Assigning a Quality Measurement to Matching Records from Heterogeneous Legacy Databases: A Practical Experience

L. Robert Pokorny
XSB, Inc.
Stony Brook, New York
pokorny@xsb.com

Abstract. In this paper we present the results of a practical exercise in matching records from two different legacy databases. This exercise is a portion of our work for the US Defense Logistics Agency, (DLA). It addresses the problem of identifying parts in the database of DLA managed items that match parts available from a commercial online catalog. A method of automatically generating approximate matching records is described and a procedure is presented for assigning scores to these matches. Finally a statistical quality sampling standard is applied to these match results to generate a quality measure for each score. This gives DLA an objective means to determine whether or not a given part can be ordered from the commercial catalog. This methodology can be applied to any data manipulation process such as data cleaning or data mining where the process quality can be independently verified through statistical sampling.

1) Introduction. As organizations become more web aware they increasingly need to compare their own internal data to data available from outside web sources. Consider the case of ordering parts from an online catalog. The organization placing this order needs some assurance that the part it requires as specified by its internal data is the same part that is ordered. This necessitates matching the internal data for the part to external data presented by the catalog source. It also requires some quantitative measure of how good the match is.

Such a situation arose as part of our work for the US Defense Logistics Agency, (DLA). DLA wanted to determine which of the parts it managed could be ordered from commercial online catalogs. We developed a method of automatically matching DLA's internal part data to part data from online catalogs. This generated possible part matches, which were scored on how well they matched. Each score level was then assigned a quality measure using statistical sampling techniques outlined in MIL STD 105 [3], a well-known quality measurement standard. By sampling matches made at various score levels and manually examining the samples to judge if they were indeed good matches, we were able to assign an Acceptable Quality Level, (AQL), to each score. The crucial point here is that match scores were generated by the automatic matching process while AQLs for the scores were developed from a manual inspection process using parts of the data not accounted for in the automatic process. DLA can now use these AQLs to determine if a part can be ordered from an online catalog based on the risk associated with getting and using a wrong part. This is similar to ordering parts from a manufacturer based on the AQL of the manufacturing process.

The following sections of the paper describe the details of this data matching process. Section 2 outlines the matching methodology. This is the process of automatically generating matches. In Section 3 we describe the way matches were scored. Scores were based on the success of various components of the matching process. Section 4 presents the quality validation of the different score levels. This is the result of the sampling and inspection of matches and assigns an AQL to each score level. Finally, in section 5 we present our conclusions and plans for future work.

2) The Record Matching Methodology. The problem of determining which parts managed by DLA are available from commercial online catalogs exists because DLA and the online catalogs maintain different data about parts. The differences in these legacy databases makes it difficult to directly join records from the two sources to establish matches. Data about parts from DLA contains the following fields:

NSN – National stock number (Unique Federal identification number)
Cage Code – ID for registered source for part
Manufacturer Part Number – assigned by manufacturer
Nomenclature – official name of part (assigned by DLA)
Vendor Name – Entity name Cage Code is assigned to
Vendor Address – Address of entity

The specific exercise attempted to match DLA parts as described by the above records to parts from Newark Electronics, an online vendor of electronic parts. Data records about parts from Newark Electronics took the following form:

Commercial Part Number – assigned by manufacturer
Part description – field containing free text describing part
Catalog Number – Unique ID assigned by Newark Electronics
Vendor ID – Number assigned by Newark to identify source of part
Vendor Name and Address – Newark source for part

The only field that is identical in both the DLA and Newark Electronics data is the Manufacturer Part Number/Commercial Part Number. This number is assigned by the manufacturer and is identical in both databases. All other field values are assigned by the respective data owners and have no guarantee of matching. Even vendor names and addresses vary significantly between the two databases. Newark Electronics might use 3M as the manufacturer's name and DLA uses Minnesota Mining and Manufacturing. DLA may contain a street address and Newark Manufacturing a Post Office box. However, joining the two databases on the part number field does not insure that parts will match. While manufacturer part numbers are unique for parts from a given manufacturer, these numbers are not unique across manufacturers. The same number can be assigned to many parts from different manufacturers. Matching the part number is necessary to insure a part equivalence, but it is not sufficient.

The approach we take to establish matches comprises three steps. First, we join the DLA and Newark Electronics databases on the part number field. Any match produced from this join is a possibly equivalent part. The matches so obtained produce DLA Cage Code/Newark Vendor ID pairs. Second, we standardize name and address data from both DLA and Newark Electronics. This standardization process produces records that contain the following fields for both DLA and Newark Electronic data:

Company Name
Alternate Name
Street Number
Street Name
PO Box
City
State
Zip Code

For DLA, this record is associated with a Cage Code and, for Newark Electronics, with a Vendor ID. The alternate name field is only populated for DLA records since DLA often lists division and subsidiary names along with the company name. The third step in generating matches is to compare the above standardized records for a given Cage Code/Vendor ID pair and score how well the various fields match. Details about scoring the matches are presented in Section 3. The standardization of names and addresses is described in [1]. Both the standardization and the scoring of matches are implemented in XSB, [2], an efficient deductive logic engine well suited for natural language analysis.

