Data Cleaning and Matching of Institutions in Bibliographic Databases

Jeffrey Fisher[1] **Qing Wang**[1] **Paul Wong**[2] **Peter Christen**[1]

[1] Research School of Computer Science
[2] Office of Research Excellence, The Australian National University,
Canberra, ACT 0200
Email: Jeffrey.Fisher@anu.edu.au

Abstract

Bibliographic databases are very important for a variety of tasks for governments, academic institutions and businesses. These include assessing research output of institutions, performance evaluation of academics and compiling university rankings. However, incorrect or incomplete data in such databases can compromise any analysis and lead to poor decisions and financial loss. In this paper we detail our experience with an entity resolution project on Australian institution data using the SCOPUS bibliographic database. The goal of the project was to improve the entity resolution of institution data in SCOPUS so it could be used more effectively in other applications. We detail the methodology including a novel approach for extracting correct institution names from the values of one of the attributes. Along with the results from the project we present our insights into the specific characteristics and difficulties of the Australian institution data, and some techniques that were effective in addressing these. Finally, we present our conclusions and describe other situations where our experience and techniques could be applied.

Keywords: Data Matching, Bibliographic Databases, Deduplication, SCOPUS.

1 Introduction

Bibliographic databases are being used across an increasingly broad range of areas. From allocating research funding by governments, to quantifying connections between academics and institutions to determining academic promotions (Christen 2012). To support these applications, it is vital that bibliographic databases are correct, cleaned and well maintained. However, far too often, it is up to individual companies or researchers to enter their own work into these databases (Lee et al. 2007). Alternatively, many bibliographic databases are automatically created and updated which leads to a host of data integrity problems including multiple updates, missing entries, and differences in data quality and formats when drawing on different data sources (Lee et al. 2007). All these problems can compromise the quality of any analysis done on the databases, which can lead to poor decision making and the wasting of time and money.

In this paper we detail our experience and findings from a project attempting to improve the data quality of Australian institutions in the SCOPUS bibliographic database (Scopus 2009). We used a variety of established data cleaning and data matching techniques and refined them where necessary. We present an approach to extracting institution names from attribute values. We also capture and incorporate domain specific knowledge, and illustrate particular types of problems for data matching in bibliographic databases. While we developed our approach for a specific database, certain techniques and aspects of the domain knowledge could be generalised to other bibliographic databases, and potentially other application areas.

The structure of this paper is as follows: in Section 1 we provide some background on the applications of bibliographic databases and the project goals. In Section 2 we describe the main features of the SCOPUS bibliographic database that was used in this project. In Section 3 we examine data cleaning and we describe our technique for extracting institution names from the databaseand in Section 4 we discuss the two aspects of the data matching in the project, merging institution identifiers where they correspond to the same institution, and determining an institution identifier for records that do not have one. Finally, in Section 5 we present our conclusions, a discussion of other areas these techniques could prove useful, and some possibilities for extending this work.

1.1 Bibliographic Databases

Bibliographic databases such as SCOPUS (Scopus 2009) have a wide variety of applications for governments, academic institutions and businesses. Governments use them for policy development including assessing future areas of research need and allocating research funding. They also use them to evaluate research and program performance. For example, in Australia, the Commonwealth Government runs the Excellence in Research for Australia (ERA) program to assess the research performance of academic institutions. The ERA program relies on measures such as citation counts and article counts from the SCOPUS database (ERA 2012). The ERA program also determines the funding allocations for part of the Sustainable Research Excellence in Universities program (ERA 2012).

Academic institutions such as universities also use bibliographic databases for a wide variety of tasks. Analysis of collaboration data in bibliographic databases allows universities to develop strategic partnerships and assists in identifying research and

funding opportunities. Additionally, performance evaluation of academic personnel can also be based on research output. This can be captured through measures such as the h-index, which attempts to quantify research output and quality, and is calculated from citations counts in bibliographic databases (Hirsch 2005).

Bibliographic databases are also used for commercial applications. An example is the industry that has developed around ranking universities, and scales such as the Times Higher Education ranking (`http://www.timeshighereducation.co.uk`) rely on data from bibliographic databases such as Thomson Reuters Web of Science to create their rankings.

1.2 Project Goals

Given the variety of applications of bibliographic databases it is important the data quality within them is high. To that end, the goals of this project were twofold. Firstly, it was to improve the accuracy of institution identifiers within the SCOPUS database. Secondly, this project formed part of an ongoing program of work to allow the Office of Research Excellence at the Australian National University (ANU) to better understand the SCOPUS database, and build organisational capacity for analysing and applying the data. These two goals could have a variety of practical benefits. For example, in order to analyse collaborations between institutions it is important that the institutions themselves be correctly identified. Similarly, if institutions are incorrectly identified when assessing research performance, then this may result in an inequitable allocation of funding. Improving the quality of institution data would assist with these and other applications and the lessons learned from this project could assist with similar analysis in the future.

