Astera
DATA INTEGRATION MADE EASY

# Whitepaper

## Human-Readable Reports and the Data Trapped Within

## Overview

Reports are produced for a variety of reasons, ranging from summary reports for corporate decision makers to scientific reports on experiment data to credit reports of individuals. Such reports often focus on salient points, summaries, and aggregate data. They are meant to be read and reviewed by humans and follow a structure that is ideal for reading. Originally, these human-readable reports were the final destination of the data within, but in today's data-driven world there is a strong case for liberating the data trapped in these reports so it can be used electronically for further analysis or for integration with other IT applications.

Here is a broad categorization of such report documents:

- Exported reports (PDF, TXT, Excel, etc.) from software systems, including invoices, sales reports, inventory, and more.
- Printed reports that are scanned, run through optical character recognition (OCR) software, and saved as a searchable PDF file.
- Printer spool files from legacy IBM mainframe systems.
- Machine data logs.

## Report Data Extraction

The data from these report documents that is needed for electronic processes is often manually input by human resources—a tedious, time-consuming, and often error-prone task. For digital reports, IT is often tasked with writing scripts to extract desired data.

Programmers code or write scripts that use patterns in underlying data within report documents to identify data of interest. But there are some issues with this approach. For each scenario, the user must write scripts, test them, and maintain them. Even a slight change in the next set of incoming data will render the solution partly or completely useless and more code writing and testing must be done. This solution relies on the programmer or the programming team that coded it. Typically, however, requirements for data come from the business side of an enterprise. This necessitates multiple back-and-forth rounds between the business department and the IT department.

Additionally, in most cases the extracted data needs more treatment. It may need to be parsed, cleansed, and curated, then integrated as a data stream to other applications or to data storage for business intelligence analysis. This becomes another project and needs multiple tools and programming resources to achieve.

There must be an easier way. Is there a software-based solution that can automatically extract the desired data that is trapped in report documents? What would such a solution look like?

## Anatomy of a Data Extraction Solution

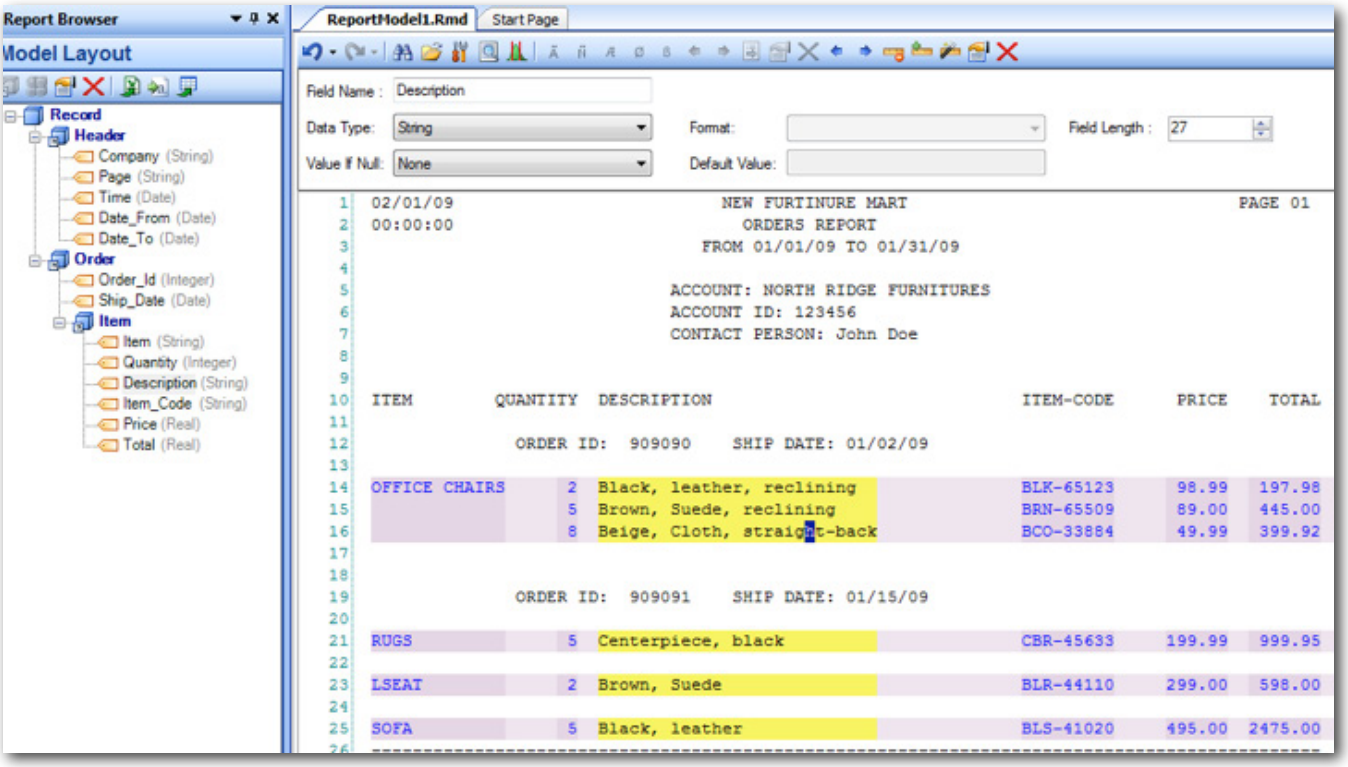
- This solution should be able to automatically build patterns or rules to identify data of interest from the document without the need for complicated programming.
- Data-pattern building should be robust enough to handle the most complex scenarios involving hierarchical (tree) data as well as flat data.
- Beyond the extraction logic, it would need to make sure that individual data fields are extracted properly.
- The field-building logic should be capable of handling fields varying in length and height, or fields floating inside the record area.
- After extraction of the desired records and their constituent fields, it should be able to perform data quality checks and conversion/transformation of the data if needed.
- Most importantly, in order to eliminate the vicious circle of business department/IT department rounds, the solution should be designed so that business users who may not be coming from programming background can do much of the work themselves.

