4.4 Variable Threshold Edit Distance Clustering

As we noted earlier, it is reasonable to assume that more than one typing error can occur in a string. Minor variants in affiliation names, for example liberal use of commas, could also manifest

Affiliation string	**Number of occurrences**
Virginia, University, Charlottesville	430
Virginia, University, Charlottesville	364
Virgina, University, Charlottesville	1
Virginia University, Charlottesville	65
University of Virginia, Charlottesville	70
University of Virginia, Charlottesville	67
University of Virgina, Charlottesville	1
University of Virginia, Charlottsville	1
University of Virginia, Charlottesvill	1
Virginia, University	59
Virginia, University	57
Virginia University	2
Virgina University, Charlottesville	1
University of Virginia	1
University of Virginia, Virginia	1
University of Virginia, Charlottesville, Virginia	1
University of VA., Charlottesville	1
Victoria, University	*224*
Victoria, University	*139*
Victoria University	*26*
Virginia University	*2*
Virginia, University	*57*
University of Arizona	*3*
University of Arizona	*2*
University of Virginia	*1*

Figure 3: Fixed threshold edit distance 1 (upper) and edit distance 5 (lower—errors only) clusters using lexically corrected affiliations

themselves as a small edit distance greater than one. However, as illustrated in the lower portion of Figure 3, even a moderate fixed edit distance threshold can be inappropriate in many cases, especially when short affiliation strings are involved.

Variable threshold edit distance clustering is a way to keep stricter control over these short affiliation strings while allowing slightly more leeway for the longer ones. Assuming that two affiliation strings differ by an edit distance of 5, the two strings are more likely to be variants of the same affiliation if they both have more than 90 characters than if they both have fewer than 20 characters. Defining the threshold as a fraction of the shorter string length allows us to take this into consideration. We define the threshold as a fraction of the shorter string to protect shorter strings from being clustered indiscriminately. The results of variable threshold edit distance clustering of the fixed threshold one (EDT1) cluster representatives are shown in Table 3.

The upper portion of Figure 4 shows the clusters formed by our example using an edit distance of 1/10 of the length of the shorter affiliation string.

	EDT1 representatives		**Lexically Cleaned-up**	
Edit Distance Threshold	**Number of Clusters**	Δ	**Number of Clusters**	Δ
1/10	11825	1702	11872	4555
1/9	11595	230	11641	141
1/7	10888	707	10926	725
1/5	9542	1346	9583	1343

Table 3: Variable threshold edit distance clustering of cluster representatives from EDT1 clusters in Table 2 (13527 affiliations) and of lexically cleaned affiliations (16427 affiliations)

4.5 Analysis of Edit Distance Clustering

In Table 3, we also show the results of performing a variable threshold clustering on the lexically cleaned-up set of affiliations, without the intermediate fixed threshold clustering step. We note that for our example the same affiliations cluster together using both methods. However, comparing the two portions of Table 3, we note that while the number of clusters at each threshold is similar, it is not identical. This is a result of our current order-dependent clustering methodology. The fixed threshold (EDT1) cluster representatives are given a count value equal to the sum of the variants in the cluster which they represent. Therefore, affiliations are considered in different orders when creating the variable threshold clusters reported in the two portions of Table 3. Performing the fixed threshold clustering first results in fewer variable threshold clusters.

The lower portion of Figure 4 illustrates that the problem of incorrectly clustering affiliations still exists, but at higher threshold levels. Examining this figure, it becomes apparent that variants of the same affiliation still differ significantly. Raising the edit distance threshold to a level that would cluster these variants together would also incorrectly cluster a significant portion of the affiliation set. We therefore consider other representations of the affiliation strings.

4.6 Affiliation String Manipulation

Many of the remaining affiliation variants were due to word permutation and duplicate words in the affiliation strings. To handle this problem, we created an alternative string representation. The representation consists of the unique words of the affiliation string in lexically sorted order. Note that this approach also provides a double payoff. Multiple affiliation strings can have the same representation, reducing the number of representations to consider. In addition, the sorted word order and lack of duplicates can reduce the edit distance between variants of the same affiliation. Table 4(a) shows the results of this activity and Figure 5 shows its effect on our example.

In addition, we noted that this approach and the standard variable threshold edit distance clustering approach reported in Table 3 attacked the problem from different directions. We then theorized that applying the two approaches sequentially would prove more effective than either was individually. This was, in fact the case. The results are shown in Table 5(a) and for our example, produced the same clusters as shown in Figure 5.

