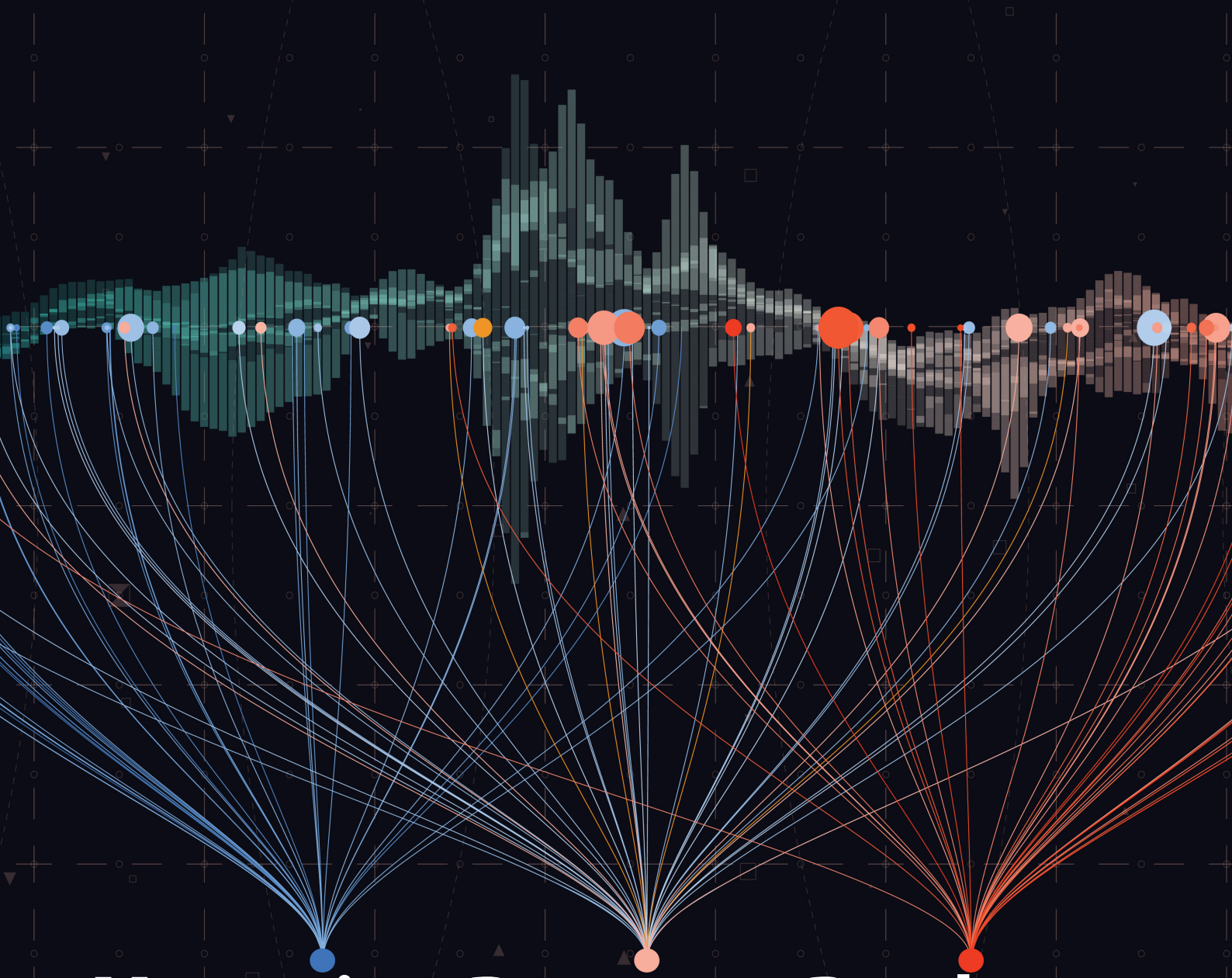


Managing Data as a Product with Distributed Reference & Master Data

Authors
Dr. Heiner Oberkampf, OSTHUS GmbH
Dr. Christian Senger, Bayer AG
Malcolm Chisholm, Data Millenium

Executive Summary

With the digitalization process, data has become the center of both traditional and research-intensive industries. Information needs from inside and outside organizations are increasing in depth and breadth. This means that only those organizations that are able to produce high-quality data products will succeed. These products should satisfy the information needs timely and at a reasonable cost.

