

# White Paper

## Successfully designing data quality projects.



**Integrated product suite for the cleansing,  
enhancement, consolidation and optimization  
of databases in batch processing!**

All company and product names  
and logos used in this document  
are trade names and/or registered  
trademarks of the respective com-  
panies.

### The effects of poor data quality

Unsatisfactory data quality adversely affects day-to-day business in a variety of ways. At the strategic level, an unsatisfactory data situation results in incorrect business decisions. At the operative level, incorrect data is often responsible for lower revenue and earnings, increased costs, dissatisfied customers, declining customer retention and a lower return on investment (ROI).

### Data cleansing: An important contribution to improving the information quality.

Data cleansing is only one step in the data quality process but it is an essential one. In this respect, it makes no difference whether it concerns data cleansing in old systems, migration projects, master data, risk management or compliance or the implementation of a customer-oriented approach with mainly operative, department-related data use, e.g. in customer retention, personalised communication or dialogue and direct marketing. Data cleansing creates optimum data quality with respect to up-to-dateness, reliability and degree of detail and therefore lays the foundations for good analysis results and business processes.



## Everything under one roof

The *DQ Batch Suite* represents a fully integrated working environment with all the requisite options for defining, executing and monitoring batch processes. All process definitions as well as the status of the running processes can be directly controlled via the graphical user interface. The time of execution of a batch process can be simply set via the integrated scheduler. Regular, recurring process flows can also be set up without problem. A particular strength of the *DQ Batch Suite* is its suitability for international projects. Customer data-related function blocks such as address conversion, postal validation and *geocoding* are available as expert systems in a large number of country-specific program versions.

Country-specific knowledge bases for consumer and business addresses are available for matching customer data for a large number of countries. The availability of country-specific individual functions is constantly being extended. It goes without saying that the user interface is available in different languages (currently German, English and French). The consistent client/server architecture of the *DQ Batch Suite* ensures that all users have the same view of process definitions and process flows. At the same time, the integrated user management can be used to control which users may open, change or start which process definitions. Effective teamwork is therefore actively supported by the *DQ Batch Suite*.

## Data Quality Batch Suite – Successfully designing data cleansing projects

With the *Data Quality Batch Suite*, Uniserv offers a unique software tool. It combines the simple and comfortable definition of process flows with seamless technical integration in a wide range of system landscapes and provides extra high-performance at the same time. As a result, the *DQ Batch Suite* copes with the diverse and highest demands of business and IT users for the efficient implementation of successful data quality measures.

### DQ Batch Suite: Data quality functions "all-in-one"

Data cleansing processes in the batch are an indispensable component of any initiative for the systematic improvement of the data quality in a company. In the initial cleanup,

- customer and address-related data is validated, corrected and standardized,
- duplicates are detected and either automatically merged directly or made available for subsequent manual processing,
- and missing data such as telephone numbers or geocoordinates can be elicited.
- Any problems with product and financial data which were discovered through data profiling are automatically cleansed if possible.

Regular cleansing activities are also indispensable in the downstream processes if only because the name and domicile of individuals or companies, street names and local government reorganizations, creditworthiness criteria, company information such as size, revenue, etc. change constantly.

The *DQ Batch Suite* integrates all the important and typical function sequences for data cleansing in the batch in a suite. The system supports you in the interactive definition of cleansing processes and in the configuration of the indivi-

dual processing steps in a graphical user interface. Defined process flows can be started and monitored in the graphical user interface.

The *DQ Batch Suite* is the hub of all cleansing projects as a software-supported assistant for the definition and control of batch processes.

