



VISUALIZING DATA QUALITY AUDIT WITH OPEN SOURCE

With Uniserv's data quality visualization technics one can underpin the cause of data owners and thereby provide a literal means to precisely highlight the data quality issues during the data science project initiation.



INFORMATION AND KNOWLEDGE WITH GOOD DATA

With global digitization, organizations are generating and using data at an alarming rate. It is so much in abundance that it can be found at every nook and corner of an organization. Still, the decision makers set aside multiple resources to manage data because they are aware of the treasure it holds within; Information & knowledge. By its application in data science, this precious resource becomes one of the decisive factors for achieving the desired outcome for an organization.

TABLE OF CONTENTS

01	DATA SCIENCE	04
02	GROUND TRUTH	05
03	DATA QUALITY AUDIT	06
04	DATA VISUALIZATION	07
05	RESOURCES	16

DATA SCIENCE

According to Hopcroft and Kannan, Computer science was introduced as an academic in the 60's with the emphasis on the programming languages, compilers and operating systems. Later on in the 70's algorithm and mathematical theories were added as an important subset. Now with increase in communication media and computation the focus has shifted to applications that can manipulate and visualize data in a meaningful way. Jeffery Stanton of Syracuse University defines data science as an emerging area of work concerned with collection, preparation, analysis, visualization, management and preservation of large collection of information. Illustration 1 depicts the necessary steps in a data science initiative.¹

Using data for predictive purposes and extracting knowledge is at the heart of a data science project, but this is only achieved during the last phase. The most crucial step in the process, is the iterative task of acquiring raw data, cleansing and transforming it in a format that is ready to be used for predictive modelling. In fact, a recent study published in 'the New York Times' states that 80% of a typical data science project is sourcing, cleaning and preparing data, while only 20% is allocated to finding and testing predictive models. To steer organizations through this bottleneck Uniserv has introduced the concept of GROUND TRUTH.