3) Scoring the Matches. The critical issue in determining whether a part from DLA is equivalent to a part from Newark Electronics revolves around how well the name and address records for a Cage Code/ Vendor ID pair match. We generate a score for this match, which assigns a quantitative value for the closeness of match. The score is composed of a number of components and is the sum of these components. The components are as follows:

> Cage Code/Vendor ID frequency – this component is a measure of how many part numbers produce the given Cage Code/Vendor ID pair. This component is scored from 1 to 4 on an approximately logarithmic scale.
>
> Name – this component indicates how well the name matches. Names are represented as strings of words and a score is based on the size of the longest common subsequence in the two name strings. This score ranges from 0 to 5 for matching company name to company name and 0 to 5 for matching company name to alternate name.
>
> Street Number – this component give a score of 0 when street numbers don't match, a score of 1 if the do match but zip codes don't match, and a score of 2 if both street number and zip code match
>
> PO Box – this component scores PO Box matches from 0 to 2 in a similar manner to street numbers.
>
> Zip Code – this component scores a 0 if zip codes don't match, a 1 if 5 digit zip codes match, and a 2 if 9 digit zip codes match.

A total score for a Cage Code/Vendor ID match is achieved by summing the component scores with the caveat that the score is 0 unless some component besides the Cage Code/Vendor ID frequency produces a non-zero result. The highest possible score for a match is 20. The best score we achieved was 14. This is a good indication of how different name and address data can be between the two databases.

There are approximately 150,000 items in the Newark Electronics catalog data. There are approximately 4,000,000 managed NSNs in the DLA database. The join on part numbers produced approximately 249,000 possible matches. Of these only approximately 30,000 had scores higher that 0. The distribution of scores is presented in figure 1.

80% of these matches scores 5 or better. Our initial assumption was that these indicated good matches, but we needed some method of quantifying how good. In the next section we describe using statistical quality sampling techniques to assign a quality to each score level.

4) Validating Scores. The idea behind assigning a quality measure to match scores is the same as that for statistical quality control. One such measure, described in MIL STD 105 [3], is Acceptable Quality Level, (AQL). AQL is defined as the number of defects per 100 items produced. In the context of producing matches between DLA parts and Newark Electronic parts AQL would indicate the number of false matches per 100 matches generated. In this case a match

is either correct or incorrect, so there can be at most one defect per match. The procedure defined in MIL STD 105 is to choose a random sample from a production run based on the size of the run. This sample is inspected for defects. To achieve a certain AQL, the sample can have at most a certain number of defects.

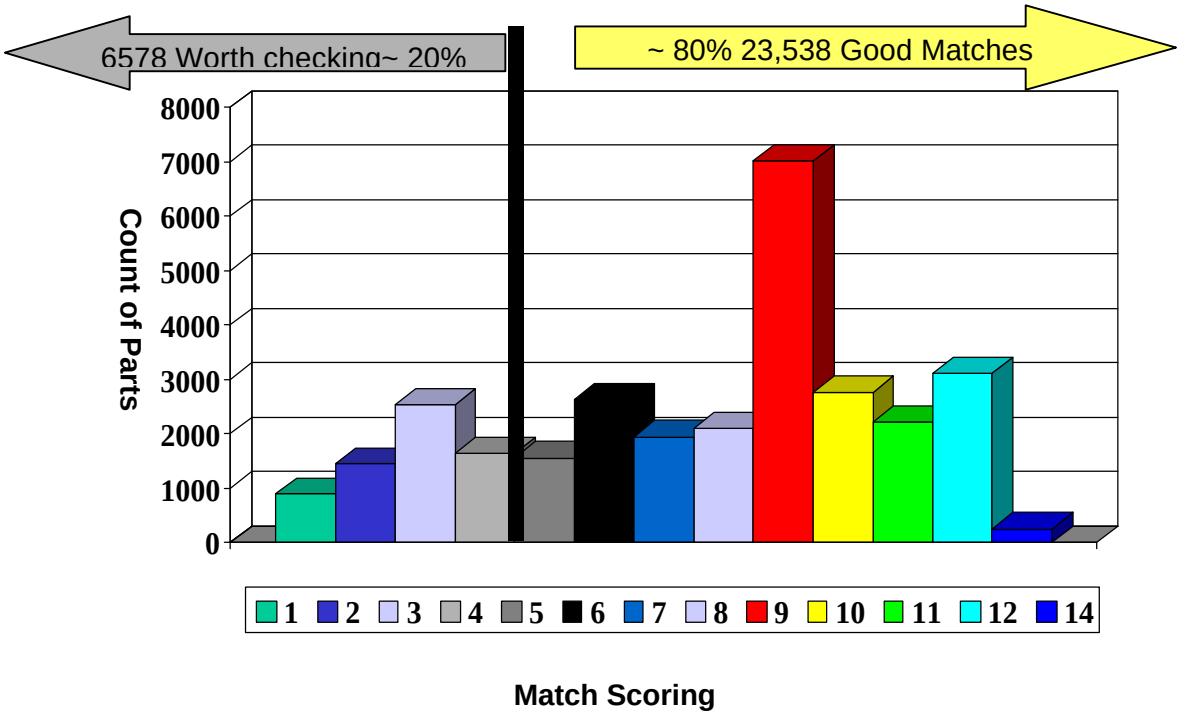

Figure 1 – Distribution of scores for part matches

Implicit in the procedure is the idea that inspection should be done by some procedure that is different from the procedures used to produce the items being inspected. One should not measure the length of a part using the milling machine used to machine the length. In our case, we validate matches by comparing the DLA nomenclature for the part to the Newark description of the part. These data were not used in producing matches.

