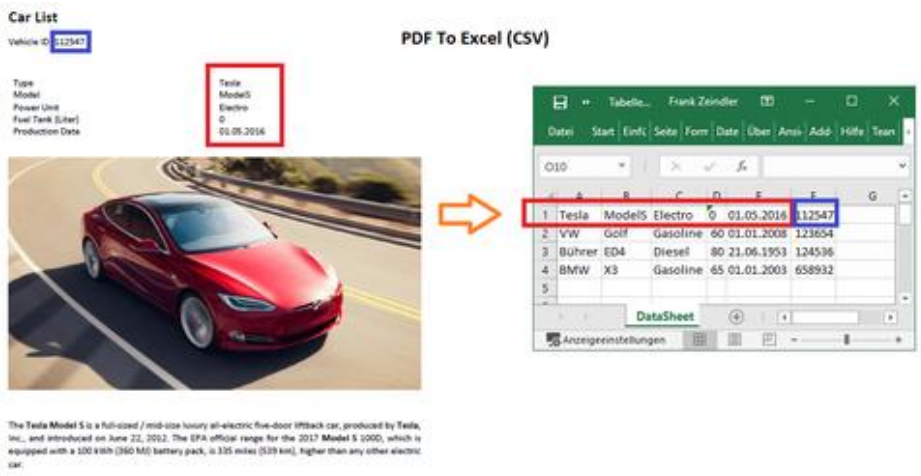
“How-to” No 1 –
Quality Assurance for Output Management

This guide describes how to compare the content of a PDF document with the underlying database content (e.g. Oracle, IBM etc.).

Step-by-step guide

Database content is usually available in a structured format and therefore easy to query. For data in a PDF, the situation is slightly different.
To make a PDF document readable by a data adapter and to compare its information with varying data sources, we first have to perform a conversion into a structured format.



The picture above shows what we would expect from a conversion to be the baseline for further processing, validation and automation.

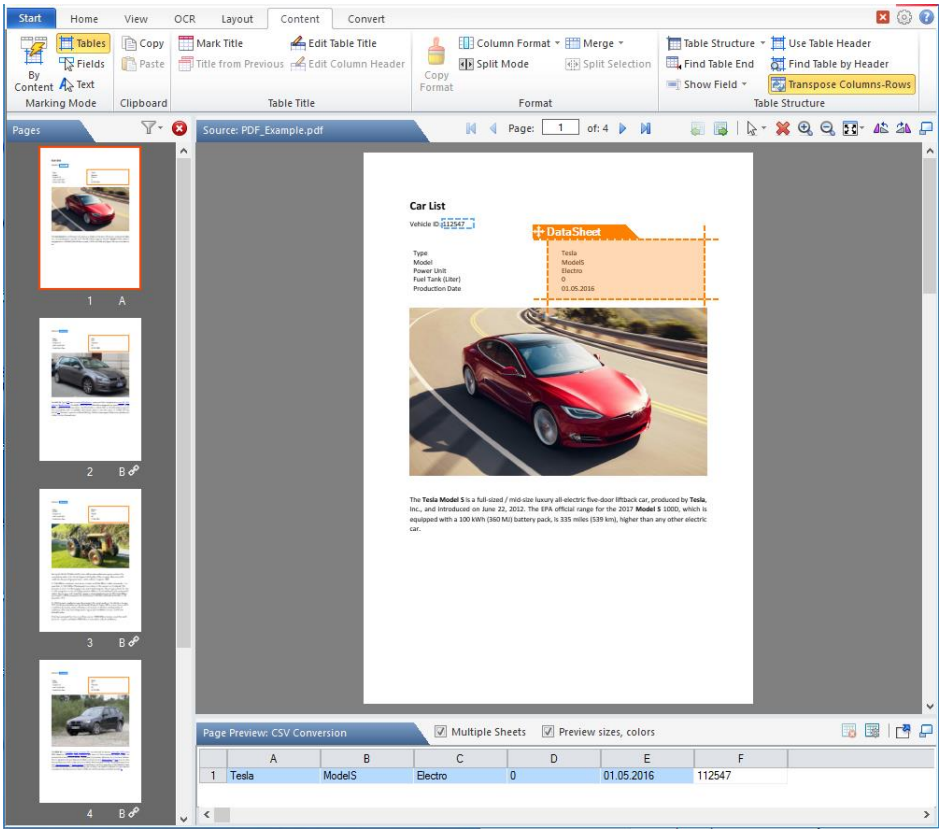
1 - Convert PDF data into structured data

To convert a PDF document to readable data we do need a converter.
Here’s the specification for such a converter:

- convert content stored in embedded tables within PDF documents
- manage predefined conversion templates
- is able to add text field content to all converted table rows as key attributes *(Why that? Because an ID makes subsequent error identification and correction easier.)*
- is able to run a conversion process on data tables with dynamic length over several pages
- is able to start conversion from batch process
- create a structured output format (e.g. csv, Excel)

2 - Create PDF conversion template

Once you have evaluated the optimal tool for your needs, your project to automatically check data integrity and data quality of the PDF output vs. the leading DB system can start.



