Solution Paper

USING DATA EFFICIENTLY

Applications for high-performance, error-tolerant search algorithms and their added value for business performance

TOP
TECHNOLOGY
Made in Germany

exorbyte

Foreword

Data-driven processes need clean data. The right decisions depend on networked data. A successful digital transformation requires a total picture of all the data.

For all areas of an organization, whether in the business world or in public administration, the successful digital transformation of the organization is a key success factor. The term “digital transformation” covers much more than just the introduction of new technologies. Upgrading and stabilizing existing IT systems, databases and data management systems are all major factors in this. Implementing an error-tolerant similarity search makes master data fit for your digital transformation and accessible centrally for use in all business processes.

Doing this, however requires that the master data be available across all systems – regardless of data quality, transliteration or format. An important step on the way to your digital transformation is creating a system that lays the foundation for networking of data and establishing a central master data hub for all business processes.

We make our contribution to your digital transformation with the approximate indexation technology matchmaker. We would be happy to accompany you on the road to your digital future. In this Solution Paper, we describe how matchmaker, the High Performance Matching Engine, works, the range of processes you can automate with it, and the added value it will ultimately mean for your entire organization. We hope that once we introduce you to some typical use cases from large and medium-sized companies and public administration organizations, and how matchmaker accelerates the efficient handling of data for them, you will be ready to start using your data’s potential to advance your own successful digital transformation!

Yours digitally,

Roland Meyer

Managing Director
exorbyte GmbH

Thilo Torkler

Head of Consulting & Sales
exorbyte GmbH

Contents

Foreword 2
Contents 3

1. A new paradigm for the use of master data 4

2. High performance in data applications 5
   2.1 High-speed data searching 5
   2.2 High-quality data matching 5
   2.3 Secure data detecting 6

3. The difference: error tolerance at all levels 6
   3.1 Semantic tolerance 6
   3.2 Phonetic tolerance 6
   3.3 Tolerance in the field mapping 7

4. Added value for total business performance 7

5. 100 % Compliance: Know your Customer (KYC) 9
   5.1 Money-laundering and the financing of terrorism 9
   5.2 PEP and sanctions lists 10
   5.3 Fraud prevention 11

6. Public safety and public services 11
   6.1 Border security 12
   6.2 Dispensation of public services 12
   6.3 Registry modernization 13

7. Getting a grip on data quality 13
   7.1 Duplicate detection and prevention 14
   7.2 Address management 14
   7.3 Data consolidation with multiple system structures 15
   7.4 New customer acquisition based on external data 16

8. Document management 17
   8.1 Multi-Channel Input Management 17
   8.2 Cross-document connections 18

9. 360° customer view 18

10. Product searches for online shops 19

11. Technology in partnership 20

12. The road to data-driven high performance 21

13. exorbyte GmbH 22

14. A peek under the hood 23

1. A new paradigm for the use of master data

High Performance Matching Engine

Data, data, data – these days, it's all about data. It's the resource of the future, and no matter what market you're in it's where you need to be looking for your decisive competitive advantage. But to get that advantage, you have to do a lot more than just collect data. You can only reap the benefits from it if you understand how to use it to create sustainable value.

As data sets become larger and larger, data matching solutions that use filters to iteratively limit the amount of data become increasingly error-prone. You need to be able to use all relevant data efficiently at all times. This requires a paradigm shift – away from the segmentation of the body of data, and towards a perspective on the data that covers it all. The exorbyte solution has been developed for just that: to use all available master data efficiently, without any pre-filtering.

Fig. 1: Performance Parameters

The High Performance Matching Engine is a universal, flexible and transparent indexing technology that has been under continuous development for twenty years. Wherever data needs to be queried, compared or searched, matchmaker can help. For fastest possible networking of the organization's data-hosting systems or external data sources, it relies on back-end interfacing with the individual processes and applications. Regardless of the number of different data sources, formats and structures, matchmaker matches both structured and unstructured queries against structured reference data that it merges into an In-Memory-Index. This means high-speed, high-quality error-tolerant results you can count on no matter how large the data volume is.

The core technology behind matchmaker is a versatile implementation of the Levenshtein algorithm and its various derived implementations, such as Damerau-Levenshtein. Layering with other powerful algorithms for dates, numerical relations, word similarity, sound similarity, semantic references, and countless other relationships gives matchmaker the ability to understand data more deeply than any other system. Exceptional error tolerance means that comprehensive and fuzzy comparisons are always based on all data in the index. With matchmaker's innovative use of standard hardware resources, several hundred million data records can be stored on a single server and retrieved in a fraction of a second.

Implementation of application-specific business rules is made simple by defining the record attributes to be considered for matching and weighting them based on matching relevance specific to the individual case. Matchmaker can perform high-precision identity resolution on any combination of available attributes for individual entity types (persons, organizations, cities, addresses, etc.). Since the data is not restricted at any point, there is no sequential processing of individual pieces of information to define. This means that every instance in the data with similarities to the defined parameters, even the records with errors, will be found. Of course, the exorbyte technology also has robust, scalable and intuitive functionality for multiple error-tolerant data queries accessed across different language areas and writing systems (Arabic, Chinese, Cyrillic, etc.).

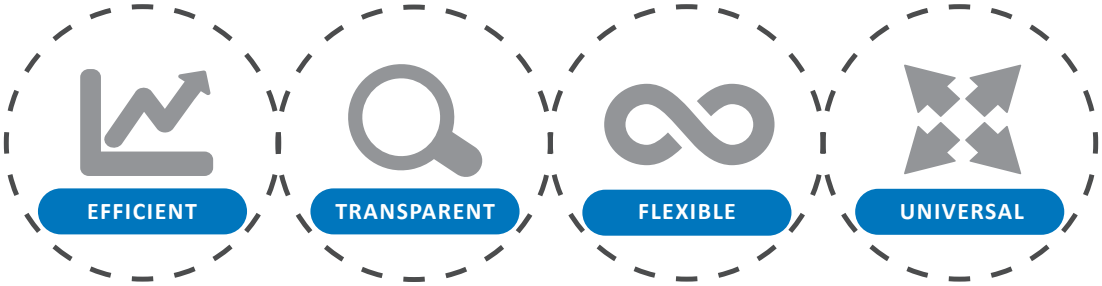

Fig. 2: Principles of the exorbyte technology

2. High performance in data applications

Search, Match, Detect

The traditional queries and business processes that matchmaker makes more robust and faster (thereby adding value) are data searches, data matching and data extraction or a combination of these basic processes applied to a data set.