SCOPUS contains a variety of different institution types. Large academic institutions such as universities produce the majority of research articles in SCOPUS, but there are many smaller research labs, companies, government departments, and even private individuals who also conduct research. Since the larger institutions are more relevant in most analysis, we only considered institutions that had at least ten records in SCOPUS.

In some cases it is unclear exactly what constitutes an institution. For the purposes of this project, we generally use the highest level organisation as the institution. For example, each separate division of a university could potentially be considered an institution, however the goal of the project was to identify all such divisions with the university itself.

2 The SCOPUS Database

The SCOPUS database is a bibliographic database containing approximately 80 million author records, and approximately 40 million academic papers, conference proceedings, theses, and other works. The snapshot used in this project covers the period 1996 to 2011. The SCOPUS database is stored in XML format with a tree schema and with a paper or article as the root of each record. The schema is proprietry so cannot be provided here. For ease of analysis, portions of the data were extracted into relational database tables containing information about entities such as authors or papers, but this does not reflect the underlying structure of the SCOPUS database. In addition, SCOPUS uses an automated data collection process which sources data from many places and in a variety of formats. This can lead to variations in data and storage formats. For example, the attributes *city*, *state* and *postcode* were all completely blank for the Australian data, despite the fact that they are appropriate. This information is often present in a record, but is usually part of the *organization* attribute, along with other information.

Since we were primarily interested in the Australian portion of the data, the *country* attribute was used to separate the data and only records with a *country* value of "aus" (correspoding to Australia), were used in the project. This reduced the dataset to 1,611,172 records. A short summary of the attributes and characteristics for the Australian records is provided in Table 1 below. The completeness column describes the percentage of records that had a non-null value for the attribute.

Attribute	Unique Values	Completeness
afid	28,306	98.2%
dptid	52,879	73.7%
organization	238,460	99.3%
city group	28,889	95.3%
address	32,038	25.1%
city	0	0.0%
state	0	0.0%
postcode	0	0.0%

Table 1: Characteristics of the SCOPUS Database.

The overall data quality was reasonably good, and the presence of the *afid* attribute, which appeared to be an identifier, was promising. However the large number of missing values reduced the usefulness of many of the attributes. From a data matching perspective, the most useful attributes appeared to be *afid*, *country* and *organization*. We examined each of these in more detail.

2.1 Attribute *afid*

According to the SCOPUS documentation (Scopus 2009), *afid* is intended to be a unique identifier for institutions. However, there were many missing values so it was not a strict primary key in the relational database sense (Elmasri & Navathe 2011). In addition, while its description indicated that it was unique, there were institutions in the database that had multiple different *afid* values.

We found that a single institution having multiple *afid* values was more common in smaller institutions. Given that bibliographic databases are important for the funding of large research institutions, they are more likely to have internal personnel making sure that research papers and journals are present in SCOPUS and correctly attributed to their institutions. In the event they determine that papers are missing or incorrectly attributed, they can notify SCOPUS to get the problem rectified. Smaller institutions are probably less likely to take these steps.

The *afid* attribute was extremely important for the data matching process. However, since it was not a perfect key and there was not a unique institution name for each *afid*, extracting a single institution name from the records for each *afid* was a significant challenge.

2.2 Attribute *organization*

From a data matching perspective this attribute was the most useful in determining the correct name for each institution. Of the attributes containing information about institutions, *organization* is the most complete with over 99% of Australian records having a value in this attribute. In addition, candidates for the institution name were often included in this attribute. It was also the best candidate for string comparison and other data matching techniques. However, the values of the *organization* attribute contained many abbreviations and acronyms that needed to be dealt with prior to the data matching as will be discussed in Section 3.3.

2.3 Attribute *country*

As described above, we used the *country* attribute to limit the data to manageable quantities. In addition, collecting all records with the same value of *country* together meant that some of the problems of trying to match across languages were removed. However, even something as straightforward as country was still not perfect. For example, there were hundreds of records with a *country* value for Australia that pertained to institutions in Austria, such as the University of Innsbruck and the University of Salzburg.

2.4 Data Summary

In summary, the most important attributes for the project were *country*, which was used to limit the data, *afid*, which was a quasi-identifier for institutions, and *organization* that contained text information about the institutions. Of the selected attributes, *organization* needed the most pre-processing to be useful in the data matching process. This involved extracting the institution names in an automated fashion as well as dealing with abbreviations and acronyms. This data cleaning process is dealt with in the next section.

3 Data Cleaning

Data cleaning is an important aspect of almost all data matching exercises (Han et al. 2012). The main objective in the data cleaning step was to extract from the 1,611,172 records for Australian institutions a unique institution name for each of the 2,910 *afid* values with more than 10 records in SCOPUS. In addition, acronyms, abbreviations and stop words could make data matching more difficult. As part of the data cleaning process we generated a list of possible acronyms and abbreviations for Australia, along with their likely expansions. Each of these data cleaning tasks is discussed separately.