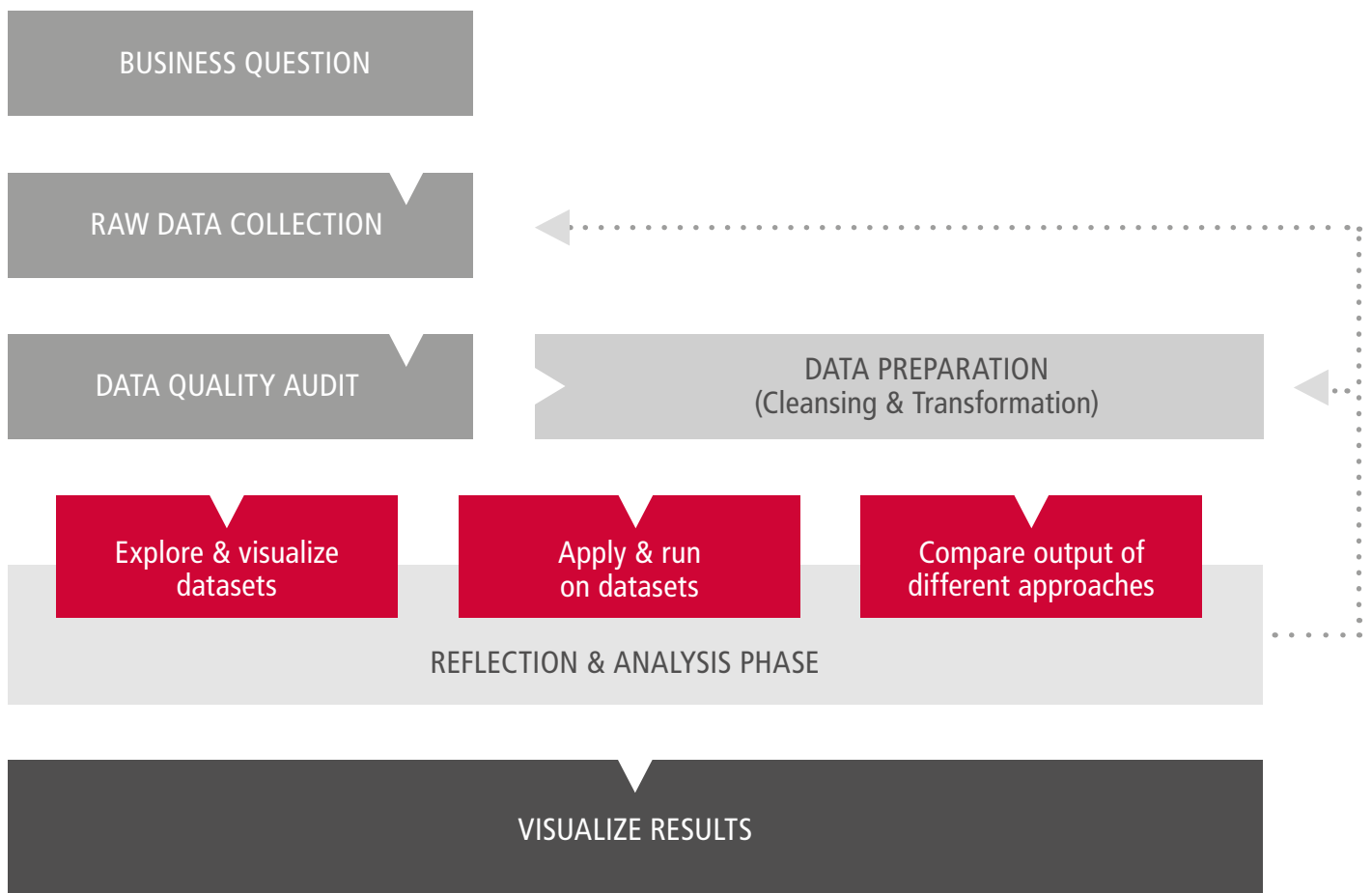


ILLUSTRATION 1: THE NECESSARY STEPS IN A DATA SCIENCE INITIATIVE.

¹ Similar to Philip Guo, 'Data Science Workflow: Overview and Challenges' (October 30, 2013), taken on 20.10.2016.

GROUND TRUTH

Data by itself cannot produce any logically correct information. To do so, the requirements have to be explored, necessary data has to be singled out and the appropriate algorithm has to be chosen. GROUND TRUTH is Uniserv's contribution to the Data Science process. It enables creation of a 360° view of the customer by:

1. Integrating various data sources in a central repository – the Smart Customer MDM.
2. Creating the so called „Golden Record“ – Merging of the master data of identical persons into a master record.
3. Creating the so called „Golden Profile“ – Enrichment of the master record with interactional and transactional data.
4. Matching (aggregated) transaction IDs to the Golden Record ID.
5. Generating complete, precise and up-to-date 360° view of the customer.

In most cases of predictive analysis, organizations are neither aware of their data quality nor of the factors affecting it. We at Uniserv believe that knowing your customer means to know your data. We therefore provide services to guide our customers through the process of getting to know their data.

1. DATA INTEGRATION

Integrating various data sources in a central repository - the Smart Customer MDM.



2. GOLDEN RECORD

Creating the so called „Golden Record“ – Merging of the master data of identical persons into a master record.



3. GOLDEN PROFILE

Creating the so called „Golden Profile“ – Enrichment of the master record with interactional and transactional data.



4. MATCHING

Matching (aggregated) transaction IDs to the Golden Record ID.



5. 360° VIEW

Generating complete, precise and up-to-date 360° view of the customer.