2.1 High-speed data searching

Master data searches in matchmaker are highly error-tolerant, real-time searches of millions of records that, if desired, can be run from a single search field (as with the classic web search engine) into which the user can enter any desired search terms or data fragments without adhering to any rules or conventions. The user simply enters what is known or suspected from the data record; this could be, for example, a first name in combination with the first two digits of the postcode. Instantly, the user sees a grouped Suggest (custom-configurable) with the most relevant records from the matchmaker index, including cross-field combinations. A categorization of the data automatically generated on the basis of the results (facets) offers additional options for trimming down the list of hits.

2.2 High-quality data matching

Matchmaker uses the same approach as searches without restrictive filtering to match records or fragments of them against reference data. Regardless of how much information is input for matching (a little or a lot), matchmaker compares the input against the complete master data and, with lightning speed, computes the matching records, including similar records, as a single query or in batch mode. Thanks to matchmaker's advanced architecture, whether the data or fragments of the data are of uniform structure (such as when seeking to identify duplicate entries, or for postal address validation) is irrelevant. The results are highly valid and presented in milliseconds.

2.3 Secure data detecting

The master data recognition in matchmaker extracts identities from complete texts or full pages (PDF documents, letters, faxes, etc.) and compares them against reference data. At the same time, matchmaker treats the entire document as a query and breaks it down into substrings. Each individual substring is matched against the entire master data set with the same high-performance matching techniques.

This opens up a number of exciting functionalities, such as efficient and right-first-time automatic routing to the digital inbox, regardless of whether the OCR results or database entries contain errors.

3. The difference: error tolerance at all levels

Seeing the truth

In most situations, data has been collected by various different people, is stored across multiple systems, and is transmitted using various different technologies. Many data records are volatile and some are dynamic by nature. These are only some of the aspects that can influence the validity of the data. The exorbyte technology addresses these issues in a multidimensional way. Error tolerance in string matching is not limited to the various types of formal errors (such as Levenshtein-based error tolerance for typographical errors or data type-specific access methods), but is applied in full in multiple dimensions simultaneously: semantic, phonetic, and field-independent.

3.1 Semantic tolerance

Semantic matching refers to the comparison of data that is different in absolute terms but which has an identity or similarity relation on the level of meaning, i.e., words or strings that are dissimilar in form but mean the same thing or something similar. With matchmaker, these synonymic relationships and aliases (such as “Heidi” to “Adelheid” or “Heinz” to “Heinrich”) are considered equivalent. This allows users to draw on deeper linguistic real-world knowledge. matchmaker’s integrated synonym and alias lists are the result of 20 years of data expertise, and company-specific parameters can be added to them at any time. Within the database analysis framework, further correlations based on the individual data can also be identified as linguistic connections and stored in the system as automatically generated semantic relations.

3.2 Phonetic tolerance

Phonetic matching allows comparison of data that is written differently but has an identity or similarity relation on the level of sound or pronunciation. Phonetic tolerance refers to the capability of recognizing and matching words or names that are pronounced the same or similar but spelled differently. Sounds that are essentially the same when spoken can be expressed very differently in written form across

different languages. This makes phonetic error tolerance extremely useful when running comparisons of international data sources.

The exorbyte team has developed a unique variety of phonetic error tolerance. For unprecedented matching precision, exorbyte uses its own in-house phonetics algorithm alongside conventional phonetics implementation (Cologne Phonetics, Soundex, Metaphone) and transformations for transliteration from other writing systems (Devanagari, Hangul, Katakana, Mandarin, etc.).

Also unique is matchmaker's outstanding performance with its special canonization functionality for Arabic names, which extrapolates common transcription variants. For example, matchmaker treats the letters and sequences of the Latin alphabet "o", "u", "w", "oo" and "ou" as equivalent alternatives to the same Arabic character, "و". "محمد" is usually transcribed as "Muhammad" in the Anglo-American language area, but matchmaker will also match this with the other common variants Mohammed, Mohamed, Muhammad, Muhammed, Muhamet or Mohammad. Common variants of "عمر" are Omar, Oumar, Umer, Umar and Ömer. The exorbyte technology reduces these variants to a normalized format in the Latin alphabet, so that the similarity of that normalized version to the letter-based similarity to a search string can then be computed on the basis of variables. This means that the data can be mapped even for different transliterations from the Arabic, or if different alphabets were used for the notes in the client meeting and for the available data records.

3.3 Tolerance in the field mapping

The unique Flexform technology enables matchmaker to assign inputs and partial inputs on certain fields to other attributes variably. This means that if, for example, first and last names are sometimes reversed, or if the house number is included in the street field, matchmaker will still identify the right hit quickly and reliably.

By combining all these levels of approximation, matchmaker can search all data fields without defining any pre-constraints and without sacrificing any of the outstanding speed of the search algorithms. This makes matchmaker the only software of its kind that guarantees the best search results in the absolute minimum amount of time.

4. Added value for total business performance

Fast, smart, secure

Speed, precision and error tolerance are the hallmarks of matchmaker's performance. The exorbyte technology has a proven track record of accelerating and improving the automation of data-driven processes exponentially, from the very first use. The immediate benefits, like drastically reducing manual post-processing of digitized incoming mail, free up important resources for the tasks in your organization that require specialist expertise. Some companies have earned back their investment in matchmaker within a few months of setting up new matching processes.