3.1 Institution Names

One of the biggest challenges for the project arose from the fact that although the attribute *afid* was intended to uniquely identify an institution, there was not a one-to-one relationship between *afid* and institution name. However, for *afids* with many individual records (10 or more) a large part of the variability amongst values of the *organization* attribute often came from capturing subdivisions of the institution. As a result, the institution name appeared to be the most frequently occurring substring. This formed the basis for our approach to extracting institution names.

3.2 Research Hypothesis

Building on some of the ideas presented by Ciszak (Ciszak 2008) and after experimentation, we determined the following hypothesis: the most frequently occurring comma-separated substring within the values of the *organization* attribute was the best candidate for the correct institution name. The ratio between the frequency of the most common substring and the second most common substring was an indicator of the likelihood that the name was correct.

To illustrate this we provide the following example which shows the name extraction process and ratio calculation for the ANU. The first step was to select the 63,389 records that had the *afid* value for the ANU. For each of these records we separated the values of the *organization* attribute into comma-separated substrings and these substrings were counted to create a frequency table. We present the four most frequent comma-separated substrings for the ANU in Table 2 below:

Substring	Frequency
Australian National University	29,311
Research School of Chemistry	5,483
Research School of Physical Sciences and Engineering	4,485
John Curtin School of Medical Research	4,447

Table 2: Substring Frequencies for the Australian National University.

The most frequently occurring substring was used for the institution name, which in this case was "Australian National University" as expected. The ratio was then calculated by the following formula:

$$Ratio = Highest\ Freq./2nd\ Highest\ Freq.$$

For the above example, this gives:

$$Ratio = 29,311/5,483 = 5.35$$

A ratio of 5.35 was relatively high and as a result we had good confidence that the name was correct. We provide some examples of extracted institution names in Table 3 below. We also provide a more detailed analysis of the relationship between name correctness and the ratio value in Section 3.5.

3.3 Acronyms and Abbreviations

In addition to extracting institution names, it was important to replace frequent acronyms and abbreviations by their expanded expressions. This improved the results of the institution name extraction, especially where the acronyms were common. In addition, during the subsequent data matching, we needed to match against common acronyms of large institutions when trying to deal with records that had no *afid* value, in case they only contained the acronym in their value for the *organization* attribute.

We used a look-up table that specified the most common acronyms along with their expanded forms. However, particularly amongst the abbreviations, there were many that had multiple possible expansions, for example "Med." could be "Medical" or "Medicine". Because of this, there were several common abbreviations that were left in their unexpanded form.

Extracted Name	Correct Name	Ratio	Notes
Australian Institute Marine Science	Australian Institute of Marine Science	18.05	None
Royal North Shore Hospital	Royal North Shore Hospital	11.98	None
Calvary Hospital	Calvary Hospital	7.00	Kogarah, N.S.W.
University Queensland	University of Queensland	6.05	None
Western Australian Institute Sport	Western Australian Institute of Sport	4.67	None
URS Australia Pty Limited	URS Australia Pty Limited	3.50	Contains Acronym
School Chemistry	Monash University	2.60	Incorrect Name
Ipswich Hospital	Ipswich Hospital	1.80	None
University Notre Dame	University of Notre Dame	1.35	None
School Psychiatry	Unknown	1.21	Incorrect Name
Innsbruck Medical University	Innsbruck Medical University	1.14	In Austria
Suite 3	Melbourne Heart Centre	1.00	Incorrect Name

Table 3: Examples of Extracted Institution Names and Ratio Values.

3.4 Stop Words

We also removed stop words such as "the", "of" and "for". Because these words contain little information, they are sometimes left out (Christen 2012) and removing them further standardised the institution names. As with abbreviations and acronyms, a lookup table was used to remove them from the values of the *organization* attribute.

3.5 Data Cleaning Results and Discussion

In this section, we present the overall results and analysis of the data cleaning. The main focus of the data cleaning was extracting a unique name for each *afid* value. We also examined whether the calculated ratio value influenced the likelihood that a name was correct.

To test the methodology, a random selection of 200 *afid* values were picked and we used our approach to extract an institution name. All of the *afid* values had at least 10 records in SCOPUS. To deal with acronyms and abbreviations, we used a manually created lookup table containing 73 acronyms and abbreviations along with their expansions. They were predominantly the most frequently occurring acronyms and abbreviations where there was little ambiguity as to what the correct expansion was. We also removed the most common stop words ("of", "and", "for", "the", "in", "at" and "on"). The results are displayed in Table 4 below:

Result	Number of *afid*s	Proportion
Correct Name	131	65.5%
Partially Correct Name	42	21.0%
Incorrect Name	27	13.5%
Total	200	100.0%

Table 4: Data Cleaning Results.