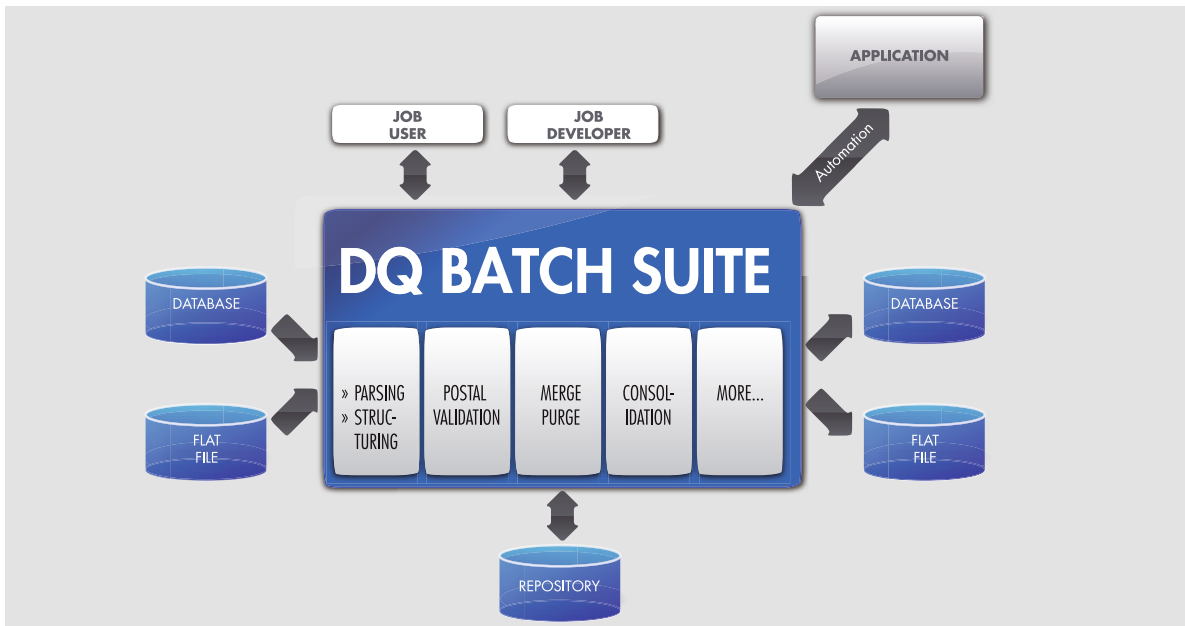
## Advantage: DQ Batch Suite!

In this way, the *DQ Batch Suite* enables cleansing tasks to be completed simply and quickly, thereby saving time. It makes the systematic IT-supported performance of cleansing processes in the batch generally possible. Consequently, the *DQ Batch Suite* enables the user to concentrate on the technical requirements of his task by relieving him of programming and routine jobs.

The *DQ Batch Suite* therefore provides data quality specialists with a powerful tool, by means of which they can define and execute analyses and cleansing operations without recourse to programmers.

In addition to this, the *DQ Batch Suite* is:

- **easy to learn**, because it is consistent in its design and can be operated intuitively,
- **open and flexible**, because it is freely configurable and can be extended by user-own functions,
- **highly scalable** through "state-of-the-art" client/server architecture,
- **user-friendly** as a result of extensive user support and ultra-easy parameterization and
- **high performance** as a result of advanced and reliable operations and algorithms.



The *DQ Batch Suite* is therefore interesting for any company which wishes to implement systematic activities for improving its data quality, data integration or data migration projects or projects in business intelligence or master data management. Data quality processes in the batch can be defined and executed more easily, faster, more flexibly and at a higher quality with the *DQ Batch Suite* - an immeasurable advantage for optimum data quality at low cost with permanent application.

