

*Investigative
Data
Mining
Limited*

41 Madeley Road
Ealing
London
W5 2LS

Telephone
020 8997 1993
020 8566 7349

Fax
020 8810 7340

Large Scale Data Mining

A case study

Background

Investigative Data Mining Limited (“IDM”) has been advising and assisting its corporate clients since it was formed in 1998. During that time it has acquired a reputation for developing and delivering innovative, cost effective solutions using data mining techniques. These methodologies and solutions assist our clients in assessing corporate risk and exposure proactively (hence satisfying corporate governance requirements outlined in the Turnbull report) and also reactively as part of an ongoing investigation.

The following case study outlines how IDM’s data mining methodology and experience was used to assist one client in a major criminal investigation where there were no existing computerised databases to analyse.

Outline of criminal case

As part of a criminal investigation, our client had raided the offices of the alleged fraudster and seized the contents of numerous filing cabinets. It was believed that the mechanics of the fraud and quantification of losses could only be established by a detailed examination of each document. Given the volume of documents, conservatively estimated as 12,000 (in reality over 26,000 were processed over a period of 75 working days), it was agreed that some form of database would be required and that information from the documents would need to be recorded before any analysis could be performed. IDM was given the task of creating the database, ensuring that the documents were treated in an evidentially sound manner and performing preliminary analysis.

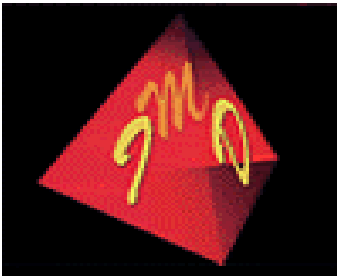
Key Objectives and Requirements

Given the nature of the case and the litigation support required by the client the following key objectives were agreed.

1. Document handling must be evidentially sound and maintain a documented evidence trail

Since the seized documents were key exhibits in a criminal case, they could not be altered or marked in any way. Some of the documents were stapled into bundles that made automatic photocopying or scanning impractical. Staples could not be removed as this could expose the process to the accusation that documents were lost, added or that the order in which the documents were presented had changed.

In analysing the information it was determined that there were eight different types of key documents. In some cases the information was typed, in others it was hand written or the bundles already contained photocopies. This ruled out automated optical character recognition (“OCR”) scanning as an input mechanism, (every scanned documents would still have to be manually reviewed to ensure that the scanning software correctly differentiated characters such as “S”, “B” and “8”). A manual process was therefore devised which agreed with the requirements of exhibit handling and the team leader kept contemporaneous notes. This process was independently verified and agreed by the Police who were working closely with our client.



*Investigative
Data
Mining
Limited*

In analysing the information it was determined that there were eight different types of key documents. In some cases the information was typed, in others it was hand written or the bundles already contained photocopies. This ruled out automated optical character recognition ("OCR") scanning as an input mechanism, (every scanned documents would still have to be manually reviewed to ensure that the scanning software correctly differentiated characters such as 'S', 'B' and '8'). A manual process was therefore devised which agreed with the requirements of exhibit handling and the team leader kept contemporaneous notes. This process was independently verified and agreed by the Police who were working closely with our client.

2. Computer Processing and Systems must be robust

As part of the chain of evidence, once information has been entered from the exhibits into a computer system the computer becomes part of the evidence chain. Given the size of the project it was decided to create four data entry workstations with the necessary software to capture the information. A deliberate decision was made NOT to network the data entry workstations. A fifth machine was designated as the "consolidator" which would merge the data from each machine on a daily basis and hold it independently. Each machine had a removable hard disk that contained the operating system and database. This ensured that at the end of the project the hard disks could be removed and sealed as exhibits with minimum cost to the client.

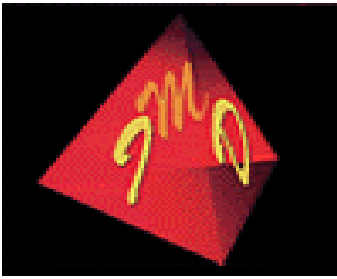
A network option was rejected for the following reasons:

- a) Prevented the possibility of cross contamination of data between each machine
- b) Reduced the requirement of a robust database management system with network security to prevent cross contamination
- c) Reduced development time and cost
- d) Ensured a higher level of information security and disaster recovery, so that if one workstation suffered a technical problem it would not affect the others

Data was transferred between the data entry workstations and the consolidator using floppy disks, the consolidator acted as the first level of system backup. Additionally, CD's were created each night from the Consolidator and held in a different location to ensure a further level backup.