We are using two different types to identify the pieces of information

- fields
- tables

Once the table and the field placeholders are configured, the tool should automatically identify similar structures in other pages and apply it to the rest of the document.

3 - Extract data from PDF

After defining the template for the PDF data extraction, we can generate the output manually or in batch mode.

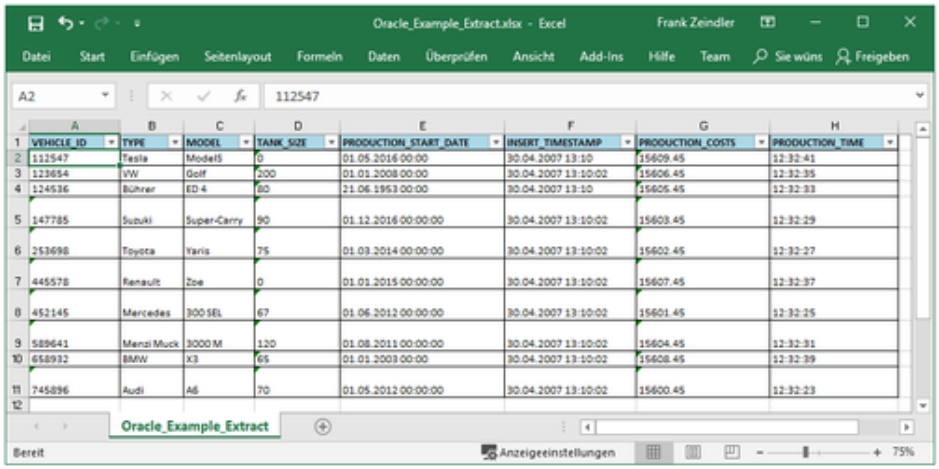
PDF_Example.csv

```
"Tesla";"ModelS";"Electro";"0";"01.05.2016";"112547"
"VW";"Golf";"Gasoline";"60";"01.01.2008";"123654"
"Bührer";"ED4";"Diesel";"80";"21.06.1953";"124536"
"BMW";"X3";"Gasoline";"65";"01.01.2003";"658932"
```

The extracted PDF data is now ready to be compared with content from a data base.

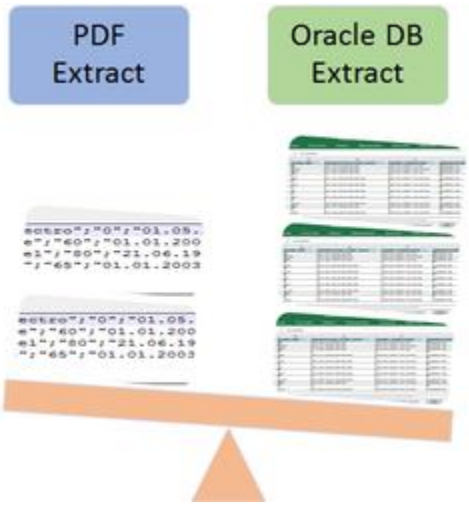
4 - Extract data from data base

This article does not describe how to extract data from a data base as it is quite easy to do so with the appropriate data adapters – and it is business as usual for data scientist.



5 - Compare data

Now here's what we wanted: we are all set to perform rule based and automated data quality and data integrity checks. This is possible because the data from the PDF is available in a structured format and therefore ready to be compared with the structured data from the database (in our example it is a Oracle database).



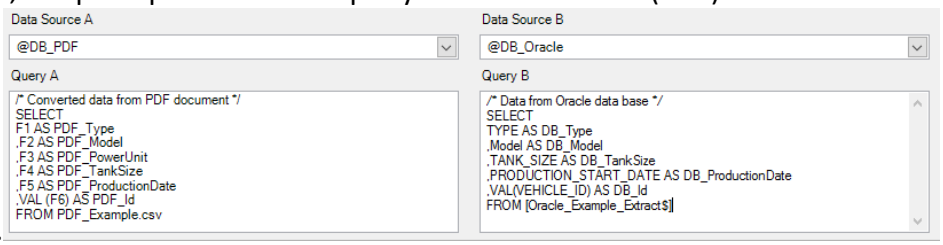
Our goal is to make sure the data in the two data sources are correct and complete. In other words: we are interested in both quality and quantity structure (coverage).

- Quality: Is the data on the PDF document exactly representing the data in the leading Oracle database?
- Quantity: Is the data on the PDF document complete, containing all models and specifications defined in the database?
- Overlap: Is there a data overlap in the PDF? In other words: is data printed on the PDF that must not be there?

To answer the above questions, you can export the data into Excel and create a compare script and do some "excel magic".
Or you can use more sophisticated tools like OMrun to run, manage and orchestrate automated comparisons. Such tools should also provide integrated reporting dashboards for management and audit and enable fully automated, recurring batch runs.