Affiliation string	Number of occurrences
Virginia, University, Charlottesville	431
Virginia, University, Charlottesville	430
Virgina University, Charlottesville	1
University of Virginia, Charlottesville	70
Virginia, University	59
University of Virginia	1
University of Virginia, Virginia	1
University of Virginia, Charlottesville, Virginia	1
University of VA., Charlottesville	1
Virginia, University, Charlottesville	431
Virginia, University, Charlottesville	430
Virgina University, Charlottesville	1
Victoria, University	
Victoria, University	*165*
Virginia, University	*59*
Pretoria, University	*1*
University of Virginia, Charlottesville	70
University of Virginia	1
University of Virginia, Virginia	1
University of Virginia, Charlottesville, Virginia	1
University of VA., Charlottesville	1

Figure 4: Variable threshold edit distance clusters for 1/10 (upper) and 1/5 (lower) length of shortest affiliation string using EDT1 representatives from Table 2

4.7 Word Extraction Improvements

After the completion of the experiments reported in Tables 4(a) and 5(a), we refined the way in which we extracted individual words from raw affiliation strings. These words are used to create the unique-word representation strings and will also be used in future experiments. In the majority of affiliation strings, words are delimited by spaces and/or commas. In our original word extraction method, only these delimiters were used to identify words. However, some string variants contained parenthesized phrases, quoted phrases, words delimited by slashes and extraneous punctuation. The refined word-extraction procedure took this into account. We also converted all words to lower case and removed numeric "words" (which were typically mail stops and postal codes). We then repeated the experiments reported in Tables 4(a) and 5(a). The results are shown in Tables 4(b) and 5(b) respectively. When comparing the tables, note that the new word extraction technique produces a smaller number of initial affiliations in both cases.

5 Recommendations

We have shown that when used together, the approaches outlined in this paper are a useful component of semi-automatic authority file generation. However, some of the approaches are motivated by particularly baroque variants found in the ADS database and may not be necessary for all applications.

	Original Extraction (a)		**New Extraction (b)**	
Edit Distance Threshold	**Number of Clusters**	Δ	**Number of Clusters**	Δ
1/10	11022	1487	10719	1493
1/9	10792	230	10484	235
1/7	10148	644	9844	640
1/5	8984	1164	8677	1167

Table 4: Variable threshold edit distance clustering of unique-word affiliations created from EDT1 representatives using (a) original word-extraction technique (13527 affiliations collapse to 12509) and (b) new word-extraction technique (13527 affiliations collapse to 12212).

Affiliation string	**Number of occurrences**
Charlottesville University Virginia	502
Virginia, University, Charlottesville	430
University of Virginia, Charlottesville	70
University of Virginia, Charlottesville, Virginia	1
Virgina University, Charlottesville	1
University Virginia	59
Virginia, University	59
University Virginia of	2
University of Virginia, Virginia	1
University of Virginia	1
Charlottesville University VA. of	1
University of VA., Charlottesville	1

Figure 5: Variable threshold edit distance 1/10 clusters using unique-word manipulation of EDT1 cluster representative affiliations from Table 2.

First, note that these approaches may not be appropriate for all application areas. These approaches are based upon the assumption that strings which differ by a small amount might be similar to one another or variants of one "correct" string. Given a situation where many unrelated strings differ by a very small amount, these approaches would not be appropriate.

Using a small fixed edit distance threshold is useful primarily for reducing the number of items to be considered by other, more computationally expensive, approaches. However, given a data set which contains few items and is known to contain relatively few spelling errors and no word permutations, variable threshold edit distance clustering alone may be a sufficient tool to bring together related items.

Using sorted-word string representations or word-set representations is appropriate when multiple spelling errors and word permutations exist.

	Original Extraction (a)		**New Extraction (b)**	
Edit Distance Threshold	**Number of Clusters**	Δ	**Number of Clusters**	Δ
1/10	10557	419	10352	514
1/9	10382	175	10165	187
1/7	9808	574	9580	585
1/5	8755	1053	8514	1066

Table 5: Variable threshold edit distance clustering of unique-word affiliations created from EDT1/10 representatives from Table 3 using the (a) original word extraction technique (11825 affiliations collapse to 10976) and the (b) new word extraction technique (11825 affiliations collapse to 10866).

6 Future Work

So far, we have only managed to cut the number of affiliations strings in half. There are still gains to be made by automated methods before human intervention is needed. However, it is apparent from Table 5 that we have reached a point of diminishing returns using our current methodology. The next step is to perform comparisons between strings on a word-by-word basis as described in [6]. This presents a number of interesting questions, the most important of which is determining an appropriate distance measure. We propose that it should contain components of total edit distance and individual edit distances between words. Note that individual between-word edit distances must be below some threshold.

We must obviously perform extensive evaluation of cluster content. We must also determine if the representatives of the clusters that we have formed are the canonical set of affiliations or if they are even close. Given this set, manual clustering will be necessary to produce the final canonical set of affiliation strings. As we noted earlier, our efforts so far have been aimed at creating a canonical set and therefore have been very conservative. Given a canonical set, we can apply more aggressive automated approaches. The results of such approaches could then be compared to the canonical set.

7 Conclusion

This has been a report of preliminary work on the problem of data cleanup in a working database. We have attacked several causes of inconsistencies in the affiliation fields of the bibliographic records in the ADS database. There is still much work to be done.

There are many opportunities to exploit existing online databases as new techniques are developed in the field of information retrieval. Before this evolution can take place, we need to find ways to improve the quality of the data and to make it easier for data providers to integrate new information into their systems. This paper has discussed techniques for improving data quality and data access by detecting variable forms of strings and collecting them together under a standard form. This can be thought of as the semi-automatic generation of an authority file. Our procedure still requires manual intervention, but after we generate authority files, we can create classifiers to reduce the burden of detecting variant forms as new data is acquired by a system.