Names were judged to be correct if they could identify the institution in question. Names were judged partially correct if they contained abbreviations or acronyms that had not been expanded or were missing a word. In general, where the name was correct, but was for a smaller part of a large institution, we deemed it an incorrect result. For example one *afid* was assigned the name "Research School of Social Sciences" but this was deemed incorrect since it was a subdivision of the Australian National University. The only exceptions were if it appeared to be a separate research centre or similar, in which case we deemed it partially correct. There was a certain level of judgment involved, however these cases were fairly few in number. For an approach that was simple and easy to implement, the overall results were reasonably promising with a correct or partially correct name extracted for 86.5% of the *afid* values.

In order to test the hypothesis that the ratio between the two most common substrings was an indicator of how likely a name was to be correct, we conducted a more detailed analysis. The results of the 200 sampled *afid* values were broken into categories based on the calculated ratio value. Table 5 below shows how the ratio affects the quality of the names generated. We counted partially correct matches as correct for this analysis.

Ratio	% Correct	% Incorrect
Ratio ≥ 4.0	98.3%	1.7%
$2.0 \leq$ Ratio < 4.0	86.9%	13.1%
$1.5 \leq$ Ratio < 2.0	88.0%	12.0%
$1.2 \leq$ Ratio < 1.5	85.0%	15.0%
$1.1 \leq$ Ratio < 1.2	80.0%	20.0%
Ratio < 1.1	60.0%	40.0%
Total	86.5%	13.5%

Table 5: Effect of Ratio on Name Correctness.

As predicted by our hypothesis, a higher ratio was a general indicator of name correctness. However, there was a significant difference between ratio values above 4.0 and ratio values below 4.0. In applications where incorrect matches would be a significant problem, then excluding everything with a ratio below 1.1 or 1.2, would reduce the incorrect names without massively lowering the coverage. For this project we retained the names extracted for all 2,910 *afid* values that had more that 10 records in SCOPUS.

We present some possible ways of improving this methodology in Section 5, along with other situations where this technique could be applied.

4 Data Matching

Data matching is the task of identifying, matching, and merging records that correspond to the same entities from one or more databases (Christen 2012). For this project, there were two main parts to the data matching process: determining which different *afid* values corresponded to the same institution and could be merged, and assigning an *afid* value to records that did not have one.

The data matching was performed using different string comparison techniques on the 2,910 institution names that were extracted during the data cleaning. For each step of the data matching, we set a minimum similarity threshold. Comparisons that returned a similarity score above this threshold were *positive matches*. Comparisons that returned a similarity score below this threshold were *non-matches*.

The positive matches fell into two categories, *true positives* and *false positives*. Matches were *true positives* if the two values matched actually corresponded to the same real world institution. Matches were *false positives* if the two values corresponded to different real world institutions. In order to determine whether matches were true positives or false positives, we conducted a manual evaluation. Non-matches could also be divided into *true negatives* and *false negatives*. However for non-matches the vast majority were true negatives and so we were unable to manually review a sufficient number of non-matches to accurately estimate the number of true negatives and false negatives.

Since we could not determine the actual number of true negatives and false negatives we were unable to calculate recall and accuracy. As a result, we used precision to evaluate the data matching results. Precision is calculated as follows (Han et al. 2012):

$$Precision = true\ positives / positive\ matches$$

True positives and *positive matches* are as defined above.

Transitive closure was also a potential problem. Transitive closure refers to the situation where if three records, "a", "b" and "c" are compared to each other pair-wise and "a" matches to "b" and "a" matches to "c" then "b" should also match to "c" (Christen 2012). However, in practice this is not guaranteed and for this project it was an issue we had to resolve. We dealt with this slightly differently for each part of the data matching so it is discussed in the next sections.

A number of different string comparison techniques were used in the data matching. Each technique takes two strings as input and returns a similarity value between 0 and 1. A result of 0 indicates the strings are completely different (what constitutes completely different varies depending on the technique). Higher values between 0 and 1 indicate more similar strings. If the two strings are identical the result will be 1. However, for some techniques different strings may also give a result of 1.0 (Christen 2012). A brief description of the techniques that were used in this project is provided below (Christen 2012).

Exact: exact matching returns either 0 or 1, with 0 indicating different strings and 1 indicating the strings are identical.

Q-gram: q-gram string matching splits the two input strings into substrings of length q using a sliding window approach, and then measures the proportion of substrings that are common to both of the original strings.

Jaro: Jaro comparison uses a sliding window approach and measures the number of characters the two strings have in common in this window and also takes into account the number of transpositions.

Longest common substring (LCS): LCS comparison iteratively removes the longest common substring from each of the two strings down to a minimum length and then computes a similarity measure based on the proportion of the strings that have been removed.

Bag distance: bag distance counts the number of characters the two strings have in common by converting them each into a multiset and then subtracting one from the other.