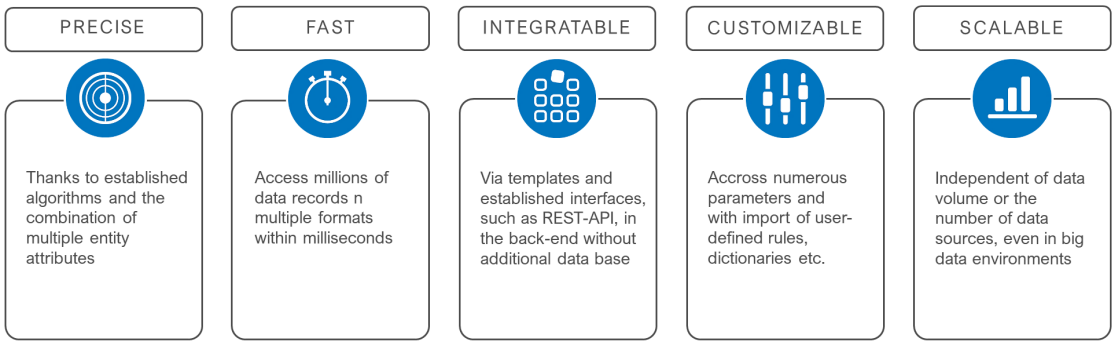

Fig. 3: Advantages of the matchmaker technology platform

matchmaker accelerates (parallelizes) processes, frees up resources, lowers costs, and minimizes risks to support your organization's core business and long-term relationship with your customers.

The time savings and quality increases that matchmaker achieves for the data-based search and matching processes boost your overall business performance.

Sustainably effective added values include:

- Lean IT
  A technology platform handles all data-supported business processes.
- System stability
  Quick integration of the back-end of your existing system landscape.
- Easy data integration
  Interfaces minimize integration costs and facilitate consolidation of data.
- Data availability
  Faster and more secure data access through a central data hub.
- Data quality
  Preventive duplicate identification is a future-oriented way of making the database more robust.
- Data enrichment
  Reference data is automatically used to enrich quality of results to maximize business potential.
- Contact quality
  Automatically standardized and valid addresses are the foundation of customer contact opportunities.
- Compliance security
  Simplifies observance of legal compliance obligations.
- User acceptance
  Uniform, user-friendly search function in all systems.
- Risk reduction
  Minimizes the frequency of errors (bad allocation) that lead to additional expenses and bad decisions.

5. 100 % Compliance: Know your Customer (KYC)

Know who you're dealing with

Do you know your business partners? This is not just an issue of contact management, but a legal duty of care that obliges you to consider economic, social and political security. Money-laundering can happen anywhere and affect anyone – businesses and private individuals alike, anywhere in the world. It is the state's job to apprehend and punish the offenders. All parties active in the global economic ecosystem are called upon to take action to preventively detect money-laundering. Meeting these legal obligations requires all business partners, suppliers, customers and employees involved in a transaction to be assessed against the PEP and sanctions lists. These are the international lists of politically exposed persons, terrorism suspects, and persons and organizations under worldwide or company-specific embargoes. Last but not least, even the most upstanding companies are exposed to a non-negligible risk of general fraud. Secure business relationships on the right side of the law can only be guaranteed with definitive identification of persons and organizations.

Processing heterogeneous and dynamic data using intelligent algorithms

5.1 Money-laundering and the financing of terrorism

Conducting legally valid transactions – that is, legal transactions with assets obtained legally – requires a sound and careful analysis of the parties to the transaction. Anyone who channels illegally generated money into the legal economic circuit, or even opens up avenues to do so as a middleman, whether knowingly or unknowingly, is subject to prosecution under German, European and international law. In implementing these requirements, German law focuses on companies in the financial sector; rules to prevent money-laundering and the financing of terrorism often associated with it are set out principally in the the Money Laundering Act (Geldwäschegesetz, or GWG), which also stipulates responsibilities and processes for diligence and supervision as mandatory organizational measures in this respect. Early detection and reporting of suspicious situations is an essential part of effective risk management in this regard. In the digital age, this is yet another example of where comprehensive linguistic data expertise needs to be coupled with fast and automated networking of all relevant data: publicly available lists of politically exposed persons (PEPs), sanctions lists, internal blacklists, transaction data, etc.

Since money-laundering tends to involve cross-border activity and very complex transaction channels, the risk management for the parties under these obligations is very challenging. To start with, managing the multilingual aspect of the relevant databases proves to be a difficult task. These are the kinds of linguistic issues that must be overcome in order to correlate the information from these relevant databases correctly:

- Uncertainty about correct spelling
- Different transliterations
- Different naming systems in international language areas
- Unclear assignment of identity attributes (titles, first name, last name, etc.)

Then there are the technical and procedural issues, such as harmonizing differing data structures and formats or simultaneous processing of multiple data sources. Likewise, implicit in a successful early detection system is the establishment of an evaluation procedure for all new and existing business partners, to be performed at regular intervals.

With matchmaker you can run your queries against all entries in reference databases with a high level of error tolerance, so you can be certain whether your business partners actually exist or not. You can rest assured that cross-linguistic data reconciliation is also being performed with high accuracy and speed automatically, digitally, and in the back-end of your systems, and that matching is being done across different writing systems, like English and Arabic. You know that matchmaker is giving you the same high level of error tolerance with Chinese, Cyrillic, or any other writing system.

5.2 PEP and sanctions lists

Business relationships with politically exposed persons (PEPs), persons who hold or have held a high-ranking national or international political office, their family members, and other persons within the close personal circle of high-ranking politicians, are considered particularly risky. Because of their influence, these individuals are more likely than the general population to be involved in corruption and money-laundering. The Money Laundering Act requires that parties to a transaction must be checked for PEP status. The PEP and sanctions list check is intended to resolve data quality-related issues, in particular:

- Variabilität
  Sanctions lists are constantly changing.
- Internationality
  There are many different country-specific directories, and they do not always use the same writing system.
- Scope
  Most sanctions list provide only sketchy information on the listed persons / organizations.
- Dynamics
  Data is, by its nature, very dynamic.
- Data structure
  Sanctions lists are found in various formats.

This means that checking the potential transaction partner once before signing the contract is not enough. Only a permanent review of customer data, vendor data and partner data carried out upon the first contact with the parties and repeated on an ongoing basis, can offer security. Having said that, it is also important to not exclude potential partners and transactions (and thus profit) with incorrect classifications. This means that whether you find a result or not, you have to be able to trust the conclusions you draw on that basis.

matchmaker offers that trust now and in the future. Recognize PEPs easily, and entirely automatically, even with different spellings, languages and transliterations: in real time (check during data entry), in the inbox (scanning) or as a regular check of all customer data (batch version). You identify unwanted transaction partners before you invest resources in this relationship and before you are exposed to risk. Because the indexing technology also works across fields, there is no need for labor-intensive reformatting and structural consolidation of your distributed data before reconciliation. All data is read out in full, in its existing structure, and compared against all relevant blacklists – first when it is created, and thereafter at regular intervals.