6 - Running a Query for PDF Data / Oracle Data

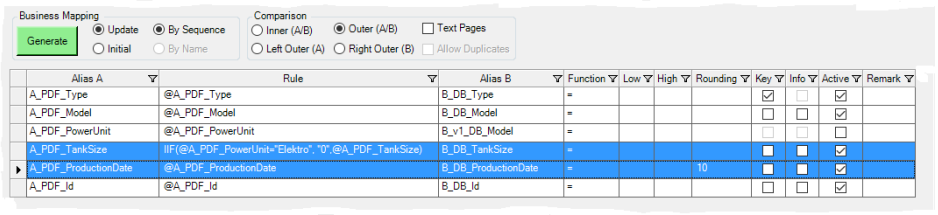
As the PDF document content is converted into structured data, it is ready to query. If you are using the framework OMrun, the principle is that one query runs on Source A (PDF) while on source B (Oracle) a similar query is executed.



7 - Define Business Mapping

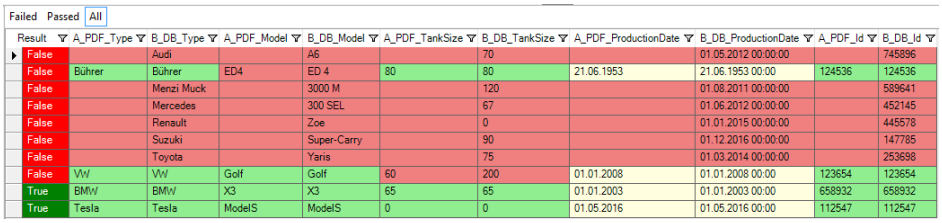
The business mapping contains the attribute mapping between source A (PDF) and source B (Oracle). For **TankSize** and **ProductionDate** (marked rows in the below screen shot) a business rule is defined and applied.

- **TankSize**
  IF PowerUnit IS "Electro" THEN TankSize = 0 ELSE TankSize = TankSize *(note: electro cars have no gas tank, so there is no size to be verified)*
- **ProductionDate**
  ROUND(ProductionDate, 10) *(note: because the data format of the two data sources is different, only the first ten characters are to be verified)*



8 - Analyze the Test Result

The result of the comparison (we also call it “report of rule violations”) between the original PDF data and the Oracle database is now ready for analysis.



Failed Passed All

Result	A_PDF_Type	B_DB_Type	A_PDF_Model	B_DB_Model	A_PDF_TankSize	B_DB_TankSize	A_PDF_ProductionDate	B_DB_ProductionDate	A_PDF_Id	B_DB_Id
False		Audi		A6		70		01.05.2012 00:00:00		745896
False	Bührer	Bührer	ED4	ED 4	80	80	21.06.1953	21.06.1953 00:00	124536	124536
False		Menzi Muck		3000 M		120		01.08.2011 00:00:00		589641
False		Mercedes		300 SEL		67		01.06.2012 00:00:00		452145
False		Renault		Zoe		0		01.01.2015 00:00:00		445578
False		Suzuki		Super-Carry		90		01.12.2016 00:00:00		147785
False		Toyota		Yaris		75		01.03.2014 00:00:00		253698
False	VW	VW	Golf	Golf	60	200	01.01.2008	01.01.2008 00:00	123654	123654
True	BMW	BMW	X3	X3	65	65	01.01.2003	01.01.2003 00:00	658932	658932
True	Tesla	Tesla	ModelS	ModelS	0	0	01.05.2016	01.05.2016 00:00	112547	112547

In this diff-report, four records from the PDF document are compared with ten records in the Oracle database.
Tolerance: Some data fields in the columns **ProductionDate** are highlighted with yellow background which indicates that the values are different but the rounding rule was applied successfully.
Two of ten records fulfill the defined business rules 100%, the other eight records do not exist in the PDF document or the data in the PDF document is not up to date.

- six car types are completely missing in the PDF document (Audi, Menzi Muck, Mercedes, Renault, Suzuki, Toyota)
- by running the queries with the function "Outer Comparison", all expected records are listed, even if there is no equivalent in the database or in the PDF document
- two car types passed the test successfully (BMW, Tesla)
- two car types are containing errors in the specification (Bührer - **Model** mismatch, VW - **TankSize** mismatch)

9 - Robotic Process Automation (RPA)

Robotic Process Automation refers to checking large amounts of data rule-based and automated with no or only little human interaction. The goal of RPA is to increase quality and processing speed and decrease maintenance effort and human interaction.
Here's the next steps in order to build up RPA for quality assurance of a PDF output versus the leading data source on Oracle.

- analyze the results from the comparison and take action accordingly
- split the results into separate partitions: one is input for RPA, one is to be tested manually because rules are too complex for low-maintenance RPA
- apply the RPA principle to the data where possible
- orchestrate the whole process in a tool with the features of OMrun to establish automated checking including reporting for management and audit.

If you have further questions with regards to your specific QA needs for your output management, do not hesitate to contact us. Don't you think that this proceeding could also be applied to check *account statements, invoices etc.?*

Happy automation,
your OMIS-Crew

Do you want to take the next step?

For further information please check out www.OMrun.ch or
ask the CTO frank.zeindler@omis.ch
ask the CEO marc.keller@omis.ch
call us +41 44 942 50 00