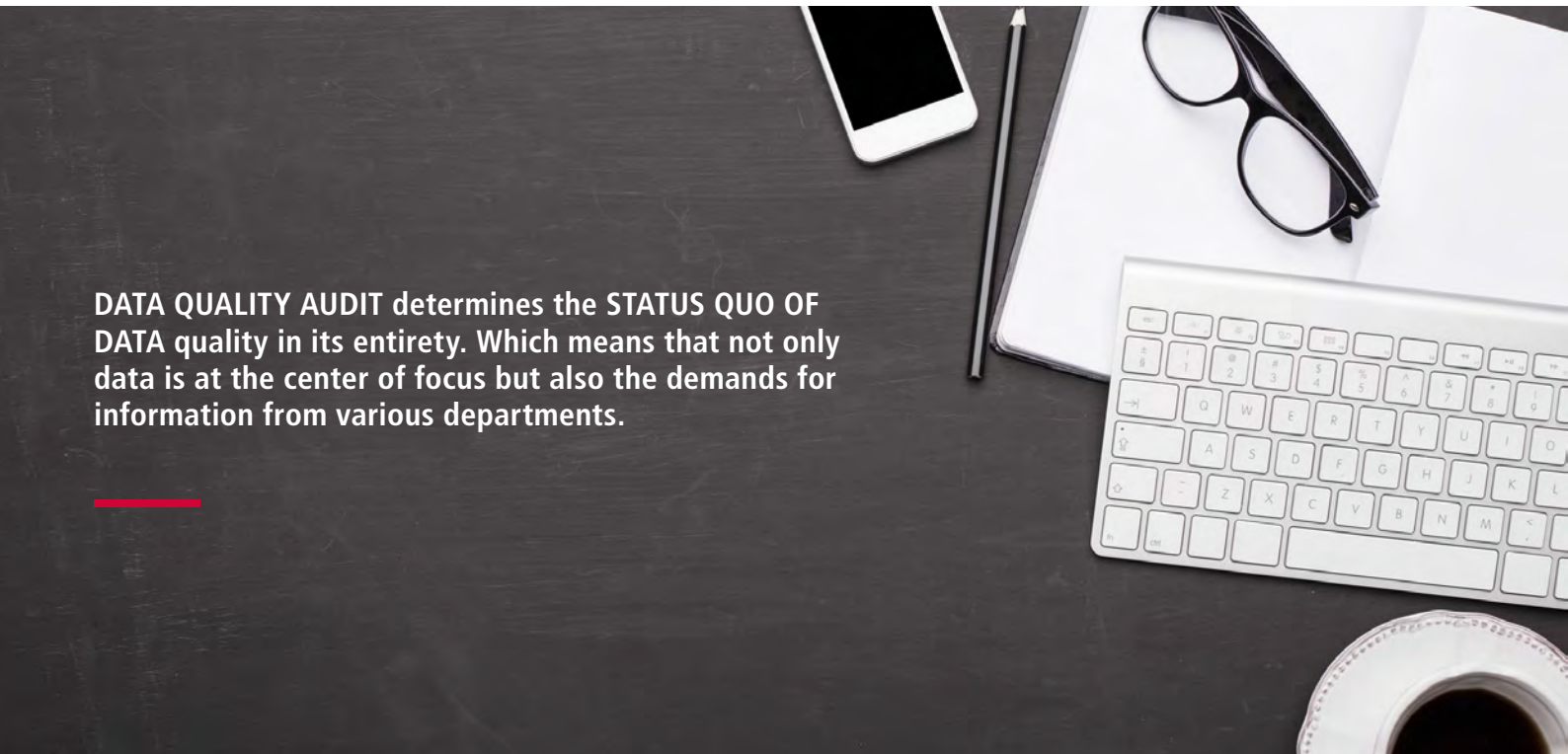
DATA QUALITY AUDIT

Everyday organizations are busy defining new strategies and taking new initiative to connect themselves with their customers. This results in each business unit working autonomously and dealing with data in multiple formats and systems. There is seldom a good overview across all business units. Therefore before data is made 'fit' to comply with different demands within an organization, a good overview must be created.

Data Quality Audit determines the status quo of data quality in its entirety. Which means that not only data is at the center of focus but also the demands for information from various departments. To achieve this, firstly the process is analyzed and the business aspect of the project and the current data resources are clarified through constructive communication. Then, the complete data set is analyzed using Uniserv's data quality tools; DATA ANALYZER & DATA QUALITY BATCH SUITE. Broadly the Data Quality Audit process can be divided into three sections: the first

is about data tidying, where we analyze the files we received from the customer. In a second step, we validate the data and compare the value added to the raw files. In the standard, we usually take a look at the outcome of postal validation and duplicate matching. Thus, the analysis may include one or more of the following:

1. Completeness of the datasets at field level.
2. Correctness of syntax.
3. Plausibility of address elements based on correctness and deliverability.
4. Singularity of data sets based on predetermined parameters.
5. Accuracy of address elements based on address relocation.
6. Consistency of data set across various source. Ex. Embargo check list or Robinson list.



DATA QUALITY AUDIT determines the STATUS QUO OF DATA quality in its entirety. Which means that not only data is at the center of focus but also the demands for information from various departments.

DATA VISUALIZATION



SAMPLE DATASET

For illustration purposes, we applied the script to a sample dataset of German households.



DATA TIDYING

The first section, data tidying, lists a few details about the general structure of the files.

Raw Data

- We received one extract from your SAP production system (In total 10487 records).
- Columns considered for postal validation are STRASSE, PLZ, STADT.
- The columns used for deduplication are VORNAME, NAME, STRASSE, PLZ, STADT.
- The total amount of data processed by postal validation and identity is 10487

Let's delve into a few special cases that have been discovered during the analysis of the raw input data:

Show entries

Search:

	NAME	STRASSE	PLZ	STADT
8333	DUMMY	Filderblickweg 26	70184	Stuttgart
8334	DUMMY	Buchhaldenstr. 2	75236	Kampfelbach
8335	DUMMY	Am Karbel 56	09116	Chemnitz
8336	DUMMY	Sybelstr. 7	04318	Leipzig
9933	DUMMY	Im Steingebiss 27	76829	Landau

Showing 1 to 5 of 15 entries

Previous 2 3 Next

ILLUSTRATION 2: SAMPLE DATASET.

The necessity to incorporate engaging visualizations in the Data Quality Audit arises from the need to share complex data with our customers: by using interactive graphs we are creating a structured interpretation path in a way that it both helps our customers to grasp the connection between their datasets as well as giving them the freedom to experience their data as they are moved by their own curiosity. Data tells stories, and visualizing it creates emotional connection to retain the information it conveys.

Advanced techniques for integrating data visualization into solution landscapes are becoming more common nowadays, with an increasing share being powered by open source technology. That is why we decided to use a varie-

ty of available open source tools to cater specific visualization needs.

When setting up a Data Quality Audit, we use R programming language to process the output of our data quality tools and connect to the visualization libraries. Originally intended for statistical purposes, the R programming language now comes with a vibrant ecosystem of libraries (called packages in R) to solve all kinds of data-related matters. R packages are constantly created and maintained by an ever-growing community. Leading software providers like Microsoft, IBM, and others have integrated R into their commercial product offering as well.