## The ReportMiner Solution

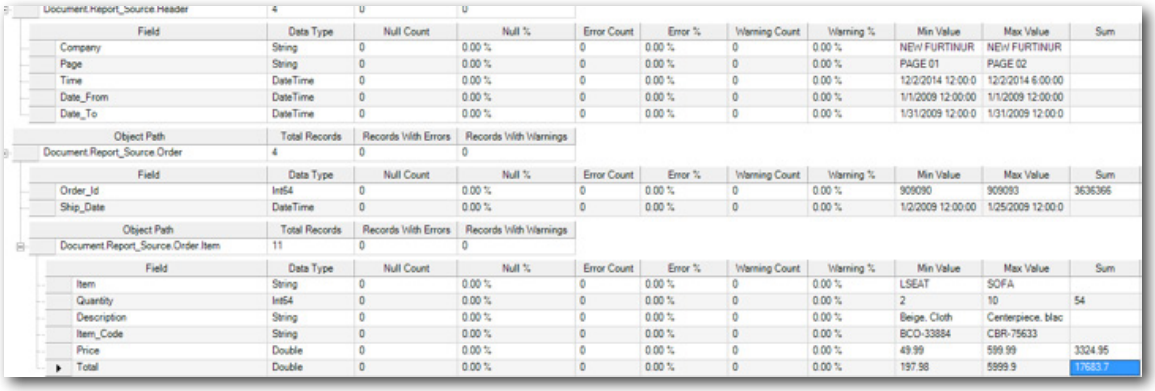

Astera took the above wish list and developed ReportMiner: a software solution that takes away all the pain from data extraction tasks.

ReportMiner automatically builds the extraction model. All the user needs to do is to point to some sample data.

ReportMiner even automatically figures out the data fields from the records, along with their data types, lengths, and formats.



Users can view the extraction model live with one click and see the resulting data.



| Document.Report_Source.Header | 4 | 0 | 0 |
|---|---|---|---|

| Field | Data Type | Null Count | Null % | Error Count | Error % | Warning Count | Warning % | Min Value | Max Value | Sum |
|---|---|---|---|---|---|---|---|---|---|---|
| Company | String | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | NEW FURTINUR | NEW FURTINUR | |
| Page | String | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | PAGE 01 | PAGE 02 | |
| Time | DateTime | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | 12/2/2014 12:00:0 | 12/2/2014 6:00:00 | |
| Date_From | DateTime | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | 1/1/2009 12:00:00 | 1/1/2009 12:00:00 | |
| Date_To | DateTime | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | 1/31/2009 12:00:0 | 1/31/2009 12:00:0 | |

| Object Path | Total Records | Records With Errors | Records With Warnings |
|---|---|---|---|
| Document.Report_Source.Order | 4 | 0 | 0 |

| Field | Data Type | Null Count | Null % | Error Count | Error % | Warning Count | Warning % | Min Value | Max Value | Sum |
|---|---|---|---|---|---|---|---|---|---|---|
| Order_Id | Int64 | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | 909090 | 909093 | 3636366 |
| Ship_Date | DateTime | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | 1/2/2009 12:00:00 | 1/25/2009 12:00:0 | |

| Object Path | Total Records | Records With Errors | Records With Warnings |
|---|---|---|---|
| Document.Report_Source.Order.Item | 11 | 0 | 0 |

| Field | Data Type | Null Count | Null % | Error Count | Error % | Warning Count | Warning % | Min Value | Max Value | Sum |
|---|---|---|---|---|---|---|---|---|---|---|
| Item | String | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | LSEAT | SOFA | |
| Quantity | Int64 | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | 2 | 10 | 54 |
| Description | String | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | Beige. Cloth | Centerpiece. blac | |
| Item_Code | String | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | BCO-33884 | CBR-75633 | |
| Price | Double | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | 49.99 | 599.99 | 3324.95 |
| Total | Double | 0 | 0.00 % | 0 | 0.00 % | 0 | 0.00 % | 197.98 | 5999.9 | 17683.7 |

With Astera's sophisticated data quality and transformation features behind it, ReportMiner can be used to transform the extracted data to the required standards. The resulting data can be sent to any destination of choice, such as an Excel spreadsheet, a delimited file, an Xml file, or a database table.

The entire process of extraction to conversion to writing is automated from end to end using ReportMiner. Extraction jobs can be deployed to run based on a schedule or triggered by the arrival of a new data file.

## Conclusion

ReportMiner solves myriad problems encountered when trying to extract and integrate data from report documents. It offers automatic creation of extraction logic, built-in data cleansing, standardization, and transformation features, and automated scheduling or new data triggering. Once extracted, data can be mapped and exported to a plethora of destinations, including databases like SQL Server, Access, MySQL, PostgreSQL, and any ODBC-compatible database, as well as file formats such as Excel, fixed length, delimited, and XML. The single-click preview capability shows extracted data and any conversion or validation errors, enabling users to verify and test extraction models as they are being built. Extraction models can be saved and reused for subsequent conversions, saving even more time. The Astera high-performance parallel-processing engine processes large data volumes quickly and efficiently.

Astera
DATA INTEGRATION MADE EASY

www.astera.com
Contact us for more information or to request a free trial
sales@astera.com 888-77-ASTERA