The data matching code was written in Python 3.2 and used the Febrl library (Christen 2009) for the string comparison techniques. The code was run on an I7 2600, 3.4Ghz machine with 16 Gigabytes of memory running the Windows 7 operating system. The majority of techniques had running times of 15 minutes or less when calculating similarities and clustering institutions. However, the LCS comparisons took longer, with running times of up to two hours. Since we only ran each technique a small number of times, this was not a significant issue, but if they needed to be run repeatedly or with larger data sets, then alternative languages or libraries could be investigated, along with a possible parellelisation of the algorithm.

While there were similarities between the two different data matching tasks, matching between *afid* values and determining an *afid* value for records that did not have one, they each had unique characteristics so are treated separately.

4.1 Matching Between *afid* Values

The purpose of data matching between different *afid* values was to determine where they corresponded to the same institution so they could be merged together.

The approach used an agglomerative hierarchical clustering technique (Han et al. 2012, Naumann & Herschel 2010). Initially, each *afid* was assigned to its own cluster, and each cluster was also given the name we had extracted for the *afid* as a second attribute. The data matching compared clusters using the name attribute and merged clusters where the similarity score between the names was above the assigned threshold.

To deal with transitive closure, we conducted pair-wise matching between clusters and recorded all successful matches. All clusters where there was a successful match between the names were merged. In some cases two clusters were merged even though the similarity score between their names was below the threshold, for example when they both successfully matched with a third cluster. The evaluation examined all pair-wise matches from merged clusters, even where individual matches were below the required similarity threshold.

The data matching was conducted iteratively. After each step of the data matching, clusters were merged where they had been matched successfully and a new comparison technique was tried, generally with a lower similarity threshold. The initial techniques were exact matching with a threshold of 1.0 and q-gram matching with a threshold of 0.9. Several techniques were tested for the third step. The results of the data matching are described in Table 6 below.

4.1.1 Evaluation

We provide a brief description of the results of each iterative matching step, along with specific examples

Comparison Technique	Similarity Threshold	Other Parameters	Clusters Matched	Clusters Formed	Precision
Exact matching - step 1	1.00	None	566	230	85.9%
Q-gram matching - step 2	0.90	q = 2	75	35	87.0%
Jaro - step 3	0.80	None	1,332	165	< 50%
Jaro - step 3	0.90	None	169	63	< 50%
LCS - step 3	0.80	Shortest length = 3	465	160	< 50%
LCS - step 3	0.90	Shortest length = 3	69	34	72.2%

Table 6: Data Matching Results (part 1). Note that some comparison techniques yielded precision results that were clearly less than 50% and were not fully evaluated.

where they are relevant. We started with 2,910 clusters, corresponding to the 2,910 *afid* values that we extracted a name for in the data cleaning. Exact matching merged clusters if they had exactly the same name. This matching technique merged 566 clusters down to 230 new ones representing a reduction of 336 clusters, which was 11.5% of the initial 2,910.

To evaluate precision, a random selection of 100 new clusters was reviewed. Because in some cases three or more clusters were merged into a single new one, there were more than 100 matches to evaluate. Of the 220 pairwise matches, 189 or 85.9% were correct. Of the 100 clusters sampled, 88 were completely correct, i.e. every cluster merged was part of the same institution, one was partially correct where two of the clusters were actually the same institution, and the third was different, and 11 were incorrect with none of the matched clusters referring to the same institution. Of the 12 clusters that were incorrect or partially incorrect, four of them had the correct name for the institutions, but it was a common name, e.g. "Calvary Hospital", while in the eight other cases, at least one of the institution names was incorrect.

The second technique was q-gram matching with a similarity score of 0.9 and a q-value of 2. This largely resolved institutions that were present multiple times, but with minor variations in their names. This technique merged 75 clusters down to 35, which was a reduction of 40, or 1.4% of the initial 2,910. Since there were fewer than 100 new clusters they were all evaluated.

Of the 46 pair-wise matches that were generated in this step, 40 of them were correct which is a precision of 87.0%. Of the 35 values formed, 29, or 82.9% were completely correct and 6 were completely incorrect. It is worth noting that this step combined several different divisions of the "Commonwealth Scientific and Industrial Research Organisation" (CSIRO) into a single cluster, and these were treated as correct matches. This occurred because the expansion of the acronym "CSIRO" is so long that it heavily skews the similarity scores when conducting string comparisons. We discuss this further in our conclusion.

Of the techniques tested in the third step, only LCS with a similarity threshold of 0.9 had a precision that was reasonable (greater than 70%) and even for these matches, the increase in incorrect matches would not be justified in many applications. In addition, some comparison measures gave results where the precision was clearly less than 50% were not further evaluated. The results from these techniques indicated that we might be reaching the limits of what could be achieved with string comparisons.

To try and assess the number of true matches that remained, a matching round was explored using q-gram matching and a low similarity score of 0.75 and q = 2.

