Preparing Clean Views of Data for Data Mining

Paul Jermyn[+], Maurice Dixon[+] and Brian J Read[$]

[+]CISM, London Guildhall University, 100 Minories, LONDON, EC3N 1JY, UK
Jermyn@lgu.ac.uk and M.Dixon@rl.ac.uk

[$]ITD, CLRC Rutherford Appleton Laboratory, Chilton, DIDCOT, Oxon. OX11 0QX, UK
Brian.Read@rl.ac.uk

ABSTRACT

Good data preparation is a key prerequisite to successful data mining [P99]. Conventional wisdom suggests that data preparation takes about 60 to 80% of the time involved in a data mining exercise [R97]. There have been good reviews of the problems associated with data preparation [F97, HS98 and MS97]. However the data cleaning aspect of data preparation is regarded as involving major human input and often has been neglected in practice.

This paper reports work undertaken in support of a data mining programme at Rutherford Appleton Laboratory (RAL). It proposes a *Clean Views Model* for data cleaning. The action taken to clean data depends on the business purpose of the analysis. The cleaning action also depends on what you believe about the data expressed in the business rules that the data need to satisfy. These rules are expressed in a *validator function*. What you believe about the data can change during the data mining process and also business purposes change through time. This means that data cleaning should be described as an iterative exercise rather than a one-off procedure. We contend that the original, unclean data should be retained and that the cleaning produces a particular view, a *Clean View*, for a particular validation function.

We have developed the Clean Views approach after looking at several datasets. We report some work on the flow of funds data associated with transitional economies [Ba99]. These data were generated by the central banks of Eastern European countries and are used to assess their movement towards a market economy. We also report work done on data in an international mailing list concerned with a survey of alumni of business schools who have graduated as Masters of Business Administration.

KEYWORDS: Data Cleaning, Data Mining, Data Preparation, Data Validation, Mailing Lists, Transitional Economies

1. INTRODUCTION

Good data preparation is a key prerequisite to successful data mining [P99]. Experience suggests that data preparation takes 60 to 80% of the time involved in a data mining study [R97]. What you do about cleaning data depends on what you believe about the business and on what features seem relevant to your particular study. Data cleaning is expensive and labour intensive; therefore, it tends to be done once, if at all, before analysis or data warehousing [L98, N97]. There are two ways of improving flawed data. Either, new data must be found or it must be possible to reason over the existing data to infer clean values. Usually a combination of these ways is adopted.

In this paper, we describe work undertaken as part of RAL's data mining programme. We introduce an approach to data cleaning called the *Clean Views* model in which we propose that the cleaning actions are based upon a validator function that is determined by the business needs of the analysis.

We have explored some issues associated with data cleaning by looking at a range of business domains. These include:

- Flow of funds within transitional economies of eastern Europe [Ba99]
- Statistical returns on students for the UK's Higher Education Statistical Agency [HESA98]
- International mailing lists used for in questionnaires [J99]

The main data problems that occur in datasets arise from missing data, erroneous data, duplicated data with inconsistencies, and heterogeneities. The information structure in a dataset is quite subtle and inappropriate cleaning can damage the data [P99]. The main process problems relate to managing and auditing data cleaning decisions.

We introduce a *Clean Views Model* that asserts that the original data should be retained intact. The model requires that the business rules that the data should satisfy for the intended analysis should be specified before cleaning. These rules are expressed in a validator function. The model is iterative and allows for different combinations of cleaning actions and validation functions.

The *Clean Views Approach* moves the emphasis of data cleaning from it being a human activity of hand cleaning to being a system development activity. This means that data cleaning becomes scalable as the size of the datasets increases. It makes explicit the need to define a validation function for acceptance of data. However the problems in detecting what to clean remain, especially for erroneous data. We are building a tool to structure the iterative processes.

2. PROBLEMS IN DATA AND THE CLEANING PROCESS

2.1 Past Experience

Practical experience shows that data preparation can take between 60% and 80% of the effort in a typical business data mining project [R97]. It would seem reasonable to assume that the same is true for any non-trivial analytical exercise. The difficulties that underpin this resource cost can be understood from at least two perspectives: data problems and process problems. First, data problems will be briefly described, then process problems. Finally, a list of requirements for a cleaning process to solve the process problems will be given.

2.2 Problems with Data

Recent papers [F97, MS97] describe well the problems in data recorded in datasets. Here we just note some key problems:

Missing data occur two ways

- where data are expected but are absent
- where data are appropriately not available or inapplicable in the real world

Detection of the missing data is often relatively straightforward.

Erroneous data occur when an incorrect value is recorded for a real world value [F97]. Detection can be quite difficult. (E.g. the incorrect spelling of a name)

Duplicated data occur in two ways

- repeat records, perhaps with some values different
- different identifications of the same real world entity

Repeat records are common and usually easy to detect. The different identification of the same real world entities can be a very hard problem to detect [HS98].