In this whitepaper we argue that large organizations need to embrace distributed approaches to reference and master data management, as a means of collaboratively aligning their vocabulary without a forced top-down standardization procedure.

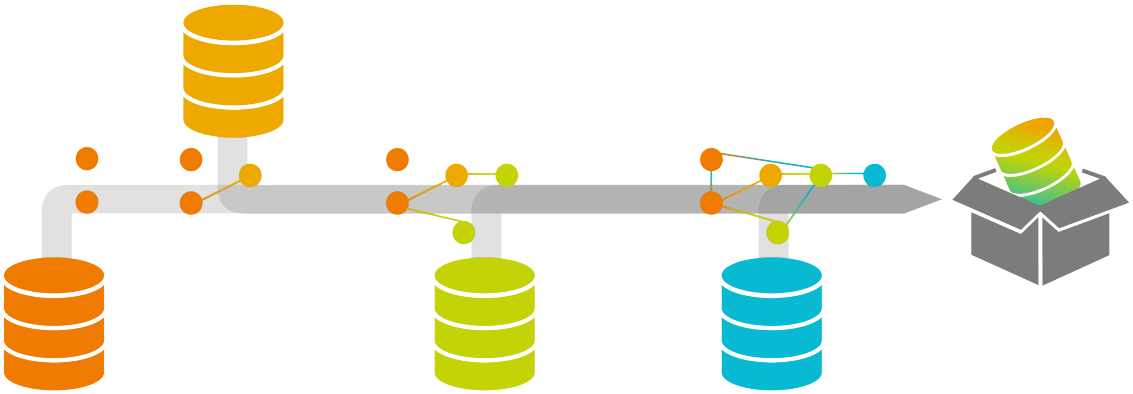
Introduction

Digitalization essentially means that the role of data is changing in traditional and research-intensive industries such as pharma and life-sciences. From being a supplementary part of physical products, data is now a product itself grounded on data innovation. Many such data products are difficult to predict much in advance, though they can be required at any time. While pharmaceutical companies are great at producing sophisticated and high-quality pharmaceutical products, producing high-quality data products at scale has been recognized to be challenging. As we will show in this whitepaper, one key reason for this is lack of shared reference and master data (RMD) and corresponding high integration and cleansing efforts. Two aspects are to be considered at this stage are:

1. Since business value is realized through data-driven or data-informed decision making or insight generation, accuracy and quality of a data product are critical. High-value data products are typically assembled from different internal and external data sources for specific data consumers to serve different use cases.

2. To minimize costs, the creation of high-quality data products needs to be (semi-)automated as much as possible. This is however only possible with good reference and master data which allows to integrate data automatically based on common or mapped terms and identifiers.

The classical top-down reference and master data management (RMDM) approach with one central monolithic system is not able to address the dynamically changing data and application needs. In a digital enterprise, data innovation occurs in different unforeseen groups across the organization and to larger parts from external contributors. This results in a (guided) collaborative network of loosely coupled interconnectivity. From this perspective only distributed management approaches are suitable to overcome this challenge. Furthermore, master data management is no longer only an enterprise internal topic – a scalable model needs to address cross enterprise challenges and needs to consider public sources from i.e. governance, authorities, academia or open source communities.



In this whitepaper, we describe the motivation and requirements for a modern, distributed approach to RMDM, which creates an eco-system or platform for participation and innovation across organizational boundaries. When managed properly, RMD can drive business models because they enable different groups to collaborate more effectively on many different use-cases – contributing to a growing data and knowledge eco-system.