The Data Quality Audit framework is written in R Markdown, which allows converting its contents to HTML or any other format, thus facilitating the results to be shared with our audience.² One single R Markdown file contains a standardized script to visualize the data we have already processed and validated with Uniserv's data quality batch software – DQBT.

For illustration purposes, we applied the script to a sample dataset of German households (names were anonymized). The first section, data tidying, lists a few details about the general structure of the file(s) and the contents used for the upcoming analysis (illustration 2).

As a part of our consultancy work we detect systematically occurring anomalies in the data. The R package DT

provides a means to display these anomalies in a concise table.³ These DataTables provide functionalities to filter, paginate, sort and search through the contents (illustration 3). Hence, they are supportive in giving an overview of the polluted data and creating awareness for common data quality pitfalls.

As a part of our consultancy work we detect systematically occurring anomalies in the data. The R package DT provides a means to display these anomalies in a concise table. These DataTables provide functionalities to filter, paginate, sort and search through the contents (illustration 3). Hence, they are supportive in giving an overview of the polluted data and creating awareness for common data quality pitfalls.

Raw Data

- We received one extract from your SAP production system (in total 10487 records).
- Columns considered for postal validation are STRASSE, PLZ, STADT.
- The columns used for deduplication are VORNAME, NAME, STRASSE, PLZ, STADT.
- The total amount of data processed by postal validation and identity is 10487

Let's delve into a few special cases that have been discovered during the analysis of the raw input data:

Show 5 entries

Search:

	NAME	STRASSE	PLZ	STADT
111	tbd	Herzog-Wilhelm-Str. 9	84034	Landshut
112	tbd	Printzenhof	41366	Schwalmtal
113	tbd	Tulpenstr. 51	47906	Kempen

Showing 1 to 3 of 3 entries (filtered from 15 total entries)

Previous 1 Next



OVERVIEW OF DATA

The DataTables provide functionalities to filter through the contents. They are supportive in giving an overview of the polluted data.

ILLUSTRATION 3: GETTING AN OVERVIEW OF THE POLLUTED DATA.

2) <http://rmarkdown.rstudio.com/lesson-1.html>

3) The R package DT provides an R interface to the JavaScript library DataTables (<https://rstudio.github.io/DT/>, <https://datatables.net/>)



FILLING RATIO AS KPI

The quality of validation and matching highly depends on the filling ratio of the columns used in the steps. That is why we include this ratio as a KPI.



CHARTING LIBRARY

The methods of the library leverage hover and zooming abilities of the plotly API.