Heterogeneities arise when data from different systems are brought together in one analysis [DKCJR96]. Two of the problems are

- structural heterogeneity which arises when the data structures reflect different business usage
- semantic heterogeneity which arises when the meaning of data is different in each system that is being combined

These heterogeneities can be very difficult to resolve because they usually involve quite a lot of contextual knowledge that is not defined as metadata.

There are subtle information dependencies contained in the relationships between the values of different sets of attributes. The wrong form of cleaning could result in damaging information in the data. Different analysis tools handle missing data in different ways. (For some tools, missing data leads to record discard during the data mining [P99].)

Commercial offerings exist that assist the cleaning process, but these tend to be problem specific [T98].

Uncertainty in information systems is a well-recognised hard problem [MS97]. We show in Figure 1 very simple examples of missing and erroneous data.

Subject Identifier	**Surname**	**Forename**	**Age**	**Real World Fact**	**Classify**
923C4	Ongorum	Ahmet	31	Ahmet Ongorum is 31	*correct*
782J7	MacBride	Jenny		Jenny MacBride is 30	*missing*
899P3	Abeko	Peter	23	Peter Abeko is 21	*erroneous*

Figure 1: Simple Examples of Missing and Erroneous Data

Because a problem is simple to illustrate, it does not mean that it is simple to detect or clean. In particular, semantic problems are notoriously difficult (Figure 2).

Subject Identifier	**Surname**	**Forename**	**Time**	**Duplication Fact**	**Classify**
923C4	Abeco	P A	2:55		
542J7	Rosen	Nadja	4:43		
899P3	Abeko	Peter	3:13	Same real world entity as 923C4	*semantic*
542J7	Rosen	Nadja	4:43	Identical record	*simple duplicate*

Figure 2: Simple Examples of Duplication

2.3 Problems with Process

The process of cleaning data has at least five identifiable problem areas. They are:

- One shot cleaning
- No record of cleaning decisions
- Cleaning choices made at a low level
- Expensive manual methods
- Lack of clear methodology

One shot cleaning means that the data are cleaned once to produce a “correct” database. This approach is supported by some data warehousing practices. Given the cost of cleaning data, this is understandable. The difficulty is that the way data are cleaned depends on the intended use. *For example, a record may contain sufficient information to decide on the country of residence of a person (and thus be ‘clean’ for an analysis that uses that information). However, it may not contain a complete postal address and so be ‘dirty’ for posting to that address.* If the use of a database is not well defined before cleaning - a reasonable assumption for analytical databases rather than operational databases, then there is always the possibility of the cleaning of the database being inappropriate for a particular purpose.

The process of cleaning data consists, in part, of a series of decisions. *For example, treat all cases of an empty or null value as a zero.*

2.4 Need for Audit of Data Preparation

For a large dataset, different cleaners may be involved at different times. The need for an audit trail, which records assumptions about the data and transformations on the data, becomes imperative. The audit trail will allow the analyst to be aware of the changes and to modify them. If the dataset is to be subject to quality assurance then the data cleaning needs to be auditable.

3. CLEAN VIEWS APPROACH

3.1 Introducing Clean Views

We introduce the *Clean Views Approach* that begins from the observation that data cleaning is for an analysis that meets a business purpose. A single cleaning is unlikely to fit all potential purposes. Indeed a cleaning for one analysis may be destructive of key data for a subsequent analysis. *E.g. the replacement of missing values could prevent a subsequent analysis of factors leading to missing values.* Yet, it is common in data warehouse literature to see assertions of the need to clean the data before putting it into the data warehouse [N97]. In the *Clean Views Approach,* we segregate the results of cleaning from the original data.

The steps in the cleaning are:

a. data assembly and standardisation
b. data validation which specifies acceptance criteria
c. data transformation where cleaning rules are applied to data to give a cleaner view
d. iteration over b and c
e. record cleaning decisions

These steps assert the need to be able to refine the acceptance criteria and to track the data cleaning with the possibility of undoing steps at a later stage.

Our use of the term *VIEW* combines common English usage and the understanding in relational database terminology. Namely, we are creating a sight of the original data that is achieved through filtering by selection and through focussing by transforming with the cleaning rules. We can alter our view by changing the filter and/or the focus.

We expect that the *CLEAN VIEW* will often be instantiated physically as a dataset for performance reasons to enable further work on it. However, we anticipate that there will be circumstances when the view remains virtual.

3.2 Iterative Perspective

The business purpose for which an analysis is carried out may be extended during the lifetime of the dataset. This means that the cleaning decisions appropriate to one use may have to be revisited when another analysis is under consideration. Within a given business purpose we may want to improve the cleaning based on later experience, *e.g. by adjusting a statistical threshold*.