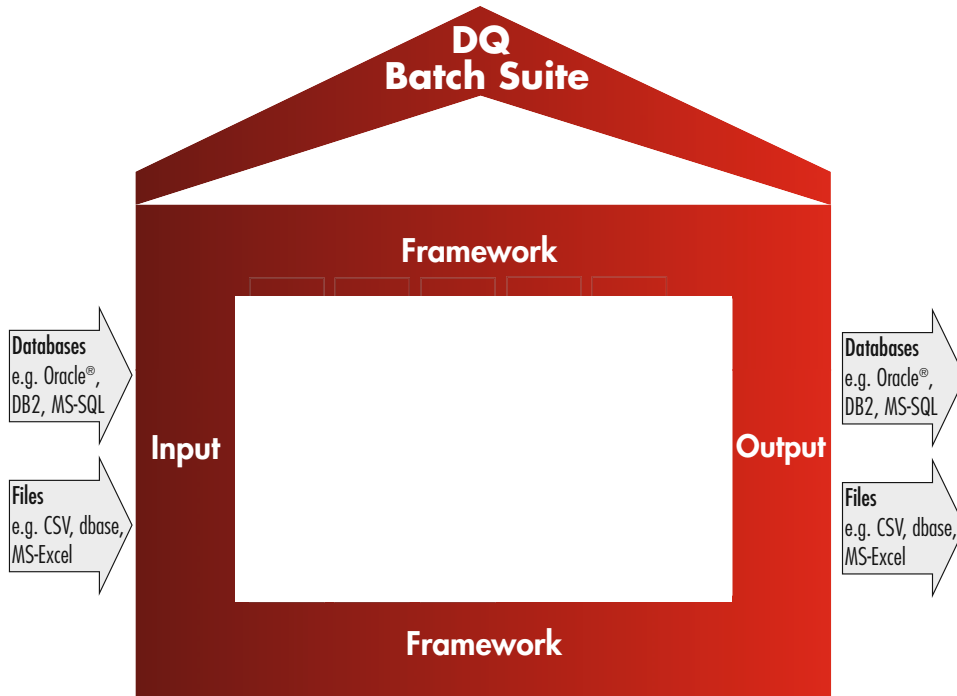
**In this configuration and uniqueness, the *DQ Batch Suite* provides tangible values and supplies directly measurable benefits which pay off immediately:**

- High productivity
- Minimal periods of training
- Comprehensive operator acceptance and satisfaction
- Faster Return on Investment



## Framework

The core component of the *DQ Batch Suite* for executing data cleansing processes is a clearly structured basic framework. This framework is equipped with everything required for the organisation and execution of cleansing processes. It comprises two basic function blocks:



### INPUT

#### Input

This function loads the data to be processed from databases or flat files to the *DQ Batch Suite*. Flat files are automatically analyzed for format, record length, content, field description as well as field beginning and end. In the process, address-related fields, such as first name, surname, title, salutation, street, house number, postcode and town as well as the telephone number and e-mail address, are automatically recognised to a large extent.

### OUTPUT

#### Output

This block at the end of a batch process writes the cleansed data directly to a database or makes it available to downstream steps as flat files.

## Function blocks

Depending on the customer requirements, the **framework** can be extended by additional **function blocks** between these two steps. The internationally tried-and-tested data quality solutions of Uniserv are used here:



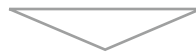
### *convert (Analysis, structuring and formatting of personal and company names)*

This function analyzes name information and makes the identified components available individually. As a result, name elements can be standardized, the gender automatically recognized and the prerequisites created for a personalized address.



### *post (Verification, correction and standardization of addresses)*

This function enables complete and reliable checking and, where required, correction of the postcode, place, street and house number of an address. Complex search and assignment algorithms, similarity analyses and associative approximation methods enable almost complete automatic cleansing. It is thereby guaranteed that hearing, reading, spelling and typing errors, with which the addresses may be stored, are reliably eliminated, and the data is available in a valid, correct and complete form. Permanent changes such as place and street renamings, local government reorganizations or post-code changes are also automatically tracked in the database.



### *geocoding (Enhancement)*

This product option of the post is function used for enhancing addresses with geographic information. This could consist of geographic coordinates, demographic information or codes of suppliers of micromarketing data, which in turn make information on purchasing power, social background and lifestyle available on the basis of these codes.