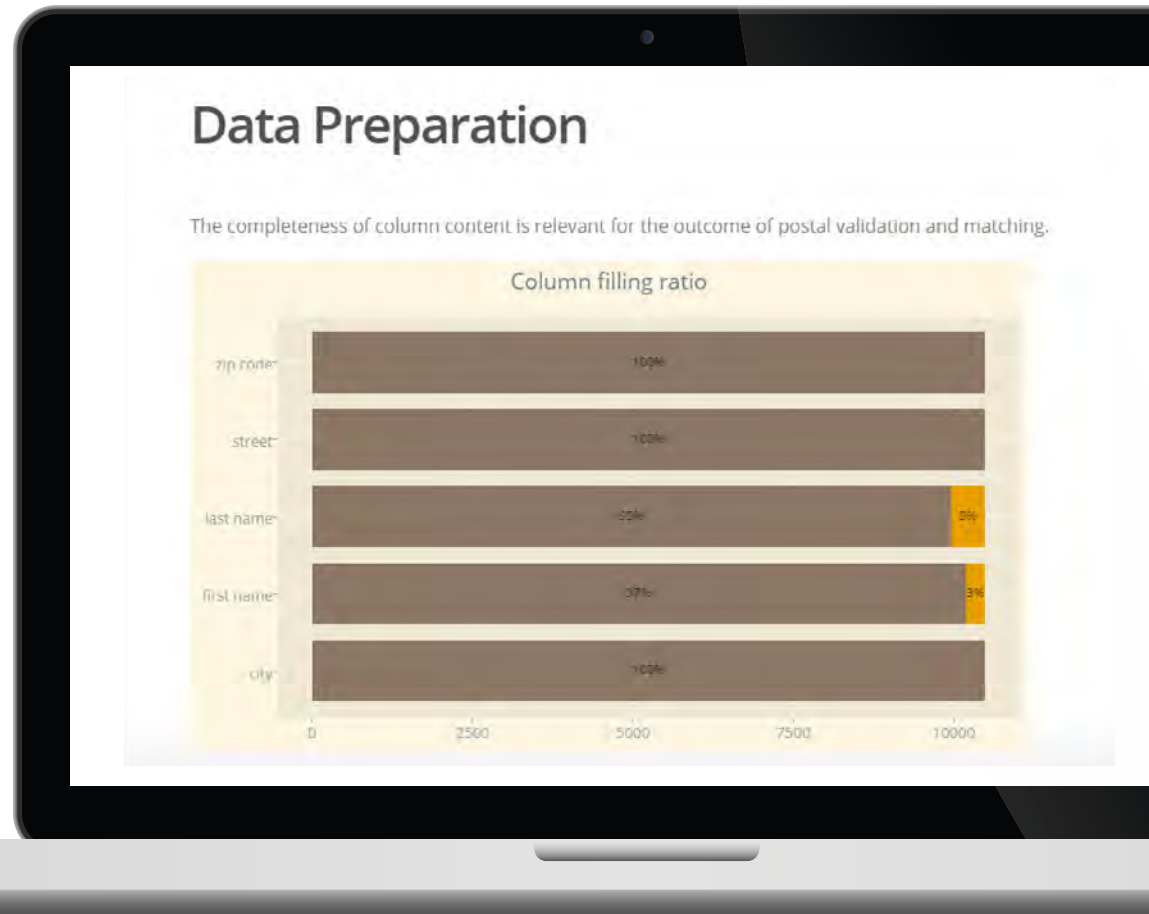


ILLUSTRATION 4: DATA PREPARATION.

The quality of validation and matching highly depends on the filling ratio of the columns used in these steps. That is why we include this ratio as a KPI. The horizontal stacked bar chart used to compare the proportion is powered by the plotly package for R. Plotly for R is an interactive, browser-based charting library built on the open source javascript graphing library, plotly.js.⁴ The methods of this library leverage hover and zooming abilities of the plotly API (illustration 4).

A sophisticated way to introduce these metrics on a broader scope is presented by Uniserv's DATA QUALITY SCORECARD. The scorecard provides a configurable rule engine to define enterprise wide data quality goals. Data quality KPIs can be generated over all records and all defined rules for different systems. With drill down functionality, the exact point of potentially improving data quality can be identified and actions to ensure data governance can be initiated.⁵

⁴ <https://plot.ly/r/>

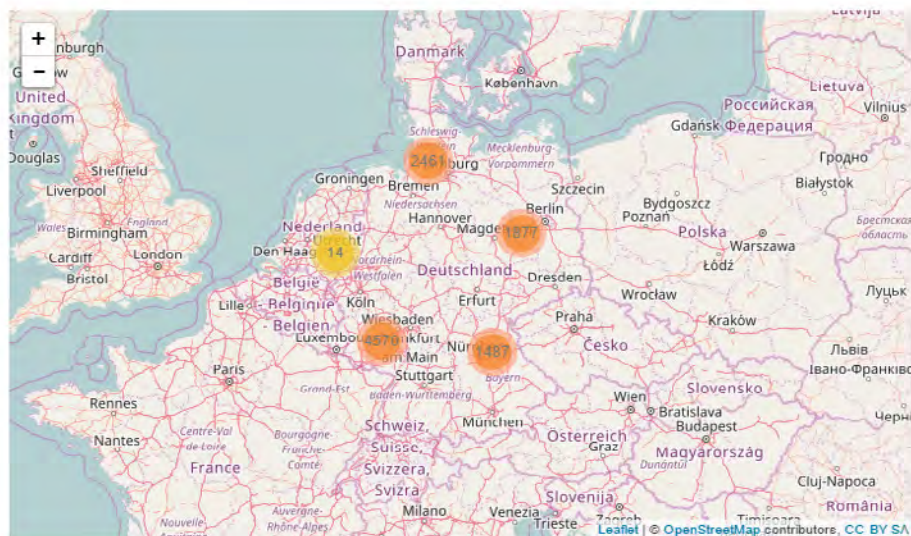
⁵ <http://www.uniserv.com/en/services/data-quality-scorecard/>

In a similar fashion, the audit pinpoints the benefits of investing in a data quality solution. To showcase the added value of having a data quality initiative, postal validation has become an integral part of every audit. In this step, we compare address data to external reference files, thereby analyzing their correctness and deliverability. In order to grasp the attention of our audience, we are wrapping the validation results in interactive maps. We chose to do this with Leaflet, which is one of the most popular open-source JavaScript libraries for interactive maps. Leaflet supports various third-party base maps⁶; for our test dataset, we chose to work with OpenStreetMap⁷.

We included three Leaflet charts to enable our customers to explore their own datasets; as they are led by their own

curiosity, customers will gain a better understanding of their data and retain the information for a longer time. The first one plots all valid addresses on a map. Since Uniserv's postal validation can enrich addresses by geo coordinates, this output is used to display the locations (illustration 5). Luckily, Leaflet comes with a functionality to cluster spatial points into an aggregated view of locations. These clusters even come with drill-down functionality: scrolling within the graph initiates zooming in and out of location clusters (illustration 6). This mode of presentation does not only deliver a breakdown into areas of commercial interest, but also evokes memories and pictures in viewers as they remember certain places, and links this attachment to geographical content.

Where is your customer located?



KNOW YOUR DATA

We want to enable our customers to explore their own datasets; as they are led by their own curiosity, customers will gain a better understanding of their data.