These reasons lead us to conclude that there is a need to preserve the original dataset for it to be available for subsequent cleaning episodes. We define a CLEAN VIEW as a dataset resulting from a transformation of the original data by a set of cleaning rules derived from a validator function. The CLEAN VIEW is always defined by the transformation from the original data.

3.3 Establishing Base Data

The source data need to be assembled in a standardised and consistent format. We refer to the resultant set of tables as base tables. Such an assembly should tackle the undisputed and unambiguous aspects of data preparation. *An example is converting dates to a standard format.* It does not restructure the tables or predetermine data cleaning choices. *For example, we do not interpret the meaning of '-' as missing data.*

3.4 Validator Functions

It is necessary to specify the validation criteria for accepting that a dataset is clean for the intended analysis. Some of the criteria are concerned with the way the analysis tools will function. Other criteria will be concerned with the business domain. The validation criteria can apply to sets of values as well as to individual values. Some data problems are always rejected as fatal for the purpose of the business analysis, whereas some data problems are tolerated up to a certain level. An important feature is that external agencies may supply the validation rules to which datasets must conform before being accepted. These rules can change over time.[HESA98]. Note that these can involve detailed business rules and are not just mechanical integrity checks.

For example, students at English Universities pay different fees according to eligibility criteria. An incoming student on the Erasmus programme is regarded as a ***special student****. The* ***major source of tuition fees*** *has to be declared as 'no fees' and the* ***fee eligibility*** *has to be asserted as 'Eligibility to pay home fees not assessed'. The actual field values are declared as numerical codes and the conditions are expressed as code values. (****special student =*** *[3|8] has* ***fee eligibility*** *= [3] and has* ***major source of tuition fees*** *= [98]. There are other credible values for the* ***major source of tuition field*** *such as [41 European Commission | 47 Other Overseas Funding] as codes 42 to 48 have the rubric 'overseas means non-UK'. It is worth noting that a local manual introduces an error in documenting one of its criteria by substituting* ***fee eligibility*** *= [8] in the constraints. Specifically in this example, the* ***special student*** *field is acting as a pointer to metadata about the correct value to insert in another field in the same record.*

3.5 Identification of Failure Type

Identification is the statement that the validation failed because of a particular class of cleaning problem that arose because the data were imperfect and inconsistent. *For example, identification of a defect as a missing value.* Validation functions can combine chains of conditions. Identification is concerned with recognising specifically where the failure has occurred.

*For example, (****special student =*** *[3] has* ***fee eligibility*** *= [3] and has* ***major source of tuition fees*** *= [41]) fails validation). The identification asserts that the failure type is inconsistent values for two of the fields. There are two ways the failure could have occurred. Either the* ***special student =*** *[3] is not correct, or* ***major source of tuition fees*** *= [41] is not correct, or both.*

The validation failure trace helps identify the best choice of a cleaning rule that addresses a specific problem with dataset.

3.6 Transformation Rules

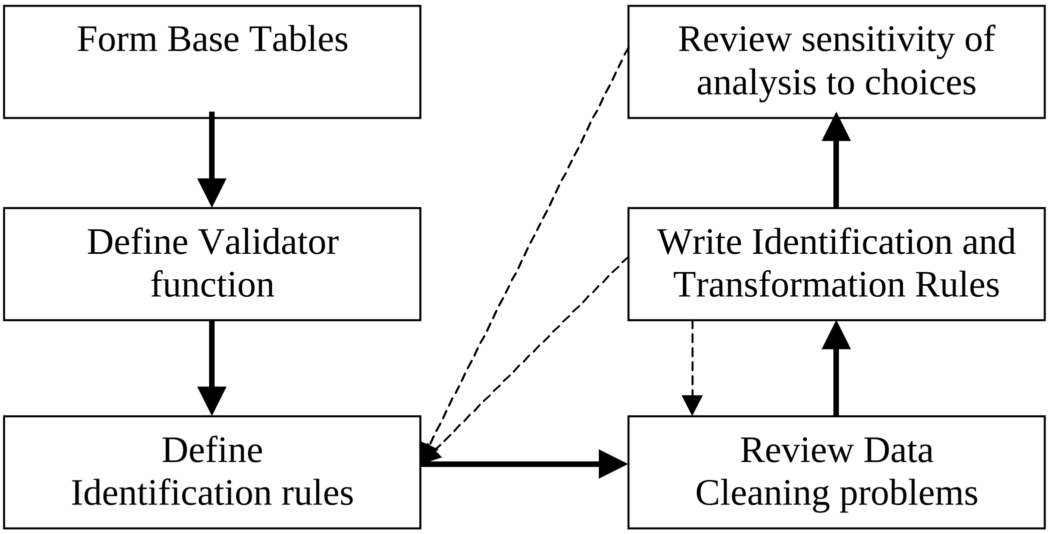
Once an identification has been made, then the choice between alternative transformation rules expresses different beliefs about the semantics of the data. The transformation rule specifies the action to be taken to transform data. Examples of the rule may be:

- substitute a value
- normalise a set of values
- omit a record
- fit a smoothing curve
- scale values with a damping function
- map dimension between scales (e*.g. ordinal to numeric)*

A choice of a different set of rules will generate a different clean view. Often validation failure leads to an improved modified clean view. However, there can be real conflicts of end user requirements identified and then alternative clean views must be specified.