The overall precision with this approach was extremely low and a few clusters that contained many institutions with similar names dominated the results. There were some true positives in the matches. It was difficult to gauge exactly how many more matches could be obtained with perfect string comparison techniques, but it is probably in the vicinity of 150 to 200. This suggests we had discovered approximately two thirds of the true matches. However, this is not counting any matches between cases where afids have completely different names but correspond to the same institution. These are unlikely to be picked up through string comparison techniques and we provide some suggestions to deal with these in our section on future work.

4.2 Records Without an *afid* Value

Out of the 1,611,172 records for Australian institutions, 29,184 or 1.8% had no value for *afid*. In some cases, this could be correct, since individuals who are not associated with an institution can perform research. However in other cases, these records had institution information present, usually in the *organization* attribute, and as a result it appeared that the blank value for *afid* was actually a data quality issue. We again used string comparison techniques to try and determine the correct value of *afid* for these records.

Of the 29,184 records for Australia that had no *afid* value, 10,376 also had no information in the *organization* attribute. These were excluded from the process since they had nothing to match against. This left 18,808 records on which to perform the data matching.

The same pre-processing steps were applied to the values of the *organization* attribute that were used when generating the institution names, such as expansion of acronyms and abbreviations and splitting the string into comma separated tokens. Both the tokens and the institution names were also converted to lower case to improve the quality of the matches.

Once the pre-processing was complete we tested different string comparison techniques and matched the tokens from the *organization* attribute for the records with no *afid*, against the institution names extracted in the data cleaning step. This was an iterative process and after each comparison technique we removed records with no *afid* that had been matched successfully from the data before trying the next comparison technique.

As when matching between *afid* values, the process began with exact matching with a threshold of 1.0, then q-gram matching with a threshold of 0.9 and then we experimented with a number of different techniques for the third step. The results of the data matching are described in Table 7 above.

Comparison Technique	Similarity Threshold	Other Parameters	Unique Records Matched	Precision
Exact matching - step 1	1.00	None	5,822	96.0%
Q-gram matching - step 2	0.90	q = 2	1,815	96.0%
Jaro - step 3	0.80	None	5,865	< 50%
Jaro - step 3	0.90	None	887	< 50%
LCS - step 3	0.80	Shortest length = 3	2,513	57.0%
LCS - step 3	0.85	Shortest length = 3	1,114	60.0%
LCS - step 3	0.90	Shortest length = 2	183	54.0%
Bag Distance - step 3	0.80	None	6,087	< 50%
Bag Distance - step 3	0.90	None	372	< 50%

Table 7: Data Matching Results (part 2). Note that some comparison techniques yielded precision results that were clearly less than 50% and were not fully evaluated.

4.2.1 Evaluation

We provide a brief summary of the results of each matching technique, along with some examples.

Exact matching checked whether one of the tokens in the organization attribute was exactly the same string as the name for an institution. This matched 5,822 records with no *afid* to an institution. This represented 31.0% of records without an *afid*, which was a higher proportion than expected. Our initial expectation was that records without an *afid* would have generally poor overall data quality.

However, transitive closure was a problem with 252 of the records matched receiving an exact match to two or more different institutions. In these cases an organization value had at least two comma-separated substrings and they had each matched exactly to two different institutions. For example "Department of Physics, University of Sydney" matched to both an institution called "Department of Physics" and an institution called "University of Sydney". The institution named "Department of Physics" likely has an incorrect name and the true match should be with the "University of Sydney". From the analysis it appeared that there was a fairly strong link between the confidence in the institution names in the data cleaning phase, and the likelihood that they were correctly matched. As a result, the ratio value calculated during the data cleaning stage was used as a tiebreaker when resolving transitive closure in these cases. An evaluation of 100 randomly selected matches gave a precision of 96.0%. For all 4 records that were incorrectly matched, the institution name was probably incorrect. Of the successful matches, a few large institutions that were frequently missing an *afid* were responsible for a large proportion of the total matches.

The second technique was q-gram matching with q = 2 and a similarity threshold of 0.9. This matched another 1,815 records. Transitive closure affected another 233 records and was dealt with as for exact matching. This step primarily matched records with minor name variations or typographical errors. An analysis of 100 random matches gave a precision of 96.0%. As with exact matching, a small number of institutions contributed the majority of the positive matches.

Unfortunately, for the third step of the matching, the results were not promising with none of the comparison techniques tried giving a good level of precision. As with exact matching and q-gram matching, we performed an evaluation on a random sample of 100 matched records. However, in many cases the matching quality was too poor to warrant a detailed analysis since it was clearly less than 50% precision. Of the techniques tested, only the LCS comparison with a minimum substring length of three gave a precision of 60%, and even this is generally too low to be useful.

Since the first two techniques had only matched approximately 40% of the 18,808 records, we performed a more detailed analysis of the results to determine why the match quality was so poor and detected three main causes for the poor precision.