ILLUSTRATION 5: CUSTOMER LOCATION IN INTERACTIVE MAPS.

6) <https://rstudio.github.io/leaflet/>
7) <https://www.openstreetmap.de/>

Where is your customer located?

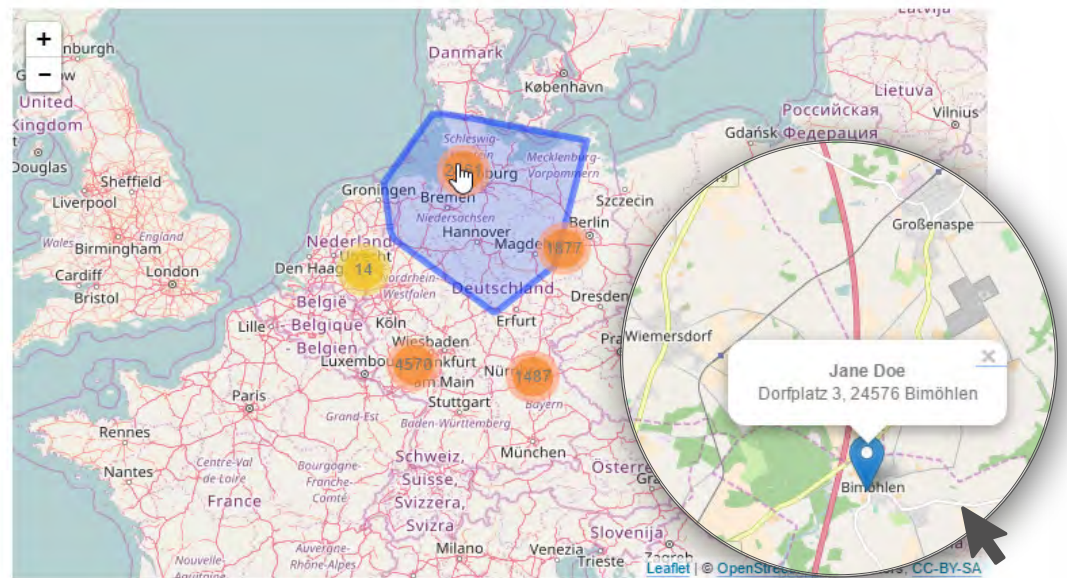


ILLUSTRATION 6: CLUSTERS COME WITH INTUITIVE DRILL-DOWN FUNCTIONALITY.



INTERACTIVE MAPS

The map comes with a functionality to cluster spatial points into an aggregated view of locations. These clusters even come with drill-down functionality.

The handling of the application becomes more intuitive for the user when familiar icons are applied; our dropped pin icon conveys the same meaning as in mobile navigati-



INTUITIVE APPLICATION

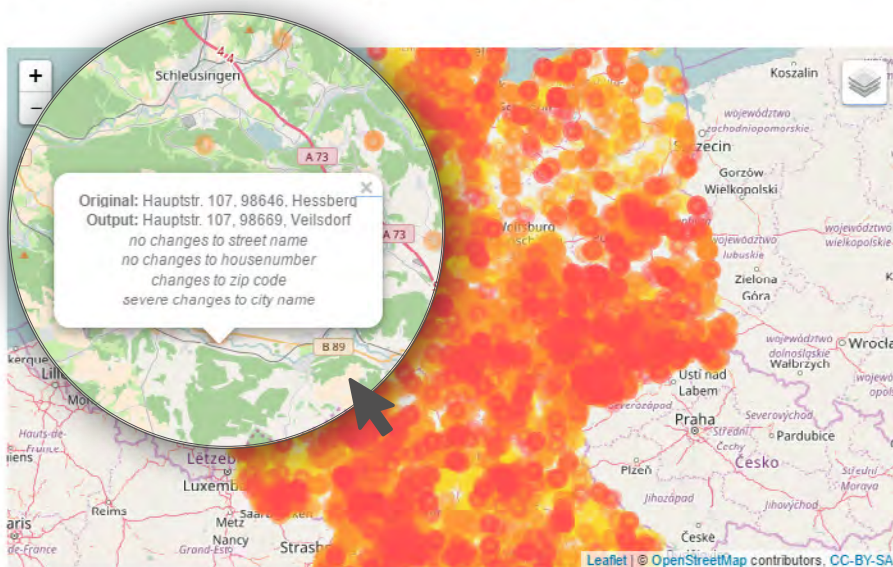
Our dropped pin icon conveys the same meaning as in mobile navigation applications; if a user clicks on it, a popup reveals the exact location.

on applications; if a user clicks on it, a popup reveals the exact location (illustration 6).

From a data quality perspective, an interesting follow-up to the question ‘Where is your customer located?’ is presented in the second Leaflet chart. Here, we are making the point that badly managed data may actually hamper one’s intention to reach that very customer. Since Uniserv’s postal validation evaluates the conformity between input and reference data, we can plot the deviations. We use the result flags from the output (indicating the degree of conformity to reference data) to create a visual, filterable distinction on the map (illustration 7). The coloring of the circles emphasizes the severity of changes.