### *mail (Matching)*

This function finds potential duplicates in a database. Pre-defined standard comparison rules are available for matching customer data. Complex in-house rules for recognizing and evaluating possible identical records can also be defined by the user without any problems. Apart from identifying duplicates, sets of customer data can also be e.g. grouped into households.

This block enables the in-house database to be matched against reference data and enhanced with additional information. In this respect, the error-tolerant matching process generates extremely high assignment rates.

Uniserv provides knowledge bases for the analysis and evaluation of the elements of personal and company addresses for a large number of countries. Apart from customer data, the sophisticated and high-performance error-tolerant processes can also be used for the systematic recognition of similar product and article descriptions.





## **Consolidate (Consolidation of data)**

Data can be condensed in an extremely user-friendly and intuitive manner via a reference number (e.g. customer number) or via the matching of results by means of the consolidation function. As a result, data from enhancing files, such as telephone numbers or target group data, can be linked to the data records of other lists. In addition, sales data can be accumulated from transaction data and used for the subsequent selection.



## **relocation (Relocation check)**

This block prevents a database from becoming obsolete through relocations. For example, more than 20,000 changes of residence take place every day in Germany. The postal address is also changed as a result. If these changes are not tracked in a database, it slowly ages. This has a negative effect on the success of the mailing campaign, because these consignments are either delayed or can no longer be delivered after a few months.



## **phone (Telephone number assignment)**

This block validates the telephone numbers for Germany or assigns them if they are missing. Out-of-date or incorrect telephone numbers can thereby be efficiently corrected and missing telephone numbers added. The telephone data is accessed error-tolerantly. As a result, telephone number can also be assigned if the customer addresses contain hearing, reading, spelling and typing errors or abbreviations. You thereby ensure that your customers can be reached.



## **Script (User-defined functions)**

This function makes it possible to insert almost any desired additional functions and rules in the data quality processes. The use of the scripting language Perl permits the definition of freely selectable complex rules, the analysis and processing of field contents by means of extensive functions for the manipulation of character strings as well as regular hard copy output and the call of Web Services, just to mention a few examples.



## **reporting (Statistics and reports)**

This function makes it possible to define individual reports about the results of an address processing operation and to generate them in various formats (CSV, Excel, XML). The results can thereby be processed in a visually attractive manner and presented.

This current portfolio of functions is constantly being extended by further blocks which can also be integrated user-specifically in the *DQ Batch Suite* framework.

## Optimising, simplifying, automating and accelerating processes

The *DQ Batch Suite* offers extremely comfortable features and properties directly via the framework. These enable users to considerably simplify, greatly automate and permanently accelerate their processes and functional sequences in all areas of data cleansing:

### User-friendly as a result of:

**Context-sensitive help:** A detailed Help Text is available for each input field. The user is thereby provided with complete and extensive support in any situation and at each step.

**Consistent operation:** The *DQ Batch Suite* is characterised by its uniform behavioural reactions, displays and formats. This makes it easier for the user to access the *DQ Batch Suite* and to operate it.

**Error handling:** The *DQ Batch Suite* provides the user with explicit, usable feedback about his errors. The user can retrieve this feedback at any time and is provided with comprehensive support for error correction. Problems can be identified more quickly and solved more directly as a result.

### Flexibility:

The user can considerably influence each individual processing operation and customize it at variance with the standards. For example, small conversions can be carried out or user-defined scripts created using a separate module.

### Extensibility:

The *DQ Batch Suite* can be easily extended by user or third party applications via its open interface. These are executed under the umbrella of *DQ Batch Suite*. User products and applications can be made available under the *DQ Batch Suite* by means of a special development environment for customers and third parties (SDK - Software Development Kit).

## Automation through:

### Scheduler for the automatic start of jobs:

A process chain can be created but the start of processing is set for a later time e.g. with lower computer utilization. The creation and the start of a processing operation are therefore decoupled, resources can be optimally used.