Roadblocks on the Path to Data-as-a-Product

To produce data products at scale, especially RMD needs to be managed as an asset since it has the biggest impact on stronger data integration capabilities required for current and future business goals. At least we must be able to:

1. Find all data related to core RMD entities such as products, compounds, clinical studies, indications, proteins etc.
2. Enable more people in the organization to utilize data effectively in their business processes
3. Answer questions about core business objects even when the data is managed in distributed sources

The following issues with RMD we observe to prevent organizations today to realize value:

1. Lack of common RMD and data standards leading to inconsistent models and definitions
2. Centrally managed RMD does not have the required enterprise-scale adoption rate
3. Missing unique global identification of RMD
4. Huge redundancies in RMD due to lack of coordination
5. Unnecessary complexity of managing and governing RMD

This leads us to the following requirements for a new distributed approach to RMDM:

1. One place to find, view and access RMD from distributed sources
2. Persistent and global unique identification of all relevant RMD (FAIR principle F1 [2])
3. Allow to manage RMD in distributed manner
4. Ability to align data models gradually
5. Aligned light-weight data governance for RMD
6. Scalable de-duplication and mapping approaches to resolve overlaps from different sources quickly

Background

What is Reference and Master Data?

In order to successfully manage RMD it is important to clearly understand what they are. In the past it was thought that all data is the same, and a single set of management practices applied universally to data. Today, we understand that, there are different kinds of data, each with its own special characteristics, challenges, management practices, relevant technologies and needs. Figure 1 summarizes this understanding.

Metadata	Information that helps an enterprise to understand and manage its data assets. As an example, this data might include authorship, creation data, and ownership, but also as much as important, sources and evidence levels of data. Metadata also includes data models.
Reference Data	Data that is used to categorize other data found in a database, or for relating data in a database to information beyond the boundaries of the enterprise. Typical examples in pharma includes locations, types of organism, indications, but also assay types and alike.
Master Data	Data about the things (core business entities) that participate in the transactions of the enterprise. In pharma, we will find the core assets as master data, such as materials, devices, healthcare provider but also products.
Event Data	Data about transactions – meaning interactions between Master Data entities where attributes change and relationships are reordered. This is, e.g., financial data, but also data collected in experiments and studies.

Figure 1: Different major classes of data.

Let us have a look more closely at Reference Data and Master Data.

Reference Data

is somewhat difficult to define but is very easy to recognize. Traditionally it is implemented as tables that consist of a code column and a description (or name) column, and very few (if any) additional columns. It often goes by the names of "code tables", "lookup tables", "reference lists" and "static data". Reference Data lists represent Reference Data concepts, like Country, Product Line, Organism, Method, Pharmacokinetic Effect, and so on. Reference Data lists usually have 200 or fewer rows and these short lists change infrequently. There are in life-science however also significantly larger Reference Data taxonomies or thesauri such as the NCBI Taxonomy with more than 650,000 entities.[1]
Typically, Reference Data is rich regarding lexical synonyms and sometimes also translations. Reference Data concepts do not have to be implemented as relational database tables, and many other options are available, so it be better to speak of Reference Data entities rather than Reference Data records or columns.

Master Data

is also easy to recognize. Master Data is traditionally implemented as tables that usually have very many columns, sometimes thousands. Such tables must be present and populated before any transactions can be processed. There are usually very few Master Data entities in an enterprise, but they are well-known, e.g. Customer, Patient, Product, Provider. Characteristic problems of Master Data are managing identities, deduplication, managing change and timely creation of entities without impeding business processes. Again, while Master Data has traditionally been implemented as relational tables, nowadays many other options are available.

Today, there is general recognition that Reference Data and Master Data need to be managed well because of its enterprise-wide impact on data quality. This has given rise to the disciplines of Reference Data Management (RDM) and Master Data Management (MDM), which we will consider together as Reference and Master Data Management (RMDM). In our experience this is a major consideration for a successful digital transformation and data governance.