Institutions where the extracted names were only partially correct or were incorrect had a significant effect on the results. In particular, a small number of institutions with a name that was similar to subdivisions of other institutions had a disproportionate impact on the number of incorrect matches. For example, one *afid* was assigned the incorrect name "Department of Medicine". This resulted in many records that contained substrings such as "Department of Renal Medicine" or "Department of Emergency Medicine" generating high enough similarity scores to achieve a false positive match.

Another problem was certain types of institutions with very similar names. For example, within Australia, many State Governments, as well as the Commonwealth Government, have a "Department of Primary Industries" and the string comparison techniques were not effective at distinguishing between them. This resulted in many records being assigned to the incorrect institution.

Finally, within the records that did not have an *afid*, a significant proportion did not have the institution name present. An evaluation of 100 records that were not matched by either the exact match or the q-gram matching found that in 40% of cases, the institution name did not appear to be present anywhere in the record. In another 13% of cases, the institution name was heavily abbreviated or shortened enough to make string matching difficult.

4.3 Data Matching Summary

Overall, the data matching led to mixed results. The process for matching between different *afid* values to determine whether they corresponded to the same institution was reasonably successful. We achieved precision of 85% or higher for the first two comparison techniques, and estimate very roughly that between them they accounted for around two thirds of the positive matches that could be found. This could be stretched a bit further using the LCS technique if a few more incorrect matches were acceptable for the end use case.

However, the matching to assign an *afid* value to records that did not have one was less successful. While the matches from the first two techniques were very good with a precision over 95%, the total coverage was only 40.6% of the records without an *afid* and after this no technique produced good results. There was a significant difference between the easy matches, and the more difficult ones. Once the records for the large institutions that were frequently missing an *afid* had been resolved, it was quite difficult to deal with the remainder.

5 Conclusions and Future Work

Overall, the project results were reasonably positive, and largely achieved the project goals, but there was still room for improvement. The data cleaning phase where we extracted institution names went well. For institutions with 10 or more records, we extracted a correct or partially correct institution name for 86.5% of *afids*. This alone was a very useful result from the project, since identifying the correct name for institutions in SCOPUS can be challenging and is often a limiting factor when using the data.

The merging of different *afids* was also reasonably successful with the exact matching and the first q-gram matching both having a precision of over 85% and between them reducing the number of afid values by 13%. The longest common substring comparison with a similarity threshold of 0.9 achieved precision over 70%, however in practice, the increase in incorrect matches may not justify the overall number of additional matches gained. An examination of the output with a low similarity threshold suggested that for the institutions where the names were correct approximately two thirds of the true matches had been detected.

Dealing with the records that did not have an *afid* was less successful. The initial steps were good with a few large institutions that were frequently missing an *afid* and were easy to resolve resulting in a precision over 95%. Between them they accounted for around 40% of the records. However, all subsequent techniques had very poor match quality.

We identified three common causes for the incorrect matches, including institutions with very similar or the same names, a small number of incorrect names extracted during the data cleaning phase resulting in a disproportionate number of incorrect matches, and many of the records not containing the institution name.

We detail in the section on future work some ways these issues can be addressed. Once this has been done to a satisfactory degree, the output from this project can be incorporated into the SCOPUS database and used to improve future analysis.

With respect to the second goal of improving organisational capacity with respect to SCOPUS, the project was also valuable. Two projects currently underway involve assessing the links between Australian institutions and those in Indonesia and India. The experience and knowledge gained from this project has been valuable for this analysis, particularly regarding the specifics of the SCOPUS database and the challenges present in the Australian institution data.

5.1 Applications and Domain Knowledge

There were many characteristics of the project that were unique to the SCOPUS database and which might not be applicable elsewhere. However, aspects of the domain knowledge could be useful in other data matching on Australian institutions or worldwide. In addition, the approach we used for name extraction and the ratio concept could be applied in other areas.

One of the biggest problems for the project was a result of the word "department" pertaining to both subdivisions of larger institutions, particularly universities, and also to institutions such as government departments. While extracting the institution names, there were a number of small institutions that incorrectly received names such as "Department of Medicine" and "Department of Physics". In most of these cases, the institution name was not actually present in any of the records, so it was impossible to tell what the institution actually was. However, when these names were used in the data matching, they caused significant problems, particularly when matching against records without an *afid* value, where they frequently caused false positive matches.

A few rules could be quite effective at resolving this problem. For example, given a country, it might be worthwhile to create a lookup table of the main government departments, and exclude any institution name that contains the word "department" which is not in that table. A small number of rules to deal with cases such as these could significantly improve the results.