5.3 Fraud prevention

According to international credit insurer Atradius' "Payment Practices Barometer 2019 – Germany", payment default by German enterprises tripled in the year under review. One cause that can be suspected for this trend is targeted or organized fraud. What may signify an existential threat to an enterprise can also cause enormous losses (application fraud) at the state level.

The challenge: be able to recognize and detect multiple, ever-changing fraud tactics. Successful fraud prevention demands clear compliance guidelines and, in its implementation, requires efficient real-time verification processes and intelligent technologies that reveal even the most sophisticated tactics. Finally, an intelligent identity check flags orders or applications with dubious identity information (false personal details or address, stolen credit card or account data, bad credit history, etc.). Any payment defaults can be evaluated with the help of recognized credit agencies or on the basis of internal blacklists on which payment defaulters or other "fraudsters" are recorded. One way fraudsters attempt to circumvent techniques that would identify them as this type of "bad actor" is by trying to obtain goods or services again using similar, but not the same, information. This is why matching of internal and external data demands both the highest possible precision (few or no mismatches), but at the same time appropriate error tolerance. Relying solely on name matching is in effect an open invitation for fraudsters. But effective fraud detection needs to go even further, by matching persons who at the name level would not appear critical at first glance, but who are actually connected with each other in a less obvious way. This requires the inclusion of additional entity attributes. A suspicion of fraud can also be derived from similarities with others, such as: multiple similar insurance claims for services from the same service provider; asylum applications registered with the same telephone number; several rebate applications from what is in practical terms the same address.

With matchmaker you can quickly and accurately match entities against reference data and embed high-level fraud detection in the early stages of your business processes, because matchmaker shows you not just the similarities between personal names, but also recognizes other common attributes such as addresses, company and personal telephone numbers, etc. as relationships. This means you can block even complex attempts at fraud with confidence and in real time, instead of in a downstream process. And this process can be carried out fully automatically in parallel with data acquisition.

6. Public safety and public services

Protect and support

Public safety and public services are tasks of the state; these are intended on the one hand to serve the welfare of the community, while at the same time are intended to benefit the welfare of the individual, as in (for example) the services provided directly by the state for a certain set of entitled persons in society. Rendering these services in an equitable way requires that public safety authorities and public service providers have reliable instruments for verifying identity. They need to be able to verify the identity of a person in a meaningful way, and know without a doubt who exactly they are dealing with. The demands of globalization constitute an ideal profile for a good testing system: mastery of international writing systems coupled with comprehensive data expertise.

6.1 Border security

While it is clear that full transparency on all cross-border activities would make a significant contribution to general security, investigative authorities still face a plethora of problems when searching for personal data across various systems, and not only at border checkpoints. All too often, this means a multi-stage, expensive process for border officials, which even then does not lead to the optimal results.

Obviously, in this realm of security a large percentage of the names to be checked will by their nature be international, and some of them transliterated from non-Latin alphabets. Consequently, name searches for border security purposes need a highly precise name matching process that can handle a large number of variations in names, in an error-tolerant way and in the shortest possible amount of time. With millions of travellers passing through airports, seaports and other points of entry every single day, border checks need to happen as fast as possible, but also as thoroughly as possible. Alongside name, other attributes will generally need to be checked against watch lists: date of birth, place of birth, nationality, organisation, address, etc. Running checks on all this data is an extremely time-consuming process for security authorities, because they must be run against heterogeneous data structures, and the entity attributes in the various possible combinations produce further check criteria.

With matchmaker you can run a name check against several hundred million data records with a response time of less than a second. Powerful search algorithms are the key to clean real-time screening. The exceptional speed of the algorithms means that you can also retrieve "fuzzy" matches. This means that you can be sure your search result will also show all possible similarities, even higher-order similarities and names (orthographic, phonetic, semantic), as well as other entity attributes. Attributes can also be weighted to reflect the most appropriate procedures and responses for border security.

6.2 Dispensation of public services

Public services like state benefits are generally linked to criteria tied to the entitlement to the benefit. Such entitlements may be subject to periodic review to assess the status of the entitlement at regular intervals. In the end, benefits come from money contributed and dispensed on the basis of the principles of solidarity and community. Unemployment offices and state welfare agencies, for example, must establish for any applicant whether the person is entitled to benefits, which benefits those are, and for how long. This means that the service provider must also make decisions on benefit fraud, for example, if an applicant for benefits is already receiving the same benefit from another institution. Reviewing this is not made any easier by the fact that different institutions within public administration use different IT systems. In addition, all systems are prone to data entry errors, sometimes due to lack of knowledge of the correct spelling, sometimes technical data transmission issues, and sometimes simple carelessness.

Making the right decision with certainty

Differences in the spelling of names, addresses or other personal identification characteristics are one reason that multiple benefits may end up going to the same person. It is not at all uncommon for duplicate entries or irregularities in address notation

to compromise the quality of data used in benefit processes in a country. To ensure that benefits are granted fairly and equitably, the institutions charged with these processes need to be able to rely on a high quality of identity resolution, and high efficiency (prompt case handling) in that processing. For this, intelligent technologies must provide the solution; manual verification of the identity features across the various systems and registers is simply too cost-intensive (the costs in terms of manpower alone be prohibitive).

But matchmaker can deliver the high-validity identity resolution these institutions need. Enter all the unstructured data you know about an individual into a single search field, and matchmaker will return highly error-tolerant results from all relevant reference databases. One of matchmaker's most powerful features is its phonetic error tolerance across multiple writing systems. Using the same algorithms that are used to reliably find individuals with foreign names, this feature allows you to generate clusters of potential duplicate entries of persons. It visualizes and interprets additional data such as family connections to identify duplicates and trigger preventive measures when new duplicate data entries are created.

6.3 Registry modernization

In an expert report on the modernization of local and national registries, the German Federal Government's National Regulatory Control Council has noted that "Modern registries are the foundation of better administrative services for individuals and enterprises." Germany's registry landscape is very diverse, with over 200 registries throughout the country, and the national government bases policy decisions on them. This is what makes a registry search a multi-step and complex process. Multiple data records across registries are also an issue. Accordingly, the quality and usability of registry data are generally assessed as poor.