1. https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=statistics&uncultured=hide&unspecified=hide

Distributed Reference Data in Life Science Industries

RMD which are required in life science industries is created in many different functions and places. This is not only inside one enterprise along the translational value chain beginning in Early Research, Development, clinical studies, regulatory processes, production, marketing, and finally Real World Evidence observation. Reference data in early research is usually fed by public literature and data created by scientific institutions. Partnering entities such as Contract Research Organizations (CROs) are involved in early research as well as clinical studies. Clinical study and regulatory affairs data are legally required to be based on reference data from authorities such as the EMA and FDA. The foundational workflows to finally come to a life science product involve many different systems collecting data and reusing reference data.

The functions fulfilling specific tasks are often oblivious of using Reference Master Data. Depending on the function, creation of the same or similar reference data points (such as a new indication) might happen in a strictly regulated environment as well as environments completely free to choose IDs and labels for data.

Thus, it is necessary for the particular function to recognize it is dealing with shared and distributed reference data (i.e., Reference Master Data) and open up for sharing labels and codes with other systems or ingesting and reuse codes from other functions or even external organizations.

Although dealing with the same kind of data – such as regions and countries, species, indications and drugs – the data is used in different contexts and often require different levels of detail or granularity. Thus, there are different perspectives on the data and the data entity itself has different roles (requiring different attributes and metadata) in different functions and workflows. Different workflows for creation of data or the conditions of use must be aligned between the different reference data producers or users along the whole data value chain.

Looking deeper into the kind of data and context, it is a complex task to recognize that data entities regarded as atomic from one perspective (e.g., a product in marketing), from another point of view is a compound model containing shared and distributed reference data parts (e.g., substances in production) which in turn, might even be regarded more complex somewhere along the value chain (e.g., agents with salts or isomers, active ingredients and excipients). Thus, it is also necessary to consider the reference data model and decompose it to their atomic reference data if required.

However, it is absolutely vital for Life Science Industries to break their data siloes between the different functions to minimize costs and effort at the interfaces where data are transferred between functions and to satisfy regulatory authorities demand which may involve multiple functions in the product's value chain.

Typical Data Challenges in Life Sciences

Missing Awareness

Internal and external functions are not aware of existing reference data and reference data standards. Alignment is necessary in order to enable integration and translation (mapping) between the sources of Reference Master Data, if multiple standards are required. This lack of awareness leads to reference data being reinvented, unintentionally leading to the creation of data silos. In the worst case, functions are aware of other systems, but claim to provide the reigning reference data while ignoring existing data or standards. Examples of this are the different functions inside the enterprise, such as R&D, production, Regulatory Affairs, Real World Evidence, as well as the work with CROs.

External Authorities

Authorities require particular detailed information from functions, though the latter is often unaware that this information is not already available to the former, due to a lack of sharing and aligning reference data. Thus, such demands can cause massive efforts for data wrangling and integration along the value chain. Diversity of regulatory requirements in different regions and product categories represent additional complexity here.

Different Perspectives

It is challenging to understand the different levels of granularity and perspectives of different functions, as well as the need for decomposing entities and their models in order to get the relevant reference data entities.

Workflow Integration

As many different functions have the need to adapt, extend, and enhance data, it is crucial to align and maintain reference data in a way that enables the subsequent orchestration of data distribution and sharing workflows. This must happen without impeding processes in other functions that also require the reference data.

Use-Case Summary & Learnings

As illustrated by the use-cases core master data entities such as product or study are managed in a distributed manner, and this is not going to change. This is so, because people work in their business applications when they capture or process data. Switching to another system or requesting the creation of a new master entity through a managed service is often not feasible for business users. Often no such system is available to them, or they do not have access to it.

By today, many enterprise data strategies aim to centralize **RMD management** to regain control. Though there are legitimate reasons and interests to centralize, we argue that centralized management alone is too slow to accommodate the fast changing IT and data landscape or avoiding the issues of of merger and acquisitions. It also fails to support the different perspectives of different business units and corresponding incompatibilities. Both aspects are increasingly important in the era of digitalization and increased need for collaboration across enterprise boundaries.

