Product Report: Triple Boost for Data Consistency and Increased User-Friendliness (part 3 of 3)

Duplicate detection with Deep Qualicision AI

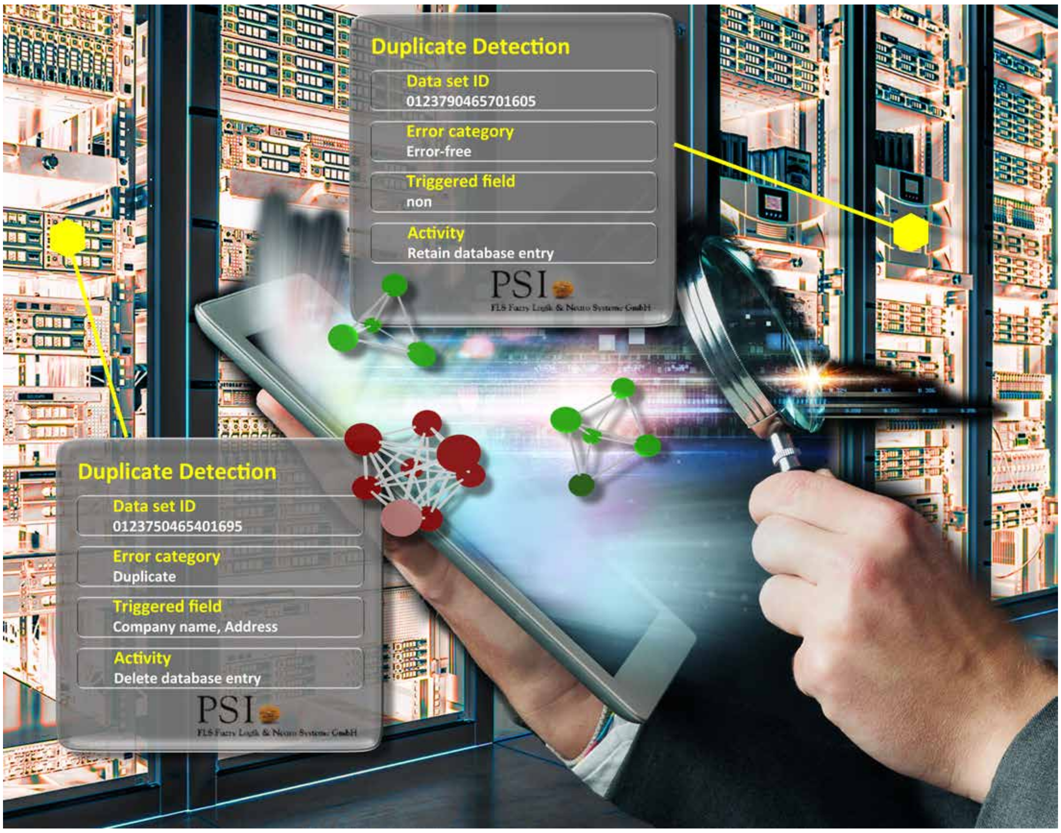
In the last two issues of PRODUCTION manager, two modules for the triple boost for data consistency and increased user-friendliness based on the Deep Qualicision AI Framework have been discussed: The auto-completion at time of data entry and the input validation during transfer to the database. Both lead to a measurable improvement in consistency of data and user-friendliness—individually but also in combination. However, the use of this combination only adds value for data that is newly collected in the process. In a database that has been in existence for many years, remaining duplicates can thus counteract overall consistency. At this point, duplicate detection based on the Deep Qualicision AI Framework comes into account. For this purpose, the syntax and semantics of the datasets learned from historicized data as well as from the data entered and checked during data entry are immediately used to detect duplicates in already existing databases.

In almost any business process today, data is the basis for acting efficient and effective. Maintaining a continuously high level of data quality is a major challenge both for editors and administrators of such databases. In the case of partially unmonitored data collection—for example without the use of auto-completion or automated data input validation—inconsistencies are generated more and more over time. Sometimes it can lead to disruptions in the process itself and those that follow. This often results in manual rework having to be carried out or even the occurrence of planning errors.

A customer use case: Address management of suppliers

For many years, addresses of suppliers operating worldwide were collected in a database. The entries were always made manually and by many different editors. Addresses that were supposedly not found, a new one was created. Over the course of time, duplicates of the same supplier were generated due to different spellings.