None of the machines were connected to the Internet and no additional programs were loaded or run on the workstations. These machines were exclusively dedicated to this project, again preventing the possibility of accidental corruption of the data.

If our client requested reports or information from the system, these would be transmitted via email only once they were encrypted using PGP.



*Investigative
Data
Mining
Limited*

3. End of Day Reporting and Preliminary Analysis was an integral part of the process

As part of the database design and system requirements, a series of key reports and link analysis charts were agreed with the client. These standard reports were programmed into the normal end of day procedures when the data from each of the four data entry workstations was consolidated. In addition to standard reports to provide information for proceeding with the criminal investigation, quality assurance reports detected data entry errors which were checked on a daily basis to ensure quality control.

Due to the nature of the system, additional reports and ad hoc analysis could be performed without interfering with the primary data capture process.

4. Database and Data Entry Process had to be fully logged and auditable

To ensure that the database and the data input procedures were fully logged and auditable it was decided to create the underlying database using ACL for Windows. As a recognised piece of audit software it had a number of advantages over Windows based products such as Access or Excel.

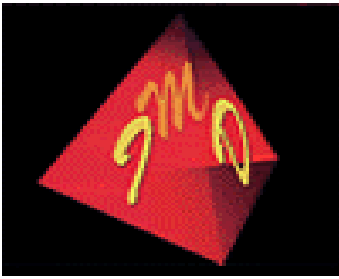
- a) Once entered into the ACL database, the data could not be accidentally altered or tampered with
- b) ACL maintained a complete audit log of each transaction including a date and time stamp for each transaction
- c) There was no upper limit to the number of transactions that could be recorded, therefore there were no sizing issues
- d) The system is fully scaleable and if it was decided to increase the number of data entry workstations, this could be achieved with minimal programming effort

In addition to system logs, manual procedures were in place to monitor and audit the activities of the data entry personnel.

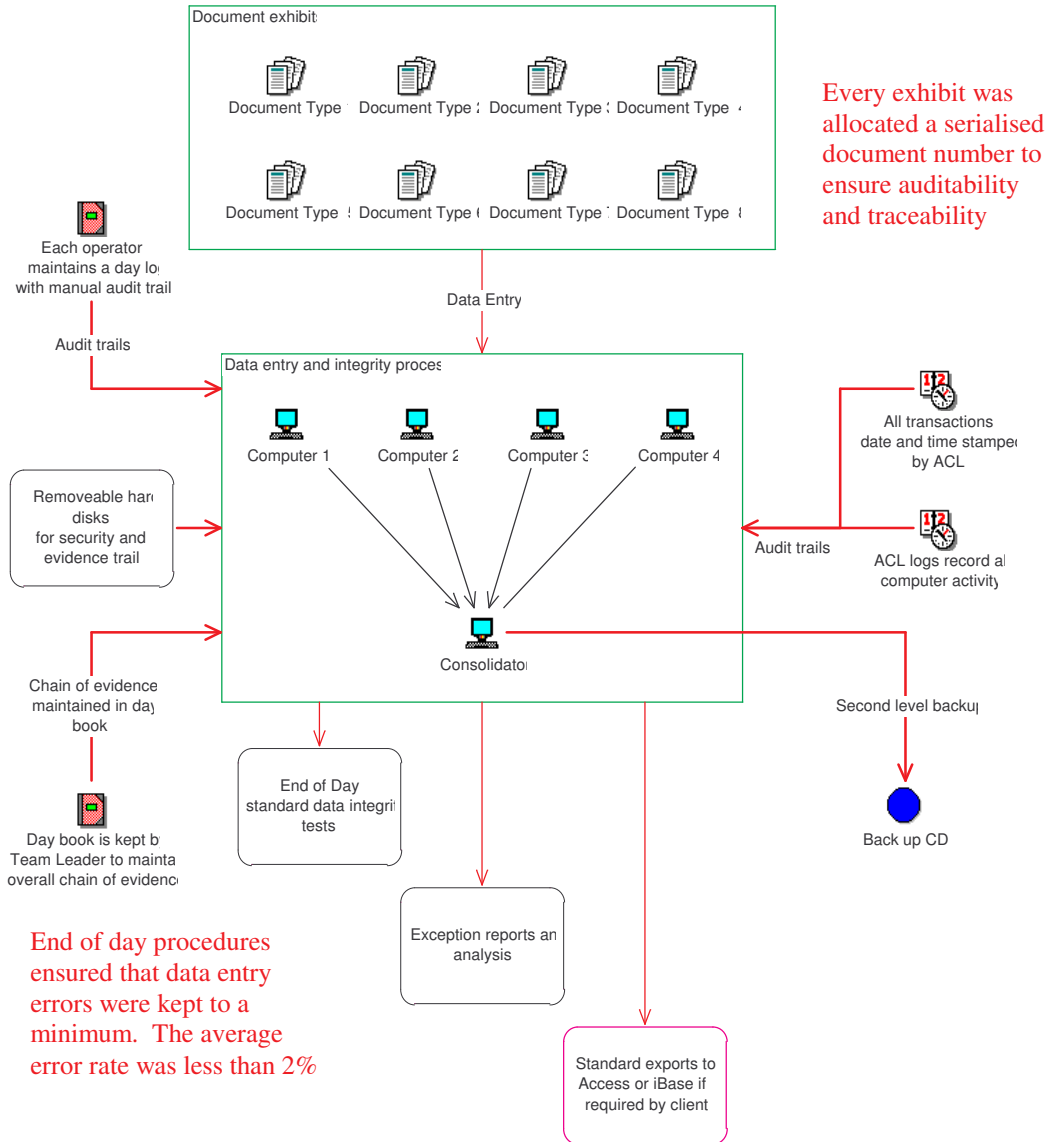
The project consisted of a team leader and four data entry personnel. The key task of the team leader was to ensure a consistent level of data entry and to conduct daily checks on the previous day's data entry. Again ACL's built in functionality was used to automatically:

- a) Ensure that the same document was not entered into the system twice
- b) Detect copies of documents that existed in different exhibits
- c) Detect any gaps in the data entry sequence
- d) Produce a random sample of exhibits to manually check
- e) Conduct "reality" checks, are dates realistic, have values been entered within appropriate ranges, etc.
- f) Produce a suite of standard reports which would be used as a basis for interviewing key suspects
- g) Produce data extracts which could be directly imported into link analysis software

The diagram overleaf illustrates the methodology adopted in the large-scale data capture and analysis project.

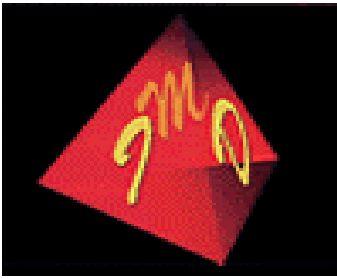


*Investigative
Data
Mining
Limited*



5. Day to Day Operating procedures had to withstand examination in court

Since IDM was providing assistance in a criminal case, all the procedures and methods employed in this investigation had to withstand cross-examination in court and be legally acceptable. All the facets of the investigation and data capture process were discussed with our client and the police, employees were vetted to ensure none had a criminal conviction; this was included in the confidentiality agreement that each team member signed. The integrity of the project was crucial and the start of the project included a training and familiarisation process for the project team members. Only software that was fully licensed was used.



*Investigative
Data
Mining
Limited*

Project methodology and its application to other key clients

We believe we have refined commercial best practice and consolidated a methodology that can be applied to other major investigations where our clients have to analyse large volumes of sensitive documents to gather useable information. IDM's experience in the field of data mining and fraud and risk profiling enables us to offer a tailor made, cost effect data capture and analysis service, which is consistent with evidence handling in criminal cases.

In the project described above the elapsed time from initial client meeting to the first day of data entry was less than six weeks. This included:

- a) **designing a bespoke data entry and consolidation system capturing 150 key fields**
- b) **testing and user acceptance**
- c) **review by the police for chain of evidence issues**
- d) **sourcing of hardware and software**
- e) **production of daily data integrity procedures and a suite of automated analyses and reports**
- f) **hiring of data entry team**

and agreeing the logistics for handling in excess of 26,000 criminal exhibits.

The overall cost for such a project will vary depending on the format of the information to be captured and number of documents to be processed. The following table will provide some basis against which a similar project budget may be established.

| Fixed start up costs | Cost |
|--|-----------------------------------|
| Database and exception reporting development (approximately) | £10,000 |
| Hardware and software rental (a one off cost irrespective of the duration of the project based on four data entry workstations and a consolidator) | £3,000 |
| Additional hardware requirement (5 removable computer hard disk drives) | £500 |
| Project running costs | Cost per week |
| Weekly costs for a dedicated team of four data entry personnel and one team leader | £6,500 |
| Overall project management and ad hoc system support and analysis | £3,000 |
| Document handling labels and other consumables (approximately £600 per 10,000 document exhibits) | £600 per 10,000 document exhibits |

For further information about this service, please contact any of the following:

London Office: Richard Kusnierz or Mandy Boylan on 020 8997 1933

Birmingham Office: Dick Price on 0121 441 3677

Edinburgh Office: Alan Livesey on 0131 225 7707

US representative in New York: Jane Bell on 0203 434 6780