3.7 The Process

The Process Diagram (figure 3) shows the stages involved in a particular Clean Views project. First, the Base Tables are formed from the source data by standardising formats and assembling the dataset. Then the Validation function for that data and a given analysis is defined. The Validation function provides the basis for the definition of Identification rules for problems in the dataset. The problems revealed by the Identification rules are reviewed. Transformation rules are defined. These produce the Clean View from the Base Tables. At this point earlier steps in the process may be revisited as the impact of the transformation rules is considered and revisions can be made. Finally, the choices made in the project can be reviewed and the sensitivity of the results of the analysis to those choices considered.


Form Base Tables
Define Validator function
Define Identification rules
Review sensitivity of analysis to choices
Write Identification and Transformation Rules
Review Data Cleaning problems

Figures 3: Process Diagram. The bold arrows on this figure show the main sequence in creating a clean view for a given validator function. The broken arrows show back tracking during the iteration process.

4. SIMPLE EXAMPLE - TRANSITIONAL ECONOMIES DATA

4.1 Transitional Economies

This study was concerned with the behaviour of the transitional economies of Eastern Europe. A transitional economy is one that is moving to a market economy from a centrally planned command economy. The subset of countries for which data were examined comprised Croatia, Czech Republic, Latvia, Poland and Slovakia. The majority of data come from the central banks of the countries concerned. Non-Governmental Organisations provide some data (e.g. industrial production figures). The analyst was investigating the banking sector and its effectiveness at distributing funds. The data formed the basis of a flow of funds analysis and the concern was with the implications for banking sector reform [Ba99]. The data used in the study were based on figures published by the central banks of five of these transitional economies.

This example was chosen for study because of the availability of an analyst with domain knowledge who was committed to studying the data. Also the data were held in spreadsheets and spreadsheets themselves are notorious for presenting problems for data quality [BPBK87].

4.2 Validation/Identification/Transformation

1. **Validation for Analytical Tool**

   The main analytical tool was regression for looking at correlations between economic factors. Regression has clear requirements. All the data must be numeric data. The correlation is between two time series with the same number of points and no missing values. For this particular problem infill was needed for Poland.

2. **Identification – Metadata**

   Here the agenda for the metadata to be used in the transformation was set by the validation. In this dataset we were mainly concerned with the following types of identification:

   - We assert that there are alternative interpretation of atoms, *such as ‘-‘, ‘N/A’ and ‘zero’.*
   - We assert that expected numeric data have been replaced by text, *e.g. on the transfer of spreadsheet values*.
   - We assert that data are missing. *E.g. middle of time series was blank*

3. **Transformation**

   Here the transformation was defined on the identified possible problem. Any given problem may have a number of alternative transformations.

   In the simple example, the problem is missing data in a time series. Two alternative transformations to solve this problem that were considered are

   - *Linear estimation between two known values (figure 4)*
   - *Moving average ignoring missing values*

   These were used because they are typical of the kinds of estimation used in this domain.