### Automation API (Application Programming Interface):

The user can easily, quickly and reliably automate additional processes covering all aspects of address processing by means of this standard API from Uniserv.

**Changing a processing operation:** The possibility of integrating external tools in the framework of the *DQ Batch Suite* to meet individual user requirements enables the user to comfortably alter, customize and modify the created processing operations.

**Automatic notification:** The *DQ Batch Suite* automatically informs the user by e-mail about the processing of the addresses, e.g. about problems in a processing operation.

## User administration:

The *DQ Batch Suite* offers comprehensive user administration as part of its administration options. As a result, a wide range of users and user groups can be quickly and simply created and defined by the administrator and given the appropriate privileges. Changes can be made at any time.

## Archiving:

The framework of the *DQ Batch Suite* provides extensive facilities for archiving jobs which have been created and/or executed. In this respect, the data to be archived is removed from the productive system, compressed and stored. Valuable resources and capacities are thereby conserved in the productive environment. The archived data can be flexibly reactivated on demand at any time. It is extracted and transferred to the productive system, where it is available once more.

## Multi-client capability:

The *DQ Batch Suite* is fully multi-client capable. This means that e.g. different departments or external partners and customers can be given certain restricted options for accessing the *DQ Batch Suite*, in order to e.g. send address data for processing or to pass on particular project-specific information of processing relevance. The respective clients are strictly isolated from each other.



The *DQ Batch Suite* can be operated in the same way as the usual Windows applications. As is customary, the menu and all the program functions are selected via icons and functional sequences, address processing operations and background programs started according to the specific requirement. The user is also guided step by step and always has extensive options e.g. to stop or modify initiated steps. Tasks can also be carried out in

parallel. External add-on programs, which can be used to extend the *DQ Batch Suite* via its open interface, profit from the product advantages of the *DQ Batch Suite* and appear in the innovative framework under the umbrella of the *DQ Batch Suite* in the same way as all the other functions.

## Further highlights of the *DQ Batch Suite* - An overview of the Framework

In addition to the features and properties mentioned above, the framework of the *DQ Batch Suite* is characterised by other central and practical highlights. These optimize the application of the *DQ Batch Suite* to a significant extent and help to improve the integration of the *DQ Batch Suite* in the business processes on the user side.

- **Selections in each step:** Each process chain (job) within the *DQ Batch Suite* consists of individual function steps. The *DQ Batch Suite* presents the user with extensive options at each step, so that the addresses which are to pass through the respective step can be precisely selected from the total stock. An optimum selective procedure with a perfect result quality is therefore already guaranteed when the job is created.
- **Conversion and script functions in all field assignments:** By this means, the user has the possibility to set and customize each field, so that the processing operation supplies the best possible result. The user can choose between standards or individually created functions here.
- **FileViewer:** The user can analyze his individual address files in detail by means of the FileViewer. Extensive filter functions are available here as well as wide ranging descriptions, e.g. of fields, their length and composition – a practical overview at any point in time. The FileViewer can also be used to check the components of external files and analyze the content of the *DQ Batch Suite* database.
- **Overall statistics:** The *DQ Batch Suite* provides the user with extensive, highly detailed statistics for each individual processing operation of his addresses. In this respect, information is available, e.g. about the executed steps, such as postal validation or duplicate matching, together with the respective result, e.g. which addresses could be reliably processed, where ambiguities were found, etc. etc. In this form, the overall statistics are a valuable resource which enables the user to obtain a compact overview of the most important results of the process chain.
- **Templates:** The user can create so-called templates for recurring complete processing operations or recurring components of processing operations by means of this feature. These can then be called on as required, so that specific repeatedly executed tasks are accelerated. Once created, the user can filter and administer the templates according to certain criteria by means of the template management integrated in the framework.
- **Cut & Paste:** By means of this feature in the framework of the *DQ Batch Suite*, the user can e.g. directly transfer descriptions of records and descriptions of lists in electronic form from MS Excel, MS Word or MS Access to the descriptions of the address processing operations.
- **Installation configuration:** Here the *DQ Batch Suite* can be administered simply, quickly and flexibly on the system side via a "cockpit", and it is also possible to intervene in processes and functional sequences. As a result, installation parameters can not only be administered, the scheduler and its contents can also be checked and the user administration accessed if required. There is also a continuous overview of the status of the processing operations currently running.
- **File formats:** The *DQ Batch Suite* can recognize and read all the standard file formats, such as csv with a header record, dbase or a fixed record layout with a "ragged margin". Unicode is also supported. In addition, databases such as Oracle®, DB2 and MS SQL Server can be accessed directly.