With matchmaker, registry data searches can be carried out in no time, regardless of the structure and quality of the data. Because of matchmaker's back-end indexing, all systems can be integrated easily and their data bundled instantaneously. All desired registers can be queried by entering all available information into a single search field. The error-tolerant matching retrieves language-independent obvious and non-obvious connections between data records, and identifies them clearly with a highlighting system based on individual weightings of the entity attributes.

7. Getting a grip on data quality

Matching, Cleansing, Enrichment

When it comes to data-driven control of business processes and solid data-driven decision-making, the quality of the data is everything. In many cases, "bad data" starts at the point of data entry, and is reinforced as it occurs across multiple systems. A company in today's data-driven, digital economy has to be able to identify and eliminate process-related weaknesses in master data maintenance while guaranteeing reliable data quality in all source systems (including deduplication, correction, enrichment, classification and consolidation of master data).

7.1 Duplicate detection and prevention

One of the most important benefits of consolidation and harmonization of master data is enhancing the efficiency of business processes and how they run. This means that you should approach any data cleansing project from the perspective of business processes. With an independent organization and smart software tools, you can set the stage for successful consolidation and harmonization of multiple data sources.

Whenever one or more of the quality characteristics critical for smooth business operations in the company are absent, quality problems need to be addressed. The causes of these problems will generally be multifaceted and complex.

One thing that excellent process flows have in common is that they maintain an ongoing process of data cleanup; after a deduplication process, the quality of the data remains guaranteed. As this implies, the ultimate goal is the prevention of duplicates for redundancy-free data. Today's data acquisition systems (like CRM packages) have solutions for avoiding duplicates when creating new data records, but these are limited to predefined parameters. If a new record differs from an existing record because of a different spelling, for example, this kind of built-in duplicate recognition fails to perform its function.

matchmaker unerringly identifies all the duplicates that have crept in over time and been propagated across distributed systems, and feeds them into a dialog application for cleanup. Cross-field variable field mappings and field-specific error-tolerance parameters reveal the typically hidden similarities between records, like misspellings, field swaps, or different formats and structures, so that these records can be clearly identified as duplicates. Multidimensional error tolerance means that matchmaker keeps your data clean over time. The Matching Engine keeps checking the database of all indexed source systems, recognizes potential duplicates, and alerts you to them. With matchmaker, data inconsistency is a thing of the past.

7.2 Address management

Bad addresses cost your business in many ways, both directly and indirectly, like when you send multiple letters to the same address or when your invoices take extra time to get there. Errors in address information arise from problems in address checks – or not having an address check at all. Directly, this increases your costs of direct marketing in real, quantifiable terms. Indirectly, bad address data ultimately damages your image. You look unprofessional when you approach potential customers and address them by the wrong name.

These errors are typically the result of data entry problems: street name and address number entered in the same field, street name outdated, incomplete or simply incorrect, numbers of postcode mixed up or otherwise incorrect, address noncompliant with postal service requirements or otherwise considered invalid. A company can accrue a surprising amount of costs as a result of this type of error. Avoiding these errors has a twofold impact, as it also saves you the process of address data cleansing.

Address data is generally very dynamic: according to a 2018 study by the German postal service, over eight million people move in Germany every year. This means not only mountains of returned mail (with all the postage and processing costs this entails), but also losing countless customer contacts

each year. In a very real way, valid address data is one of the foundations of doing business. With that in mind, change of address reconciliation should be embedded as an ongoing, automated process.

With matchmaker, you and your data are always up-to-date. It cleans up your existing address data on a fundamental level, running automated comparisons against valid reference data to correct street and city names, generate missing postal codes, replace outdated street names, structure addresses correctly and register changes of address and deceased persons. You can also use completion mode to help you speed up your address data entry. With exorbyte's fuzzy algorithms, an address fragment is enough for matchmaker to reconstruct the matching address and possible alternatives. And the completion assistant can also be integrated into online systems for use by your end customers. This increases user acceptance, reduces abandonment rates, helps eliminate fake addresses and improves the quality of the address data. Plus, an additional result that your customers will appreciate is getting your mailings and your message.

7.3 Data consolidation with multiple system structures

For any company, keeping your database accurate and up-to-date is an ongoing job. In a number of industries, like the insurance industry, the heterogeneous system landscape has emerged as a solution to the problem of structurally very different sales channels and sometimes hyper-specialized products, but this solution makes data consolidation difficult. Data structures can be diverse, and the data itself may only be available from a remote source – but do these factors have to be obstacles to future-oriented, data-driven value creation? A migration project to unify them will generally be a sweeping change process with far-reaching implications, the ultimate effects of which cannot always be clearly assessed in advance. In our experience, these processes are very time-consuming and cost-intensive, with a major factor being incalculable follow-up costs. Despite this, many IT system consolidation processes inherently have the goal of consolidating data. Maintenance and upkeep of multiple systems can tie down considerable resources. Central issues here for data quality managers and systems administrators are: How to create a shared data landscape? What systems need to be continued, consolidated, or replaced?

How can non-standard data structures be harmonized? What impact will a system redesign have on the ongoing operations?

To illustrate a few potential approaches for merging different data landscapes, we can take an example of two hypothetical companies A and B, and for each of them show the way that they benefit from the exorbyte solution:

- A and B each retain their own system landscape.
  In this situation, the diversity of the systems requires continuous data synchronization, automatic data cleansing across all systems and networking of all data for a 360° customer view.

- A and B system landscapes are merged into a new system.
  The data structures must be assessed and consolidated in a process of matching, cleansing, and ideally automatic correction of incorrect entries, and then the clean data must be transferred to a new system.

- System landscape A is migrated to B (or vice versa).
  The data structures must be converted into a uniform data format. Next, matching, cleansing and enrichment of database A with additional information from B to ensure a clean and secure merge.

With matchmaker, data quality and data availability can be ensured even in a decentralized and heterogeneous system landscape. Established legacy systems can be upgraded for the digital future with matchmaker. The platform collects all the relevant data that an agent or automated process needs for the everyday operations, across all systems. Comprehensive search parameters and multidimensional error tolerance guarantee that matchmaker will find everything virtually instantly. You know that if matchmaker doesn't find it, it's simply not there. And matchmaker applies the same principles of precise matching, automatic validation, correction and cleansing of the database and enrichment of the data records with information collected from all systems when doing the same in reverse, transferring locally saved data to a central system. Any additional systems that are added later as a result of acquisition or internal restructuring can be incorporated just as easily into a shared consolidated system landscape or a network of distributed systems.