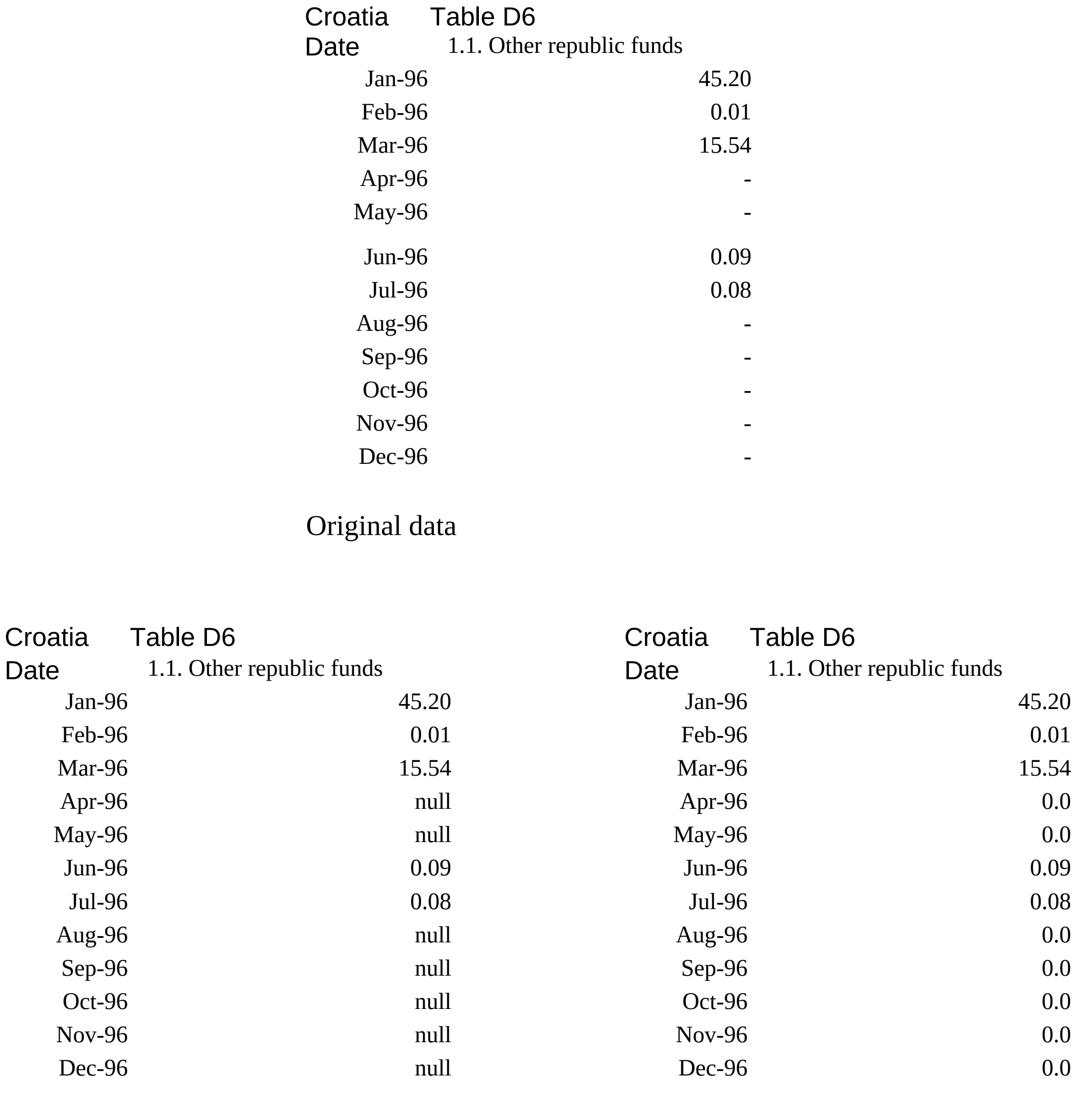
Date	Foreign assets	Credit to govt
Feb-95	19452.6	837.4
Mar-95	18547.4	1175.2
Apr-95	18529.9	1252.3
May-95		
Jun-95		
Jul-95		
Aug-95		
Sep-95		
Oct-95		
Nov-95		
Dec-95	17657.9	1577.9
Jan-96	17237.1	1142

▶

Date	Foreign assets	Credit to govt
Feb-95	19452.6	837.4
Mar-95	18547.4	1175.2
Apr-95	18529.9	1252.3
May-95	18420.8	1292.9
Jun-95	18311.7	1333.5
Jul-95	18202.6	1374.1
Aug-95	18093.5	1414.7
Sep-95	17984.4	1455.3
Oct-95	17875.3	1495.9
Nov-95	17766.2	1536.5
Dec-95	17657.9	1577.9
Jan-96	17237.1	1142

Figure 4: Polish dataset - estimating missing data. The missing values are replaced using a linear estimation function between two know values.

In a different example taken from the same domain, a time series from the Croatian data set has alternative interpretations of some of the atoms (either zero or missing). See figure 5. Depending on the interpretation chosen different transformations are needed (or not, in the case of the zero value interpretation).



Croatia	Table D6
Date	1.1. Other republic funds
Jan-96	45.20
Feb-96	0.01
Mar-96	15.54
Apr-96	-
May-96	-
Jun-96	0.09
Jul-96	0.08
Aug-96	-
Sep-96	-
Oct-96	-
Nov-96	-
Dec-96	-

Original data

Croatia	Table D6
Date	1.1. Other republic funds
Jan-96	45.20
Feb-96	0.01
Mar-96	15.54
Apr-96	null
May-96	null
Jun-96	0.09
Jul-96	0.08
Aug-96	null
Sep-96	null
Oct-96	null
Nov-96	null
Dec-96	null

Interpretation where '-' is missing

Croatia	Table D6
Date	1.1. Other republic funds
Jan-96	45.20
Feb-96	0.01
Mar-96	15.54
Apr-96	0.0
May-96	0.0
Jun-96	0.09
Jul-96	0.08
Aug-96	0.0
Sep-96	0.0
Oct-96	0.0
Nov-96	0.0
Dec-96	0.0

Interpretation where '-' is zero

Figure 5: Croatian dataset - example of alternative meanings of the '-' symbol. There are two alternative interpretations of the '-' shown here, zero and null.