## At the state-of-the-art

In addition to many other things, the *DQ Batch Suite* is mainly characterised by its extraordinary architecture:

- **Consistent use of XML** for the description of jobs, the parameter dialogs and the internal validation rules
- **Systematic use of Unicode** in the internal datakeeping as well as the graphical user interface
- **Use of a scripting language**, a portable library for graphical user interfaces and a *DQ Batch Suite*-internal database especially designed for the concerns of mass processing in the batch provide flexibility, portability and performance

## Seamless integration in your individual environment

The *DQ Batch Suite* servers are systematically implemented in client/server technology and designed so that they can be seamlessly integrated in any IT and application environment. In this respect, the greatest importance was placed on the *DQ Batch Suite* adapting to its environment and its users and not the other way round.

Above all, this simple integration is supported by the openness, flexibility and extensibility of the *DQ Batch Suite*, so that the workbench can grow with the requirements and also offers space for future developments. As a result, the investment in the *DQ Batch Suite* is highly protected.

## Product levels for different requirements

The *DQ Batch Suite* is available to the user in various configuration levels. In this respect, a fundamental distinction can be made between the standalone solution (workstation) and the multistation solution (server). The individual levels are distinguished by the

- supported platforms,
- permitted number of users,
- performance capability, data capacity
- database support and
- range of function.

In general, entry to working with the *DQ Batch Suite* is possible at any level. The *DQ Batch Suite* is designed in such a way that it can be upgraded to the next higher product level at any time. The existing data, settings and parameterizations are transferred directly without loss.

As an integrated direct marketing workbench, the *DQ Batch Suite* is available for Windows and Unix (currently HP-UX, AIX, Sun Solaris, Linux on Intel PC).





## Platform availability and requirements

The minimum technical requirements for use of the *DQ Batch Suite* are as follows:

• **for Workstation:**

256 MB main memory (recommended 512 MB)  
1 GB disk and space for the addresses and address-related work files  
Network card  
Windows XP or newer  
CD drive

• **For server applications:**

**Clients:**

256 colours  
128 MB main memory (recommended 256 MB)  
96 MB disk space  
Network card  
Windows XP or newer

**Server:**

**Entry Server**

256 MB main memory (recommended 512 MB)  
1 GB disk and space for the addresses and address-related work files  
Network card  
Windows 2000, 2003 or newer or Linux (Intel)  
CD drive

**Advanced Server**

256 MB main memory (recommended 1024 MB)  
Between 32 and 96 MB for each parallel job  
1 GB disk and space for the addresses and address-related work files  
Network card  
Windows 2000, 2003 or newer or Linux (Intel), AIX from 5.1, HP-UX from 11.11  
and Solaris (Sparc) from 2.8  
CD drive

**Enterprise Server**

512 MB main memory (recommended 2048 MB)  
Between 32 and 96 MB for each parallel job  
1.5 GB disk and space for the addresses and address-related work files  
Network card  
Windows 2000, 2003 or newer or Linux (Intel), AIX from 5.1, HP-UX from 11.11  
and Solaris (Sparc) from 2.8  
CD drive