7.4 New customer acquisition based on external data

For customer acquisition, many companies frequently pay for access to public records collected by commercial address database providers. If you do this, it is highly recommended that you run a precise comparison of the information in the database against the data available in the company's own database in order to minimize your costs of data purchasing, avoid unnecessary data quality issues and prevent scatter loss in your new marketing campaign. After all, if your goal is to attract new customers, it doesn't make sense to approach existing customers or "lost contacts". Merging an external database with the customer data available in the company generally proves to be a time-consuming process. Large volumes of data make the matching process difficult, and different data structures have to be harmonized in multi-step processes. After that, further, selective comparisons based on target group-specific criteria (clustering) require additional investments in time and human resources.

matchmaker compares external databases against your internal data with error tolerance, independently of language and the individual field structures of the databases. Using weightings for individual attributes, you can hone in on the precise data with the real business potential for you. A very useful parameter for telecommunications or public utilities, for example, is identifying households and family groups with a common attribute like a telephone number. This allows them to define the relevant target audience very precisely, which ultimately saves on marketing costs. The error-tolerant matching also delivers significantly higher quality in the results at significantly shorter throughput times. All this enables you to permanently reduce your costs of new customer acquisition and unlock the doors to further savings in resources by process automation.

8. Document management

Automatic recognition and processing

While the paperless office has not yet truly become a reality, today's business processes are based in large part on electronic document management. And because rapid availability of process-relevant data depends on the digital documentation of business transactions, electronic document management will be the key to the further automation of data-driven standard processes.

8.1 Multi-Channel Input Management

Today's businesses provide customers with quite a range of communication channels: the classic postal delivery, fax (which, as much as one may scoff, is still used every day in some fields), telephone, e-mail, and now social media channels. What all of them have in common is that they deliver information that needs to be processed in a professional way. One way of achieving uniform data storage is to convert incoming mail into digital documents. Today it is normal for incoming documents to be scanned, classified, the relevant information extracted manually and, finally, exported as documents for the appropriate processing.

Full-featured document viewers or integrated viewer tools give users a uniform view of multiple document types and facilitate business processes with note functions and fast navigation features. But information relevant to business processes will not be available, except as metadata. Risk elements like PEPs will not generally be automatically identified in the incoming mail process, or may have changed status by the time the document is processed.

matchmaker lets you automatically validate documents digitized by OCR technology and electronically transmitted data against reference data to reliably recognize, extract and check the required person or transaction information according to company-specific requirements – even with scattered/unstructured and incorrect (factual errors or OCR errors) information. This means you can automatically process over 90% of documents and significantly reduce your throughput time. The advanced matchmaker algorithms have been proven to deliver valid results even for documents that cannot be indexed with conventional methods. They enable matchmaker to automatically correct OCR errors, so that you can save valid data and perform the necessary matching processes (i.e. checks against sanctions or PEP lists) and route the documents optimally – all in one step. Upon first opening the document, matchmaker presents you with any alerts about hits on blacklists or other regulations that you may need to know.

8.2 Cross-document connections

In the insurance industry, it is normal to have multiple contracts, each for a different type of insurance, but all with the same customer. As a rule, a customer who is happy with the provider of one insurance will trust his provider with all the other insurance policies he needs. This means that assigning customers to contracts and each contract to the responsible agent (in the case of multiple agents) must always be accurate. A complicating factor in the insurance industry is that any claim will generally involve multiple parties – beneficiary of the insurance, at-fault party, witnesses, and others. This is why the identification of cross-document relationships in customer data and linking of related documents can simplify and accelerate the work process tremendously.

matchmaker enables you to extract the facts from the documents and compare them against all available master data and documents. Your agents receive a hit list with links to other documents with the same case number, other matching attributes or other relevant information such as the other current contracts of the customer or family members, or other completed or currently open cases. Also, matchmaker adds an additional level of qualification to the results with hit probabilities, so the agent can take all relevant data and connections into account when processing transactions.

9. 360° customer view

Networked knowledge

Fragmented, incomplete or unknown information makes it hard to get a total picture of the customer. In handling a customer inquiry, a customer service employee will often go through a long process chain and have to take many detours in the process, searching various systems and collecting various information fragments. With many applications, the company's search functionality is affected by variations in spelling of the names of individuals or companies that are the result of typos, different transliterations, nicknames, or any number of other discrepancies. The costs involved in this process are completely disproportionate to the output, search time is unproductive time, the search result only delivers part of the truth, and the risk of bad decisions based on the unreliable results is simply too high. The cumbersome nature of the process is only aggravated in telephone customer service situations, when it is either interpreted as incompetence on the part of the customer service employee or is the source of customer dissatisfaction (or both).

The foundation for good and consistent growth is permanent good relations between the customer and the company. To make a customer-oriented offer, you need to know who the customer is and what their expectations are. A high level of professionalism in customer service helps you retain customers in the long term. High quality of data and fast availability of data are the keys to competent results-oriented dialog with the customer. And last but not least, a 360° view of all the data in an organization is essential for well-founded decision-making.

From a plus point to a requirement: with the coming into effect of the EU's General Data Protection Regulation, every enterprise will want to have a fast search mechanism that can retrieve all data on any individual.

According to Article 15 of the Regulation, consumers have the right to require custodians of their data (controllers) to promptly provide confirmation as to whether the controller is processing their personal data. If the answer is yes, the company must also provide the details concerning the data being processed.

> ART. 15 EU-DSGVO
>
> (1) The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed […]

Fig. 4: Right of access

matchmaker gives you high-level, error-tolerant, interactive access to all your data with a search function that can be easily integrated into all applications as a single or multi-field search. For GDPR-related queries, matchmaker makes it easy to build a central search index that combines data from company-wide sources and systems. In less then a second, you can find out whether a person has any records in your systems – and if so, which ones. This allows you to give the requester an initial response immediately, and to process the request quickly without time-consuming individual searches in each one of your systems. This simple and fast search also pays off for all other customer support tasks, because it gives you a total view of all your customer data at your fingertips whenever you need to provide individual customer support.