4.3 Comparison of the Alternative Data Cleaning Schemes

For these simple examples, there was no significant difference between the alternative interpretations and transformations. That is, for the Croatian data, it makes little difference if you treat a dash as a zero or as a missing value. For the Polish data, the alternative methods of estimation made no real difference. For the analyst the conclusions drawn were the same as with the cleaning done by hand. This reflects the simple nature of these examples and the uncontroversial nature of the estimation used for the missing values. In the case of the Croatian data, many of the known values were close to zero so the alternative interpretations make little difference.

The principle benefit was the explicit representation of the choice made in cleaning the data. In comparison, data cleaning choices embedded in spreadsheets are quite opaque.

5. COMPLEX EXAMPLE – INTERNATIONAL MAILING LIST DATA

5.1 Survey of MBA Alumni

A market research organisation wished to carry out a survey of MBA alumni by questionnaire on behalf of a client. The business schools were in many countries and each school could have alumni in many different countries. There was not an opportunity to carry out a prototyping run. This section documents what was done in a practical situation with urgent deadlines and was not a planned scientific experiment. In this case, there were multiple uses of part of the data. Name and address data received from the business schools were used to distribute the questionnaire. The mailing lists supplied were in several file formats. They differed in name, number and type of fields for each record. However about half the business schools were unwilling to release directly alumni information so they distributed the questionnaire.

In this case, there was an operational need for the data, namely for use to mail questionnaires initially. Since the questionnaire pack was expensive, there was a high price to pay for a mail shot that contained a substantial number of failed deliveries. The primary aim of the cleaning was to produce a single mailing list with a reasonable chance of the questionnaire reaching the addressee.

The analytical needs were to provide estimates of:

- sex balance of schools
- the degree of ‘internationalisation’ of business schools

These estimates are to be cross-referenced with data from other sources to check for broad agreement of results and flag any further discrepancies for further study.

This example was included in our study because of the different purposes, and hence different validation rules which the same data needed to satisfy in different circumstances. In addition, one of us (PJ) was active as an analyst in this survey as well as being responsible for the cleaning.

5.2 Validation Alternatives

Distinct validation functions were defined for each of the three intended uses. The two analytical aims are intended to support conclusions drawn from other data. Therefore, the cleaning can be quite flexible. In the case of estimating the sex balance, it is reasonable to allow for a number of unknown values. In the case of internationalisation, we can accept as valid large numbers of unknowns by making strong assumptions during analysis.

The operational use of the data has slightly different criteria. The aim is to reach as many of the alumni as possible without obviously wasting money. So a record with no address lines can be excluded, but one which may have some flaws but which looks like it might arrive is included. The vast majority of addresses were American followed by the UK. American and UK addresses were validated by the presence of a post/zip code pattern. For example, UK postcodes consist of letters and digits. One pattern is “LLDD<space> DLL”, another is “LLD<space>DLL”, and so on. Therefore, if an address has a legal post/zip code pattern for its country, it is treated as a good candidate for posting. Non-American and non-UK addresses were not subject to validation of post code and simply tested for the presence of some kind of address.

It is often the case that there are a number of possible validation functions with greater or lesser degrees of similarity. The preference of one over another would reflect the specific aims of a project. If this project had required that an address be only used if it were known to be correct (maybe because the cost of each mailed packet was high), we could have defined a stronger validation function. For example, the record must have a post/zip code pattern and the values must be in a database of legal post/zip codes.

5.3 Identification and Transformation

The Identification and Transformation are different for each of the validation functions. We consider each in turn:

Internationalisation
For each mailing list, the aim is to infer a country field. This is used to provide a metric to estimate the degree to which a school is global rather than national in its recruitment and the willingness of its alumni to work internationally. For some mailing lists, a country field is defined. For many of them, a country name is embedded in fields containing the address. (See figure 6.) These require a pattern match against country names. There is a strong assumption made that if a country is not mentioned then the address is for the country of the business school. Note that address data can be missing or incomplete in this case. Missing address values are treated as having the same country as the business school. An alternative would be to treat them as unknown.

Sex balance
For each mailing list, the aim is to infer a sex field with one of three values: male, female and unknown. Broadly, the mailing lists were split between those that had some kind of title field and those that did not. (See figure 7.) Different titles are associated with different values for the sex field. For example, Mr goes with male, Ms with female, Dr with unknown. The mailing lists with no title field can be sub-divided into those with a title embedded in a name field, those with a separate first name field and those with neither. Possible title patterns are identified in the first group and the value of the sex field is inferred as before. First names field values are sorted and unique values listed. Then a person evaluates each case and associates a sex field value with it to create a lookup table. This is then used to infer a value for the sex field. In the final case, because of the limited resources available for this project (i.e., time) these were set to the default value of "unknown". Note that missing records were not a significant issue in this case because the name was present for all records.