Motivation for Distributed RMDM

As illustrated by the use-cases, we acknowledge that important reference master data is managed in different systems by different people - **this is what we call distributed RMDM.** Before we describe how the solution to the related challenges, we give a more detailed description of distributed RMDM and why it is needed.

What is distributed RMDM?

Based on our experience in working with large organizations in different industries we see that the biggest challenge lies in (missing) standardization and corresponding high mapping efforts. On the one hand this is related to organizational change and consensus building and on the other hand to data modelling. By having data in a distributed manner (or the wish to have that), **data models are of increasing importance.** Distributed RMDM is not only about the same (and unique) ID for integration and translation, it also can be strongly enforced with the ability to translate, integrate, and share data models – in optimal case with explicit context. This is comparable to not only translating words from one language to the other, but also taking care of the grammars.

Distributed RMDM can be characterized as follows:

- Management takes place in many systems instead of one central system
- There is an actual overlap in domains and entities, e.g., product information is maintained in many systems

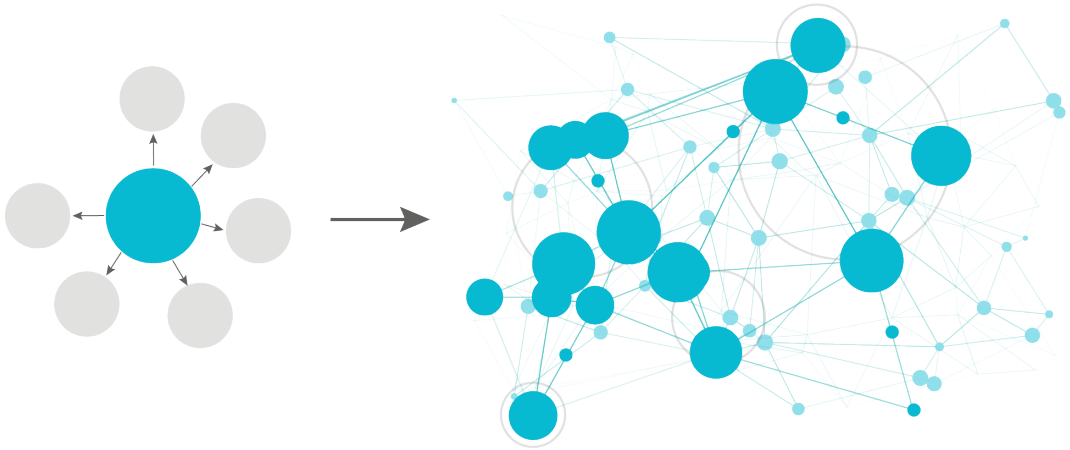


Figure 2: In a central approach all data is created in on place and then distributed to other systems. In a distributed environment, any authorized source can contribute to a common reference and master data eco-system.

The Need for Distributed Reference Data Management

There are several strong drivers for distributed Reference Data Management.

Reference Data Creation Requires Unique Processes and Data

While each Reference Data entity usually requires just a code and description, the creation process of a new entity may be very complex. For instance, establishing a new Product Line entity may be an initiative that originates in Corporate Strategy, requires a number of sign-offs from various departments, requires acceptance of several studies and reports, and finally, needs board approval. Tracking the progress of such an initiative involves managing a great deal of information. This information will be unique to the initiative, and relatively complex. But at the end of it, all that is generated is just one new Product Line entity with a code and a description.

Different Organizational Units Create Different Reference Data

Following on from the example just discussed, it is easy to understand that the organizational units involved in the creation of new Reference Data will be different on a Reference Data concept-by-concept basis. Strategy may lead the creation of a new Product Line, while Marketing leads the creation of a new Customer Segment. This means that many different organizational units are involved in creating Reference Data entities, and that in reality it is distributed. It is far better to embrace distribution rather than fight it by assuming that the structural similarity of Reference Data must mean it all be managed centrally.