10. Product searches for online shops

Taking your customer by the hand online

Far from being just buzzwords, “Usability” and “User Experience” have become key benchmarks for successful websites. Visitors to online retailers need to be able to find their way around quickly and collect the focused information they are looking for. With the growing importance of the platform economy, the number of online items and, ultimately, the total volume of data, is growing exponentially. It has already reached the point where a good search function cannot sort through this mass of data without the help of intelligent algorithms. Sometimes the customer can’t remember the name of the product or brand exactly; sometimes the customer is looking for specific suggestions on a broad spectrum of themes; sometimes, we simply enter an incorrect search term in our haste to find the product we want. No matter what, as users of the website we expect it to understand our input correctly. We might give it a second chance, but if that doesn’t produce the result we are looking for we are going to be highly inclined to leave the site very quickly.

matchmaker is also the ideal search engine for online shops, because it gives both online retailers and online customers what they want. Even with hundreds of millions of products, incorrect entries and users from different language backgrounds, matchmaker applies its algorithms and error tolerance to deliver the right hits in a flash. For the online retailer, this delivers quantifiable economic benefits: modern algorithms correct erroneous user input, find search-relevant suggestions and present an individual (marketing-driven) ranking of search results in a simple or grouped hit list (Suggest). With custom configuration, the search engine can serve sales promotion functions to demonstrably boost the central key performance indicators of your e-commerce platform: Usability, conversion, sales.

11. Technology in partnership

Combining technologies and enriching value creation

The digital transformation of an enterprise brings processes together, transforms standard processes into automated workflows, and builds technological bridges between sub-processes for smooth, seamless value creation. As a universal and flexible indexing platform that bundles data across systems and makes it available company-wide, matchmaker is a valuable addition to a wide range of automation technologies, particularly where the collection and processing of master data is needed. Using REST and native interfaces, the platform can also be readily integrated with other platforms for process-specific workflows, for a result that benefits all parties, enabling users to accelerate their added value (integrated processes) and save resources that would otherwise require an individual interface with the other platform, and for the technology partners opening up new customer markets in the form of integrated process solutions.

In particular, companies with intensive contact with the end customer through multiple channels (insurance companies, banks, energy utilities) have much to gain from these integrated process solutions. The clear advantages for fast digitization coupled with simultaneous data networking of the information in the file systems also make these partnerships ideal when digitizing archives of paper files to merge historical data into current systems.

In these platforms, exorbyte enriches the digital value chain in partnership with solution providers at the international level:

- Scanning technology
  Automatic comparison, validation and classification of digitized / scanned documents

- OCR solutions
  Validation of imported scanned documents including attachments (unstructured data) against reference databases

- Document viewer
  Enrichment of documents with relevant data and connections (links to documents, processes, persons)

12. The road to data-driven high performance

Customized solutions for individual requirements

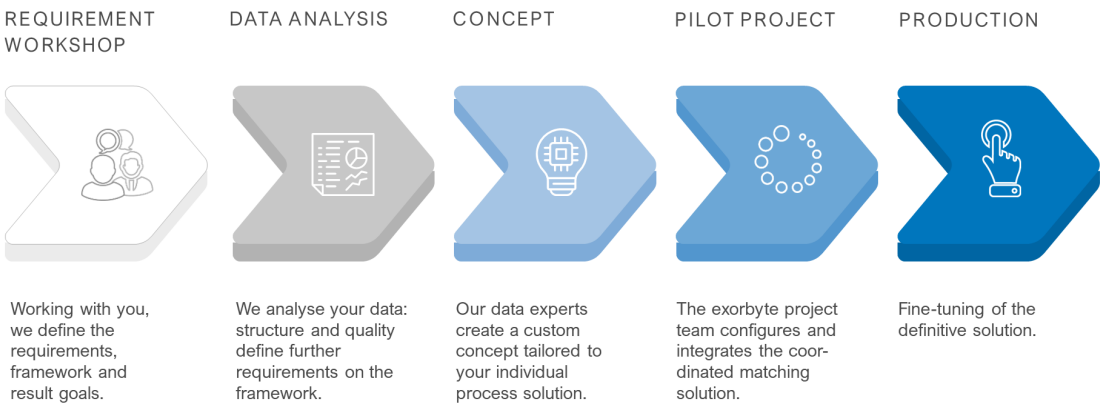

Fig. 5: Phases of the matchmaker implementation

exorbyte offers you more than just exceptionally powerful software. We are data specialists with an interdisciplinary team of software experts, linguists and project managers to advise you on how to solve the particular challenges you face. We see ourselves as a development partner for the implementation of your data strategy, and we walk you through your digital transformation and beyond. In our experience, the full potential of your data-driven processes can be unlocked in five phases (Fig. 5).

Because data is a complex subject and the basic conditions and requirements of each company are unique, the optimization of data management for efficient availability of all master data can vary greatly from company to company. The exorbyte data specialists have extensive experience with search, matching and recognition in data-driven projects. Let us give you a personal presentation to convince you of our comprehensive expertise in data.