In addition, an improved methodology for dealing with acronyms could also be worthwhile. As mentioned in the analysis, all the divisions of CSIRO were combined by the early string comparison techniques since the expanded form of CSIRO is so long that it dominates the matching. For CSIRO this was not a problem since they are all part of the same institution. However, similar situations occurred to a lesser degree with other long expansions such as CRC for Cooperative Research Centre, or NSW for New South Wales. In practice, it could be worthwhile to change the data cleaning approach to detect the expanded forms of acronyms, perhaps allowing slight variations, and then reduce them to their acronyms for data matching purposes, rather than the other way round. This would prevent these terms causing too many incorrect matches.

While we have not tested it elsewhere, there is no a priori reason why the frequency based approach that we used to extract institution names couldn't be applicable in other areas. In particular, any application where the domain is relatively small in relation to the number of records would be a good candidate for this approach. Examples could include suburb names for a country, or potentially company names or product names. Alternatively, the domain could be restricted to the larger examples, as we did in this project, in order for the technique to be used. For a simple and easy to implement technique, it was surprisingly effective.

The ratio concept could also be used in these situations as an indicator of confidence in the correctness of the result. A high ratio value indicates that there is only a single good candidate for the correct value, whereas a ratio value close to 1 indicates that there are two or more candidates for the correct value and it may be difficult to pick between them. This technique could be extended by creating a probability distribution from the results, rather than using the ratio value, which only considers the two most frequent values. Doing this could better capture the variability, especially if there are three or more candidates for the correct result. However, care would need to be taken in these situations to not overemphasise the impact of the low probability results. For example, in this project, very few institutions with a ratio value above 4.0 had an incorrect name ex-

tracted. However, for the Australian National University, the correct name only made up 23.7% of the comma-separated substrings. This situation was also common in other large institutions so using a probability distribution could risk more incorrect results rather than less.

Finally, small variations to the methodology would also be worthwhile in many practical applications. If coverage is less important for the analysis being undertaken, then for the SCOPUS data, accuracy could be increased to over 90% with a reduction in the coverage of approximately 12%. In practice this is probably a worthwhile tradeoff, since even a small number of incorrect names from the data cleaning step significantly increased the number of incorrect matches in the data matching. Similarly, modifying the technique to also incorporate the number of records could improve the result, since generally the large institutions were more likely to be correct.

5.2 Future Work

There are a number of ways this work could be extended in the future. The use of more sophisticated data matching techniques such as incorporating TF-IDF (Term Frequency - Inverse Document Frequency) (Christen 2012) could improve the quality of the matching, particularly for determining an institution for records without an *afid*. When dealing with institutions that had long names such as "Department of Natural Resources and Mines", where a small difference in the name is actually important from a data matching perspective, a TF-IDF approach could be quite effective. However, even these techniques would not assist in cases where the institution name is simply not present in the record.

Based on our evaluation of the results in the data matching phase, a few different situations were responsible for a large proportion of the incorrect matches, both when data matching between *afid* values and when trying to determine an *afid* value for records that did not have one. The creation of a small set of domain specific rules could significantly improve the quality of the institution name extraction, and the subsequent data matching.

Finally, a collective data matching approach (Christen 2012) that attempted to do data matching on articles, authors and institutions simultaneously might be very successful though it would also be complex and computationally intensive. The data could be treated as a network capturing links between articles, individuals and institutions with the weights of the links measuring the frequency of the connections. This type of approach could potentially handle missing values in the data, and would also be very good at dealing with situations where a few records had incorrect values.

References

Christen, P. (2009), 'Development and user experiences of an open source data cleaning, deduplication and record linkage system', *SIGKDD Explorations* **11**(1), 39–48.

Christen, P. (2012), *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*, Springer.

Ciszak, L. (2008), Application of clustering and association methods in data cleaning, *in* 'International Multiconference on Computer Science and Information Technology', IEEE, pp. 97–103.

Elmasri, R. & Navathe, S. B. N. (2011), *Database systems: models, languages, design, and application programming*, Pearson.

ERA (2012), 'Excellence in Research for Australia 2012 National Report'. Australian Research Council.

Han, J., Kamber, M. & Pei, J. (2012), *Data mining: concepts and techniques*, 3 edn, Waltham, MA: Morgan Kaufmann.

Hirsch, J. (2005), 'An index to quantify an individual's scientific research output', *Proceedings of the National Academy of Sciences of the United States of America* **102**(46), 16569–16572.

Lee, D., Kang, J., Mitra, P., Giles, C. L. & On, B.-W. (2007), 'Are your citations clean?', *Commununications of the ACM* **50**, 33–38.

Naumann, F. & Herschel, M. (2010), *An introduction to duplicate detection*, Vol. 3 of *Synthesis Lectures on Data Management*, Morgan and Claypool Publishers.

Scopus (2009), *Scopus Custom Data Documentation*, Elsevier, Amsterdam.

Smalheiser, N. R. & Torvik, V. I. (2009), 'Author name disambiguation', *Annual review of information science and technology* **43**(1), 1–43.