One example is a supplier in Italy whose street name can be entered in many different ways: In the local language as “Via delle Fabbriche” or as a German translation in the variants “Fabrikstr.”, “Fabrikstrasse” or “Fabrikstraße.” In addition, the company name can likewise be entered in the local language or as a German translation as well. In this way alone, there are eight possible entries for the same information. Additionally, variants with upper and lower case letters can also be created. Consistency in address management thus steadily decreases over time, which reduces user-friendliness as well as the process itself.



Duplicate Detection

Data set ID
0123790465701605

Error category
Error-free

Triggered field
non

Activity
Retain database entry

Duplicate Detection

Data set ID
0123750465401695

Error category
Duplicate

Triggered field
Company name, Address

Activity
Delete database entry

PSI
FLS Fuzzy Logik & NeuroSysteme GmbH
PSI
FLS Fuzzy Logik & NeuroSysteme GmbH

Duplicate detection with Deep Qualicision AI.

Searching for duplicate data on the basis of similarity metrics

In case of a constantly growing database that has existed for years, a manual search for duplicates to maintain consistency is out of the question due to the amount of time involved. A first approach is to use similarity metrics. For this purpose, the contents of data sets are interpreted as text objects with a sequence of letters and then the distances between them are calculated. If this deviation does not exceed a specified rate, the two checked objects are treated as duplicates. However, this represents a method approach searching for well-defined anomalies. In essence, this is a threshold check for a similarity comparison, which also depends on the length of the word. In addition, such processes have weak runtime behavior for large amounts of data, which limits applicability in the context of Big Data. Moreover, similarity metrics are sometimes unstable regarding semantics when processes change over time. Instead, a mechanism is needed that automatically detects anomalies in the structures of a data set comparison and can continuously adapt to the current conditions.

Data-based duplicate detection using Qualitative Labeling combined with Machine Learning

In most business processes, a broad base of historicized data already exists. Through Qualitative Labeling combined with Machine Mearning based on the Deep Qualicision AI Framework, the structures of an entire database can be learned from past data in a process-specific manner. Data-driven methods offer many advantages, especially for detecting multilevel relationships and complex similarities in data—such as finding a supplier that is listed with several entries in address management.

KPI-based self-learning duplicate detection as part of a Deep Qualicision complete AI system

The fundamentals for duplicate detection based on the Deep Qualicision AI Framework are the combination of Qualitative Labeling with a knowledge base of historicized data trained by Machine Learning. In addition, similarity metrics are used to make comparisons between text objects. However, the framework also enables decision support by simply giving preference to different evaluation KPIs. In this way, not only syntactic similarities, but also semantic analogies—as with different spellings of street names or company names—can be included for detecting duplicates. This kind of KPI-based self-learning inspection mechanism can thus provide an automatic way for continuously detecting duplicate data entries, based on a data history and including a knowledge base that is constantly growing in the process. For the process itself and those that follow, this ensures that planning can be carried out with consistent data to reduce manual rework and avoid errors.

Deep Qualicision-based duplicate detection as an extension of auto-completion and data input validation

A system already in operation with auto-completion and data input validation can be extended in a modular manner using the common Deep Qualicision AI Framework for duplicate detection. This enables a further measurable increase in user-friendliness and data consistency.

Benefits of duplicate detection

- + Detection of duplicates as anomalies in an entire database
- + Automated detection of duplicate data sets
- + Significant time savings and planning reliability in downstream processes
- + Consistency across the entire database
- + Qualitative standardization and plausibility analyses
- + Continuous relearning of the knowledge base to maintain a current data status

A complete AI system with triple boost for data consistency and user-friendliness

The modular linking of the auto-completion, data input validation and duplicate detection components—which can also be operated individually—creates a constantly self-learning and expanding knowledge base for automated support in data collection, verification and storage. This provides the triple boost for date consistency and user-friendliness based on the Deep Qualicision AI Framework.

PSI FLS
Fuzzy Logik & Neuro Systeme GmbH
Dr. Jonas Ostmeyer
Consultant Supply Chain Optimization
ostmeyer@fuzzy.de
www.deepqualicision.ai