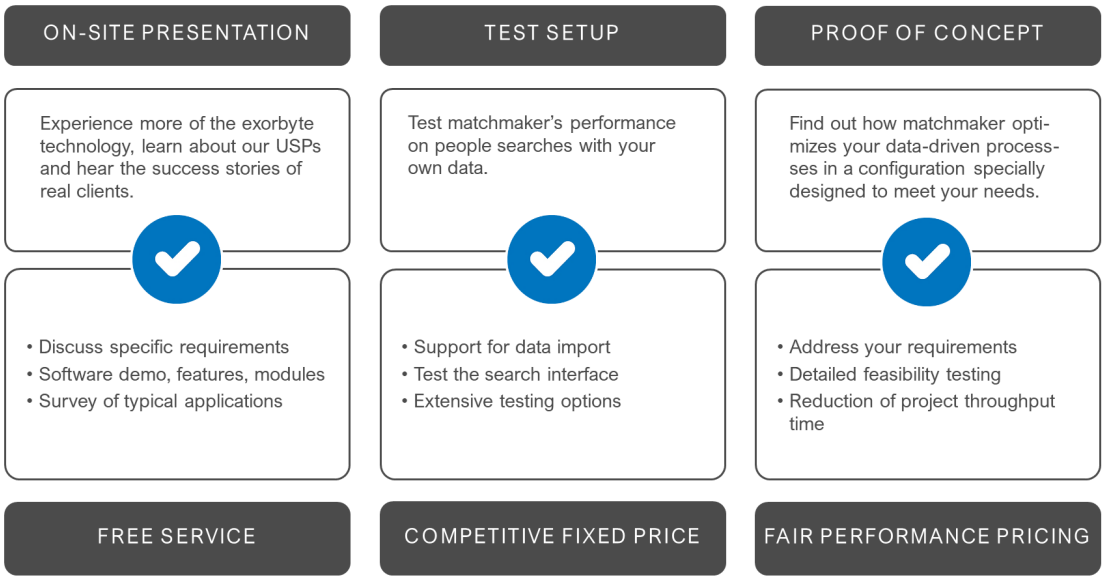

Fig. 6: Opportunities to experience matchmaker

13. exorbyte GmbH

Your partner for empowering your data

exorbyte GmbH, with its offices on the shores of Lake Constance, Germany, has developed matchmaker, the High Performance Matching Engine, to leverage the value of your master data. Its clean and scalable client-server architecture delivers flexible, versatile, high-availability solutions that give you an total view of your master data across all business processes.

The unique combination of speed and error tolerance means matchmaker can seamlessly connect heterogeneous system landscapes – independent of font, language, format and data quality. Our USP is the special ability to search master data without culling, restricting or prefiltering (also known as match prospecting, blocking or pruning), making matchmaker a true approximate indexing technology that ensures reliable, direct and fast identification of all similar entities. This opens up the possibilities for numerous use cases between data availability, data quality and data input for large and medium-sized companies.

Insurance companies, public utilities, governmental authorities and public-sector service providers, financial institutions, telecommunication service providers, commercial organizations and online platforms are all relying on exorbyte's customized solutions for handling master data. Let's talk about your data-intensive processes.

Contacts

Roland Meyer
Managing Director
+49 7531 3633900
roland.meyer@exorbyte.com

Thilo Torkler
Head of Consulting & Sales
+49 176 13633906
thilo.torkler@exorbyte.com

14. A peek under the hood

Phonetics and transliteration

One of the most important tools for approximative search is phonetic coding, which is usually used in the form of alternative search keys. This is not error tolerance, but just another key, for example:

- Stemming
  The word stem is used for the search.

- N-Gram techniques
  Groups of two to five letters are used as search keys, or as keys composed of the first letters of several fields.

The disadvantage of these techniques is that they are for the most part an equivalence relationship. That means two words are "similar" if they have the same code, and they are "not similar" if that is not the case. This runs into problems at the edges of the equivalence classes: if a word is just outside the class, you won't find it, and if it is just barely within the class, the hit will be just as "good" as any other hit anywhere else within the class.

matchmaker also uses phonetic codes – but only as a gradual support of word similarity. This also means that the code itself is found and evaluated with error tolerance, while the overall quality is determined by a voting procedure. This is done with both conventional coding methods and coding methods specially developed by exorbyte.

The difference with real audio or phonetic transcriptions is that a phonetic code generally simplifies the string, in other words, abstracts it to make it less susceptible to error. In phonetic transcription, the string is specialized by adding more information. This makes it even harder to find good matches.

Soundex is the best known, but still a very simplistic phonetic coding. It uses only the first letter and the next three consonants (which are only divided into six different groups). For example, "Müller" becomes "M460". Only „M" and two consonants are considered (after deduplication). The "0" is used to fill out the code. This code is not a good approximate code, because it always involves a significant amount of reduction. The inaccuracy of this model is made even clearer by the following example: The soundex code for the German last names "Meier", "Mayer", "Mair" and "Meyr" is the same: M600. However, this is also the soundex code for: Mr., Meer, Mara, Marr, Muro, Mirra, and many other German last names. On the other hand, "Meyers", "Mejer", "Majer" and "Mejerl" are not coded in the same way. And there are many other examples of permutations with problematic coding: "Tapes" and "Tieves" have the same code, while "Dapes" and "Dives" do not!

Metaphone is a length-dependent code that does not recognize vowels and is widely used in multilingual applications. It does a better job at matching some variants of the surname "Meier", but is also very short and contains vowels only in special combinations with adjacent consonants. Fricatives with "h" are preserved, something that soundex lacks.

The Cologne Phonetics (Kölner Phonetik) algorithm is similar, but optimized for use with German. It also disregards the vowels.

The main phonetics algorithm used in matchmaker is the proprietary phonetics algorithm exoPhone, which is specially adapted to the way matchmaker processes phonetics, i.e. it is less generalized and consequently more definitive for "fuzzy" phonetics hits or those requiring error tolerance. This combination makes the phonetic search robust, and assigns errors that have no phonetic impact less weight than phonetically relevant errors. It also considers vowel groups. In this coding, "Müller" becomes NOLEL.

With server-side scripting, additional phonetic encodings (including self-programmed ones), can be linked, and this can be done without a negative impact on the runtime of the query thanks to clever integration into the matchmaker algorithms.

A unique transformation (which is not a real phonetic coding, but behaves very similarly) is the coding of the word shape. This proprietary exorbyte technology is selectable as SHAPE transformation in matchmaker. What phonetics is to the spoken word, shape is to the written word. For example, in an OCR application, the shape of the characters provides a rough encoding of the characters – representing only ascenders and descenders, strokes or curves to approximate the shape of a word without knowing the exact meaning (which would involve a much longer coding). Here, “Budapest” would be encoded as “Errotoyooot”.

Thanks to its modular architecture, additional transliterations are easy to add to matchmaker. The exorbyte technology also supports a number of phonetic algorithms and additional functionality for phonetic matching of transliteration variants from Arabic.

Standard transliterations included in matchmaker:

- Devanagari (ITRANS)
- Hangul (RR)
- Katakana (Hepburn)
- Hiragana (Hepburn)
- Mandarin (Pinyin)
- Cantonese (Pinyin)
- Arabic (ALA-LC)
- Cyrillic (ALA-LC)
- Greek (UN/ELOT)
- Hebrew (UNGEGN)

Publisher
exorbyte GmbH
Turmstraße 5
78467 Konstanz, Germany

Version
March 2020

Printing
winz.druck, Rielasingen, Germany

Design
Media Lab GmbH, Konstanz, Germany

Image credits
istock/Ani Ka (title page)