We are currently applying these methods to the bibliographic data in the Astrophysics Data System (ADS) and the Networked Computer Science Technical Report Library (NCSTRL)[5] databases, two real-world online databases. The methods described in this paper were used to collect the 21 unique affiliation variants of Figure 1 into the four clusters shown in Figure 5. These are high quality clusters, containing no misclassifications, that are used in the second stage, manually supervised, clustering phase of our methodology. The methods are effective and efficient enough to be used in production environments. In the study reported here, we extracted 20,168 unique affiliation strings from our sample of 146,000 records. Preliminary results from the application of approximate word matching [6] show that we are able to further reduce these to 8,928 strings by using extremely conservative thresholds. Although we do not yet have a quantative measure of the misclassification rate, our qualatative assessment is that the results are excellent. The error rate is so low that we could choose to ignore it.

We agree with Siegfried and Bernstein that "as data files grow in size and as information is exchanged the role of automated merging devices will become increasingly important."[13, p. 220] One of our goals is to automatically generate auxiliary access structures for browsing online databases. The techniques described in this paper represent necessary steps to that end.

Acknowledgement. We would like to thank Stephen S. Murray, Guenther Eichhorn and Michael J. Kurtz of the Smithsonian Astrophysical Observatory for providing us access to the data in the Astrophysics Data System (ADS).

References

[1] H. A. Abt. Institutional Productivities. *Publications of the Astronomical Society of the Pacific*, 105:794–798, 1993.

[2] A. Accomazzi, G. Eichhorn, M. J. Kurtz, C. S. Grant, and S. S. Murray. The ADS Article Service Data Holdings and Access Method. In G. Hunt and H. Payne, editors, *Astronomical Data Analysis Software and Systems VI*, A.S.P. Conference Series, 1997. in press.

[3] L. Auld. Authority Control: An Eighty-Year Review. *Library Resources & Technical Services*, 26:319–330, 1982.

[4] C. L. Borgman and S. L. Siegfried. Getty's Synoname and its Cousins: A Survey of Applications of Personal Name-Matching Algorithms. *Journal of the American Society for Information Science*, 43(7):459–476, 1992.

[5] J. R. Davis. Creating a Networked Computer Science Technical Report Library. *D-Lib Magazine*, Sept. 1995.

[6] J. C. French, A. L. Powell, and E. Schulman. Applications of Approximate Word Matching in Information Retrieval. Technical Report CS-97-01, Department of Computer Science, University of Virginia, January 1997.

[7] P. A. V. Hall and G. R. Dowling. Approximate String Matching. *Computing Surveys*, 12(4):381–402, Dec. 1980.

[8] K. Kukich. Techniques for Automatically Correcting Words in Text. *Computing Surveys*, 24(4):377–440, Dec. 1992.

[9] R. Lowrance and R. A. Wagner. An Extension of the String-to-String Correction Problem. *Journal of the ACM*, 22(2):177–183, Apr. 1975.

[10] E. T. O'Neill and D. Vizine-Goetz. Quality Control in Online Databases. *Annual Review of Information Science and Technology*, 23:125–156, 1988.

[11] E. Schulman, J. C. French, A. L. Powell, S. S. Murray, G. Eichhorn, and M. J. Kurtz. The Sociology of Astronomical Publication Using ADS and ADAMS. In G. Hunt and H. Payne, editors, *Astronomical Data Analysis Software and Systems VI*, A.S.P. Conference Series, 1997. in press.

[12] E. Schulman, A. L. Powell, J. C. French, G. Eichhorn, M. J. Kurtz, and S. S. Murray. Using the ADS Database to Study Trends in Astronomical Publication. In *Bulletin of the American Astronomical Society*, volume 4, 1996.

[13] S. L. Siegfried and J. Bernstein. Synoname: The Getty's New Approach to Pattern Matching for Personal Names. *Computers and the Humanities*, 25(4):211–226, 1991.

[14] A. G. Taylor. Authority Files in Online Catalogs: An Investigation of Their Value. *Cataloging & Classification Quarterly*, 4(3):1–17, 1984.

[15] V. Trimble. Postwar growth in the length of astronomical and other scientific papers. *Publications of the Astronomical Society of the Pacific*, 96:1007–1016, 1984.

[16] R. A. Wagner and M. J. Fischer. The String-to-String Correction Problem. *Journal of the ACM*, 21(1):168–173, Jan. 1974.

[17] M. E. Williams and L. Lannom. Lack of Standardization of the Journal Title Data Element in Databases. *Journal of the American Society for Information Science*, 32(3):229–233, May 1981.

[18] J. Zobel and P. Dart. Phonetic String Matching: Lessons from Information Retrieval. In *Proc. 19th Inter. Conf. on Research and Development in Information Retrieval (SIGIR'96)*, pages 166–172, Aug. 1996.