Transforming data2knowledge

D2K Distil Solution Brief

Focus on Financial data

Source: Company Annual reports and other financial filings

Target: Key knowledge locked in tables and footnotes

Destination: Oracle 9i database[1]

D2K Distil, custom extraction tools are rapidly developed on a proprietary ETL platform, to extract specific information required by the customer. This solution brief describes how numerical data and the accompanying notes are quickly identified and extracted from financial statements in relevant source documents. The technology is able to automatically:

Ø Identify relevant tables and sections within the document
Ø Recognize rows and columns in tables
Ø Normalize terms and match notes to terms
Ø Map cells from table to database (70-90% automatic)

Human error is eliminated by highly accurate automatic processes during the repetitive data extraction stage. Supervision is introduced during the proofing and quality control stage to ensure 100% or near 100% accuracy, as proven in existing customer implementations from which this case study is taken.

D2K Distil Refinement Process

Step 1: Download
Step 2: Processing
Step 3: Quality control
Step 4: Delivery

Stage 1: Download

D2K Distil has the capability to access and automatically download dynamic data and documents from any website. In this case all DEF14 and 10-K filings are downloaded from the SEC Edgar website on a regular daily basis for processing. D2K Distil administrators are able to dynamically manage document sources and types, download frequencies, processing priorities and other variables, via specially designed interfaces.[2]

Stage 2: Processing

D2K Distil has a massively parallel infrastructure that allows immense scalability and speed of processing. Proprietary software supports the processing and proofing of any number of selected document batches by any number of data experts concurrently. All operators logged in to the system may see the work of others at a click, enabling immediate and efficient supervisor and peer review.

Identification (of relevant content)
Extraction (of relevant content and associated notes)
Normalization (of extracted data)
Biography extraction and Persons assignment

Identification of relevant content

Proprietary software automatically identifies all relevant tabular and textual financial information[3] within the document with around 85%[4] accuracy. This is then quickly refined to near 100% accuracy by D2K Distil experts.

Extraction of relevant content and associated notes

Variations in layout are automatically recognized and tables are split into header, rows and columns. This layout identification stage is visually verifiable on screen by D2K Distil experts. Then column contents are automatically recognized. The fiscal year is identified automatically from the header information.

Despite financial data values being exhibited in numerous formats; millions, thousands, full values, negative values in brackets, etc. D2K Distil is able to normalize these into full values and organize by currency if required. True or logical negative values are quickly recognized and stored as negative values in the destination database.

Notes may be in the form of an explanation of the figures in text form or a table, detailing the figures in the main table. Firstly D2K Distil splits notes into their individual parts. Secondly individual notes are associated with their rows in the appropriate table. For text, the process is complete. For detailing tables, the data is identified and extracted as for main tables.

Normalization of extracted data

D2K Distil uses an extensive financial statement terminology library to normalize extracted data. D2K Distil has learning capabilities, as a new variant of an item is encountered it is stored for future recognition. This process is supervised by a financial data expert. The system utilizes underlying knowledge of financial statement structure to establish an item hierarchy and the logical linkage between break-down and aggregate items.

D2K Distil captures all relevant content, any items not assigned during early processing are assigned a normalized collective term based on their position in the item hierarchy e.g. if an unidentified item is one of the break-down items of "short term investments", it is normalized as "other short term investments", if it is an asset item unassociated with any other, it is normalized as "other assets".

Each financial statement row will have two item descriptors stored, the original as used in the statements and the other normalized. Also the logical relationship between item rows will be fully mapped and stored, with each row having a pointer to the next level aggregate row.

Biography extraction and Persons assignment

Individual biographies are identified automatically and separated from notes or other text using specially designed D2K Distil interfaces. The same individuals from all extracted tables are automatically identified and mapped to a single person, together with biographical data. Human supervision is easily integrated at this stage to catch any variations in name, errors and formatting issues.

Stage 3: Quality control – the Human touch

Quality control is a combination of automatic functions and visual verification of data with feedback to continuously improve extraction parameters. There are automatic quality control steps built in to every stage of the extraction process, enabling D2K Distil operators to verify their work before moving on to the next document. At the end of processing the system uses several sophisticated automatic rule-set checks of data consistency to ensure the highest levels of accuracy.

Following the completion of a batch, all documents are visually verified by a D2K Distil expert supervisor, re-applying all automatic proofing rules. Weekly QA meetings amongst, experts, operatives and developers ensure any errors spotted during quality control procedures are quickly eradicated from future processing. This is done through an improvement of the automated data recognition algorithms or by the addition of newly defined proofing rules.

Stage 4: Delivery and communication

Feedback is sent daily from the D2K Distil to the customer and with greater frequency if desired. To ensure maximum security for the customer’s newly refined knowledge set, each project is assigned a dedicated supervisor. In this case the D2K Distil supervisor assigned to the financial information extraction project performs the following to customer specification:

- Saves the database to the desired data transfer format.
- Delivers the database through agreed channels, to a preferred location.
- Sends data processing, operational event and quality control reports.

In this particular case study of a US client, the location of d2k’s D2K Distil means that by the time office hours start on the US east coast, 6 hours of work has been accomplished on any previous day’s received documents. Delivery is typically twice a day, 8am and noon EST. In peak load periods, processing is extended to 8-9pm CET, 2-3 pm EST, with a third delivery at that time.

D2K Distil; Data accuracy & quality standards

D2K Distil provides the highest attainable data quality for the lowest cost. This is achieved through the automatic collection of data and the automatic application of proofing rules, as well as the continuous feed-back loop for software improvement.

It is not possible to give a single measure of data accuracy, as data quality is composed of several interlinked components. For financial statements data, the D2K Distil typically achieves the following:

Completeness – 100% error free

In the selection process all rows and all notes are captured, making the missing of any single data item technically impossible

Data item identification – close to 100% error free

Financial statements have a clearly defined fixed column structure with variation only in the row structure[5]. There is very little room for error and any identification errors are evident in the proofing phase if the statements don't balance or add up. Such errors are clearly highlighted in specially designed D2K Distil interfaces.

Data capture – 100% error free

Data is extracted from the source documents without any human touch, eliminating any possible copying errors.

Data item normalization – better than 99% error free

Two potential pitfalls which D2K Distil experts are trained to spot are mismatched and unmatched data items or terms. If the D2K Distil expert cannot make a positive match between mismatched or unmatched items it will be assigned to an appropriate "other" category. This is judged by a financial industry data expert and a new "normalisation rule" added to the financial terminology library.

Text matching; notes – close to 100% error free

Notes must be matched to either a row in one of the statements or to a particular cell. Note matching is a reliable and tested process with better than 99% success rates achieved with automation. In the unlikely event an item is mismatched or unmatched; this is corrected using human intelligence as described above.

[1] The database platform may be specified as required; in this case Oracle 9i is used.

[2] PDF, Word, HTML and image files can all be transformed with near 100% accuracy.

[3] Income statement, balance sheet, statement of cash flows and associated notes.

[4] Typically 70-90% accuracy, 85% representing a median percentage point.

[5] limited to the naming of a row and the implied underlying meaning of the data