Posting decision
For each mailing list, a postable field is set to either yes or no. Where the mailing list was for a non-American or UK school, this is set to yes unless there are no address data. For the American and UK mailing lists, if there is no indication that the address is for another country, i.e., a country field has a value or there is a country name in an address field, then the address fields are tested for the presence of a postcode pattern. If a post/zip code pattern is found then the postable field is set to yes. Otherwise, it is set to no.

5.4 Comparison of different Data Cleaning Schemes

Different cleaning scenarios can lead to the discard of different records:

Sex balance
Rather than generate different conclusions, the differences in the rules used produced different numbers of unknown values in the sex field. For example, just using the title field produces results for a few mailing lists that we can feel very confident about Unfortunately many of the mailing lists do not have this field. Looking for a title in a

name field results in fewer unknowns but a greater chance of error in the inferred values.

Internationalisation

The main area of error is amongst addresses outside the country of the business school but with no reference to that country. Typically, the city is obviously foreign (with respect to the business school) but there is no country name. So a mailing list for an American business school includes Istanbul but does not mention Turkey. This was treated as acceptable error for the purpose but is an obvious weakness compared with human cleaning. Likewise, the difference between missing addresses being treated as "home" addresses or as unknown, where unknown means the record is ignored, provided an indication of the range for our internationalisation metric rather than alternative results.

Posting decision

Instead of a range of alternative posting lists, the rejected records were considered by a human being for inclusion despite failing the validation. However, because the number of failed records is much smaller than the complete mailing list this task is made much easier.

In summary, we found that the adoption of the Clean Views model made it easier to focus on the particular purposes we had in generating a particular cleaned dataset. Iteration using revised transformations for the same identified validation failure retained clarity of purpose and process.

6. TOWARDS A DATA CLEANING MANAGEMENT SYSTEM

6.1 The Utility of Clean Views

Clean Views concerns recording explicitly the decisions taken in cleaning the data. For small datasets, employing the Clean Views framework can be more expensive than hand cleaning and this may prove to be a practical hindrance. However, effort is now channelled into software development rather carrying out the data cleaning manually. Having an effective support system means that you can afford to reclean data.

6.2 Using an Appropriate Language

In the examples used in this paper, the cleaning rules were coded in VBA. Earlier explorations of these ideas used Prolog and Clips. An example of a cleaning rule would be the one that replaced "-" with a zero for a given time series. In the work done here the code for specifying the action and the conditions is integrated with the code to apply the rule. Ideally, we would like to separate out the code for the rules and their conditions from that for applying them.

Having separated the application from the expression, one might consider how easy it is to express cleaning rules in different languages. Ideally, data cleaning rules should be expressed in a language in common use for manipulating databases, i.e., SQL or similar. However, we can easily imagine situations where SQL is limited. One example is expressing patterns in the form of a regular expression (where a language like Perl is very good). The issue here is identifying what additional power a language like SQL needs to enable rule-based data cleaning using Clean Views.

6.3 Organising Data Cleaning Schemes in a DBMS Framework

Within the Clean Views Approach, all data cleaning schemes share the following elements: validator functions, base tables, and transformation rule sets. This means that a standard framework can be adopted. Also data cleaning is a mainstream task, so the aim should be for it to be supported by extensions to DBMS functionality and be implemented where possible using DBMS functions. There is a need to be able to show the effects on a Clean View of alternative Data Cleaning Schemes. This means that it should be possible to retain the transformation functions across cleaning iterations. It should be feasible to create a template for the validation rules, and maybe the transformation rules that relate the analysis tools, and tune these to particular implementations. It should be feasible to extend this use of templates for rules appropriate to common business fragments. An effective user interface will be essential since tracking iterations of business aims and of alternative cleaning rules could become prone to substantial confusion.

Ideally a deductive database is needed, but they are not readily available or in common usage. SQL itself does not have sufficient power and therefore depends on being embedded in a procedural language. There is then the problem of rules embedded in the procedural code rather than the explicit rule set. Further consideration is needed of the extent to which one can settle for triggers placed in the database.

7. CONCLUSIONS

What you do about cleaning data depends on what you believe about the business. Any documented plan for cleaning data has to be better than no plan at all. The Clean Views model introduces a structured framework for carrying out data cleaning in a systematic, documented way. The investigations carried out so far indicate that the basic ideas are sound and worth developing into an effective data cleaning tool that is user tuneable.

The Clean Views model moves effort from cleaning the data themselves into a system development framework of formulating criteria and rules for use by a tool. People are unlikely to get faster at data cleaning, whereas we can expect them to get better at developing cleaning systems with appropriate tools.