Similar to the map pin, an annotation pops up when the user clicks on a colored circle. The annotation juxtaposes original and validated input, providing additional information about the adjustments at a glance (illustration 7). In this example, a city has been incorporated into a bigger municipality, resulting in both changed city name and zip code.

Can you reach your customer?



VISUAL DISTINCTION

We use the results flags from the output (indicating the degree of conformity to reference data) to create a visual, filterable distinction on the map.

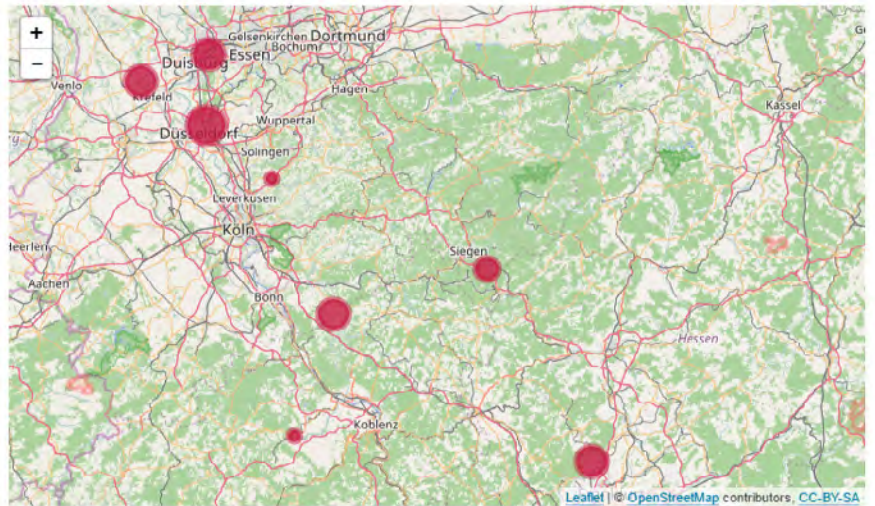
ILLUSTRATION 7: VISUAL, FILTERABLE DISTINCTION ON THE MAP.



FINDING DUPLICATES

The duplicate groups are plotted on the map, indicating amount and geographical distribution of duplicates.

Do you know your customer?



Red dots indicate duplicate groups. The bigger the dot, the bigger the group.

ILLUSTRATION 8: SCANNING DUPLICATES.

The audience is facing another provoking question in the matching section. In this step, Uniserv is scanning the data for potential duplicates. Uniserv gathers the results and assigns them to duplicate groups, based on their frequency. These groups are then plotted on the map, indicating amount and geographical distribution of duplicates (illustration 8). Since most duplicate groups are likely to appear

in urban areas, this chart is very good at pointing to outliers and creating transparency about systematical errors that happened at the point of data entry. It furthermore presents a good opportunity in the data quality audit to start a discussion on data ownership and data-entry processes.

Frequency of duplicates

Uniserv's matching engine is able to execute a pairwise record comparison based on an underlying, configurable ruleset. The outcome of the matching results in so-called cleans and duplicates.



CONCRETE NUMBERS

Plotly is used to display the relative and total frequency of duplicates.

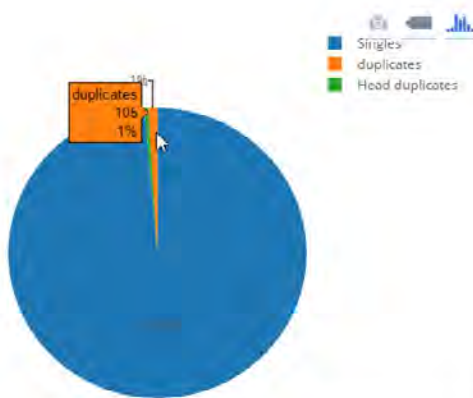


ILLUSTRATION 9: STATISTICAL VISUALIZATION.

While spatial visualizations are a great means for data exploration, other visualization techniques are more suited to provide concrete numbers. Hence, we apply native R-functions on datasets to calculate statistical character-

istics which we render with visualization libraries. For example, plotly is used to display the relative and total frequency of duplicates (illustration 9).

BETTER DATA. BETTER BUSINESS.

Data science has the capacity to influence every aspect of an organization and with Uniserv's data quality visualization technics one can underpin the cause of data owners and thereby provide a literal means to precisely highlight the data quality issues during the data science project initiation.

Using open source for this purpose makes it feasible to adjust the visualization technics such as charts and figures, according to the one's needs. Moreover the vibrant open source community guarantees that the final data quality verdict is delivered using an up to date and state-of-the-art technology.



Tim Kettenacker is working as a data scientist for Uniserv. He is experienced in merging data sources and ensuring consistency in several large-scale projects for customers in the insurance, banking and logistics sectors. Tim uses a mixture of tools to unlock hidden value in messy datasets and communicate data insights.