Our methodology for assigning AQLs to match score levels proceeds as follows. We choose a random sample for a given score level based on MIL STD 105 tables and the number of matches generated at that score level. We choose an initial AQL and inspect the sample manually. If we find an acceptable level of defects we choose another random sample and a lower AQL. This continues until we get a sample with too many defects for the chosen AQL. We then assign the last successful AQL to that score level. If, on the other hand, our initial sample fails, we choose another random sample and a higher AQL. We continue this way until a sample succeeds and assign the chosen AQL to that score level.

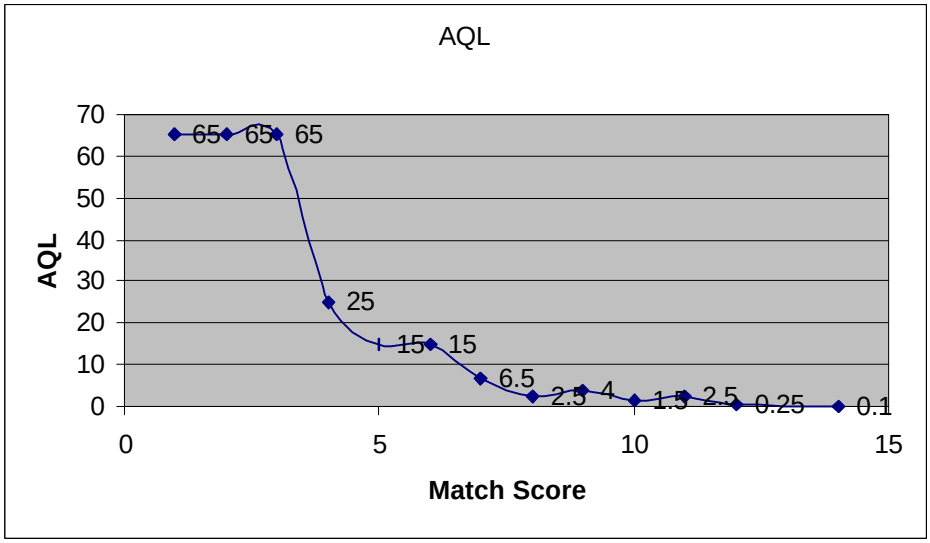

Figure 2 – AQL results for Match Scores

Figure 2 illustrates the results of this process. Note that scores of 5 or better have AQLs of 15 or less. This guarantees that, with a score of 5 or better, at least 85 % of the matches are good. At a score of 14, only 1 match in 1000 is bad. This gives DLA a straightforward way to evaluate the risk of buying a part from Newark. If the part is expensive or critical to safety or performance they would choose a very low AQL before deciding to order. If this is not the case they could choose a higher AQL. By applying statistical quality control to our data processing we are able to give DLA a quantitative measure of the quality of part matches.

5) Conclusions and Future Work. The procedures outlined here can be applied to many data processing quality problems. They assume that there is some method to produce new information from underlying data. This could be a data mining technique, a data cleaning process, or some other data transformation. In our case, we produced data matches between two legacy databases. They also assume that there is an independent way of verifying the results in addition to the method that produced them. We were able to validate matches using descriptive data not used in producing the matches. When these conditions exist, a quality measure can be assigned to results based on statistical quality control methods.

We intend to extend this to other data manipulations we are performing for DLA. We are also using DLA data to evaluate manufacturers' capabilities and group parts by common features. These tasks involve applying reasoning rules to the data to produce new knowledge. By applying this quality technique to these processes we can give DLA assurances as to how valid this new knowledge is.

Acknowledgements. We would like to acknowledge the support of Don O'Brien and Tony Monteleone at DLA for their support and insights. We also want to acknowledge the work of Michael Epstein at XSB, Inc for his programming efforts to set up the data validation procedures.

References.

1) B. Cui, T Swift, and D. S. Warren. Preference logic grammars: semantics, implementation, and application to data standardization. In *Proceedings of the 1999 International Conference on Logic Programming and Non-Monotonic Reasoning*
2) K. Sagonas, T Swift, and D. S. Warren. XSB as an efficient deductive database engine. In *ACM SIGMOD Conference on Management of Data.* ACM Press, 1994
3) US Department of Defense. Military standard 105 E, sampling procedures and tables for inspection by attributes. May 1989