ACKNOWLEDGEMENTS

It is a pleasure to acknowledge the support, help, and data provided by Parminder Bahra of Loughborough University with the transitional economy data and by Tak Lo Tsim and Jo Parry of London Guildhall University for the HESA data. PJ was supported by a Teaching Research Studentship from London Guildhall University.

REFERENCES

[Ba99] **Simulation experiments and the microeconomic effects of financial reforms in Poland and Estonia,** P Bahra, C J Green and V Murinde in *Economic performance and financial sector reform in Central and Eastern Europe*, ed. A.W. Mullineux and C.J. Green, 1999, Edward Elgar, Cheltenham, England.

[BPBK87] **Implications of Data Quality for Spreadsheet Analysis,** D P Ballou, H l Pazer, S Belardo and B Klein, Data Base 18 (3) 1987 pp13-19.

[DKCJR96] **Managing Heterogeneity in Medical Information Systems,** M Dixon, J Kohoutkova, S C Cook, K G Jeffery and B J Read *in Proceedings of the 10th ERCIM database research Group Workshop on Heterogeneous Information Management*, Prague 1996, ERCIM-96-W003:7.

[F97] **Data Preprocessing and Intelligent Data Analysis**, A Famili, Wei-Min Shen, R Weber and E Simoudis, Intelligent Data Analysis, 1 (1), 1997, Elsevier (http://www-east.elsevier.com/ida/).

[HESA98] **Coding manuals and Circulars**, Higher Education Statistical Agency, 1998 (http://www.hesa.ac.uk/manuals/home.htm).

[HS98] **Real-world Data is Dirty: Data Cleansing and the Merge/Purge Problem**, M A Hernandez and S J Stolfo, Data Mining and Knowledge Discovery 2, p1-31, 1998.

[J99] Private Communication - Paul Jermyn 1999.

[L98] **A Step by Step approach to Data Warehousing**, K Lewin of Ardent Software, Unix and NT News Supplement: Data Warehousing, p2, 1998.

[MS97] **Uncertainty Management in Information Systems: from Needs to Solutions**, Ed. A Motro and P Smets, 1997, Kluwer Academic Publishers, ISBN 0-7923-9803-3.

[N97] **Strategy and the Role of IT in Insurance,** Figure 6.3, A Newman, 1997, Pearson Professional Ltd, p105-109.

[P99] **Data Preparation for Data Mining**, D Pyle, 1999, Morgan Kaufmann Inc., ISBN 1-55860-529-0.

[R97] **Data Mining Leading Edge: Insurance & Banking**, D Romano in *Proceedings of Knowledge Discovery and Data Mining*, Unicom, Brunel University, 1997.

[T98] **Trillium Software System,** Harte-Hanks Data Technologies (http://www.trilliumsoft.com/1998).

Table 1

UID	Name line	Address1	Address2	Address3	Address4	Address5
834	Carlo Alveri	Viale Menido 6	I 20566 MILANO	ITALY		
845	De Heer Allard Blommers	Consultant	The Chicago Consulting Group	J.F. Kennedylaan 100	NL 3751 EH Baarn	THE NETHERLANDS
789	Ms. Ann R. McIntyre	P.O. Box 384	Delaware	OH 43025		

▼

Table 2

UID	Name	Country
834	Carlo Alveri	ITALY
845	De Heer Allard Blommers	THE NETHERLANDS
789	Ms. Ann R. McIntyre	USA

Figure 6: Complex example 1: Cleaning data for country. Table 1 is an example of a source table for the country data. Table 2 is an example of the clean table with data for country (inferred is the case of the third record).

Table 1

QID	Extrnl Indiv Name
345	Mr. William T. Aming
456	Ms. Ana P. Kansas
567	Mr. Brad Cohen
678	Mr. Peter F. Haas
789	Ms. Ann R. McIntyre
799	Dr. R Bahra
823	Mr. Kyriakos Amkouris

Table 2

QID	Mailname	Sal	Fname
834	Egr.Sig. Carlo Alveri	Egr.Sig.	Carlo
845	De Heer Allard Blommers	De Heer	Allard
867	De Heer Arthur Filius	De Heer	Arthur
897	De Heer Bart Creyghton	De Heer	Bart

▼ ▼

Table 3

QID	Name	Sex
345	Mr. William T. Aming	M
456	Ms. Ana P. Kansas	F
567	Mr. Brad Cohen	M
678	Mr. Peter F. Haas	M
789	Ms. Ann R. McIntyre	F
799	Dr. R Bahra	U
823	Mr. Kyriakos Amkouris	M
834	Egr.Sig. Carlo Alveri	M
845	De Heer Allard Blommers	M
867	De Heer Arthur Filius	M
897	De Heer Bart Creyghton	M

Figure 7: Complex example 2 - Cleaning data to infer sex. Tables 1 and 2 are examples of the source tables for the sex data. Table 3 is an example of the clean table.