TIM KETTENACKER

Customer Data Scientist,
Uniserv GmbH

Ovia Raj Kumar is working as a technical data quality consultant for Uniserv. As a technical consultant, she works alongside customers to analyze their data quality issues and guides them along the path to ensure their data-centric processes are running smoothly.



OVIA RAJ KUMAR

Technical Consultant,
Uniserv GmbH

RESOURCES.

1. John Hopcroft and Ravindran Kannan , 'Foundation of Data Science' (04.09.2013).
2. Jeffrey, Stanton, 'An Introduction to Data Science', Syracuse University (2012).
3. Philip Guo, 'Data Science Workflow: Overview and Challenges' (October 30, 2013)', taken on 20.10.2016.
4. Data quality audit, <http://www.uniserv.com/> taken on 21.10.2016.
5. Uniserv GmbH, Data Quality Audit: The status quo of data quality, taken on 21.10.2016.
6. Haug, Anders, Frederik Zachariassen, and Dennis Van Liempd. „The costs of poor data quality.“ *Journal of Industrial Engineering and Management* 4.2 (2011): 168-193.
7. Figueiras, Ana. „How to tell stories using visualization.“ 2014 18th International Conference on Information Visualisation. IEEE, 2014.
8. Lee, Bongshin, et al. „More than Telling a Story: A Closer Look at the Process of Trans-forming Data into Visually Shared Stories.“ *IEEE computer graphics and applications* 35.5 (2015): 84-90.
9. <https://cran.r-project.org/>
10. http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0
11. Wimmer, Hayden, and Loreen M. Powell. „A Comparison of Open Source Tools for Data Science.“ *Journal of Information Systems Applied Research* (2016)



TRUE ↗

↙ FALSE

© 2016 Uniserv GmbH or a Uniserv subsidiary. All rights reserved. The distribution and reproduction of this publication or parts thereof for any purpose and in any form whatsoever are prohibited without the express written permission of Uniserv GmbH.

Ground Truth from Uniserv as well as the associated logos are trademarks or registered trademarks of Uniserv GmbH in Germany and other countries. All other product and service names mentioned are trademarks of the respective companies.

Ground Truth is the foundation for revenue growth and new business models against the background of digital transformation, e.g. for the optimization of sales and marketing activities, campaign management, blacklist matching, Compliance or also Customer Relationship Management. The Golden Profile of each customer is at the heart of this process, i.e. the Golden Record enhanced with interaction and transformation data (transaction data). The Ground Truth means that decision-makers no longer act in a vacuum but have grip on reality for decision-making.

CUSTOMER DATA MANAGEMENT

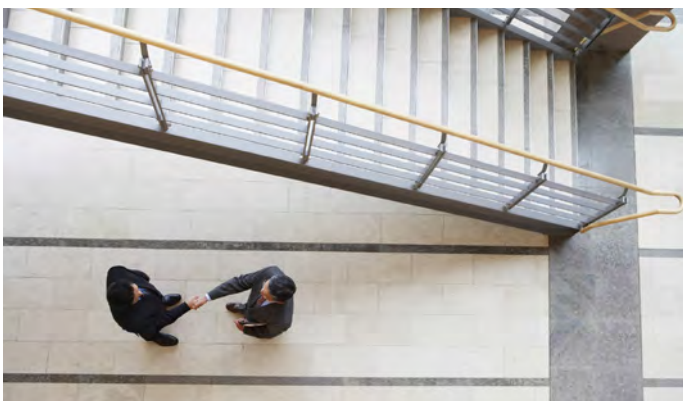
from the European Market Leader

Uniserv is an expert in successful customer data management. Smart Customer MDM, the MDM solution for customer master data, combines data quality assurance and data integration in a comprehensive approach. Customer data is at the focus of initiatives for Master Data Management, data quality, data migration and data warehousing, e.g. in the environment of CRM applications, eBusiness, direct and database marketing, CDI/MDM applications and Business Intelligence.

With several thousand installations worldwide, Uniserv serves the expectations of a comprehensive solution for all business and customer data over the entire data lifecycle. The company employs more than 130 people at its head-

quarters in Pforzheim and its subsidiaries in Paris, France, and Amsterdam, Netherlands, and serves a large number of prestigious international customers from all sectors of industry and commerce, including Allianz, Deutsche Bank, eBay, EDEKA, E.ON, France Telecom, Lufthansa, Otto, Siemens, Time Warner, TUI and VOLKSWAGEN. The Commissioner for Data Protection for Baden-Württemberg recently confirmed that Uniserv structures its business processes in compliance with legal data protection requirements.

You will obtain further information at:
www.uniserv.com



UNISERV GmbH

Rastatter Str. 13, 75179 Pforzheim, Germany

T: +49 7231 936 - 0

F: +49 7231 936 - 3002

E: info@uniserv.com

www.uniserv.com

© Uniserv GmbH, Pforzheim, All rights reserved