External Authoritative Sources

External Reference Data is Reference Data that is produced outside the enterprise, such as FDA Structured Product Labeling standards. The enterprise does not create these, but sources them from the external authority. Again, this means that in reality there is a distributed Reference Data architecture that the enterprise must deal with. The need to comply with external authoritative sources will increase as can be seen by e.g. the FDA's new pharmaceutical quality initiative on Knowledge-aided assessment & structured applications (KASA) [4] or the EMA's driven ISO-standard on the Identification of Medicinal Products (IDMP)[2].

Governance of Reference Data

We have discussed diverse processes, information, organizational units, and external sources in the creation of Reference Data. In addition, there may be roles and responsibilities that are unique for specific Reference Data concepts. Very often, these are related to the requirement for expert knowledge about the content of Reference Data. For example, an 'Intent of Use' concept may have entities for "Treatment", "Prophylaxis", and "Diagnostic", and each of these terms will require an expert to provide a definition, and, perhaps, an editor to standardize the definition and make it easily consumable. A different Reference Data concept may require quite different roles. Such variable governance considerations make centralization difficult and provide a sound basis for adopting a distributed approach.

So far, we have discussed the creation, or production, of Reference Data Values, but what about their distribution? It is here that the idea of a central hub storing all Reference Data for use in an enterprise seems to have the greatest value. But does it? Effectively, we are duplicating copies of the Reference Data from the environments where it has been created into another environment. This requires additional processing, which takes time and means that Reference Data that has been created may not be immediately ready for consumption. There is also risk in that we have a single point of failure for Reference Data in our architecture. The question arises, therefore, as to whether such a central hub is the best option for distribution. Surely, obtaining Reference Data directly from the distributed sources where it is produced is a better option. It may be argued that we lack the technology to do this effectively, but that is a practical consideration which should not prevent us from recognizing that the distributed architecture is the best option.

Business Value

If we accept that data is managed in distributed sources we can concentrate on shaping how corresponding efforts can be aligned and governed. The positive outcomes of this distributed and shared management manifest mostly as

- **First time right:** Reused data leads to higher quality from the source[3]
- **Avoid redundancies:** Each entity and attribute is managed once. This means that distributed/shared governance leads to shared workload (less redundant efforts).
- **Manage by the right expert:** Multiple experts maintain data assets - it is not necessary that one person, with only a bit expertise on everything, maintains everything.
- **Manage at the right place:** Data assets are created and maintained at the point where it makes most sense, e.g., within the right workflow. Different functions have different input systems to maintain there RMDM. Without accepting distribution, people have to switch systems in their work for maintaining RMD – with distributed management one enables the use of different systems for RMD maintenance.
- **High availability:** With a truly distributed system there is no single point of failure.
- **Lower integration costs:** When different sources share a common set of RMD it is easier to integrate data from them
- **Higher rate of innovation:** With the data ready to be used to build understanding about how to best exploit data collected in clinical studies, select the best drug candidates or predict toxicology
- **Faster time to market:** assembling the data for regulatory submissions is one of many processes where shared RMD will increase quality and speedup processes. This is especially important in the context of patient-tailored medication and corresponding more frequent product releases.

2. https://www.ema.europa.eu/en/human-regulatory/overview/data-medicines-iso-idmp-standards-overview
3. https://hbr.org/2020/02/to-improve-data-quality-start-at-the-source

Framework for Distributed RMDM

The following describes the key functional components of a distributed RMDM architecture:

1. Distributed authorized sources of RMD
2. Central registry (technical precondition)
3. Central lookup & resolution (technical precondition)
4. Model repository including standards, alignments and extendibility
5. Operating model with roles, accountability and a new mind-set

High-Level Functional Architecture for Distributed Management of RMD

We have made the case for the need for RMD to be distributed. However, what would a distributed RMDM environment look like?

In order to answer that question, we need to consider what elements are necessary for distributed Reference Data Management to work. Figure 3 illustrates that some degree of centralization, or federation, is needed – not for management, but to facilitate governance and accessibility. Any business unit that wishes to become an authorized source for some Reference or Master Data concept can do so by approval of the governance body. Any business user can discover authorized sources for a given a RMD concept by looking it up in the Registry.

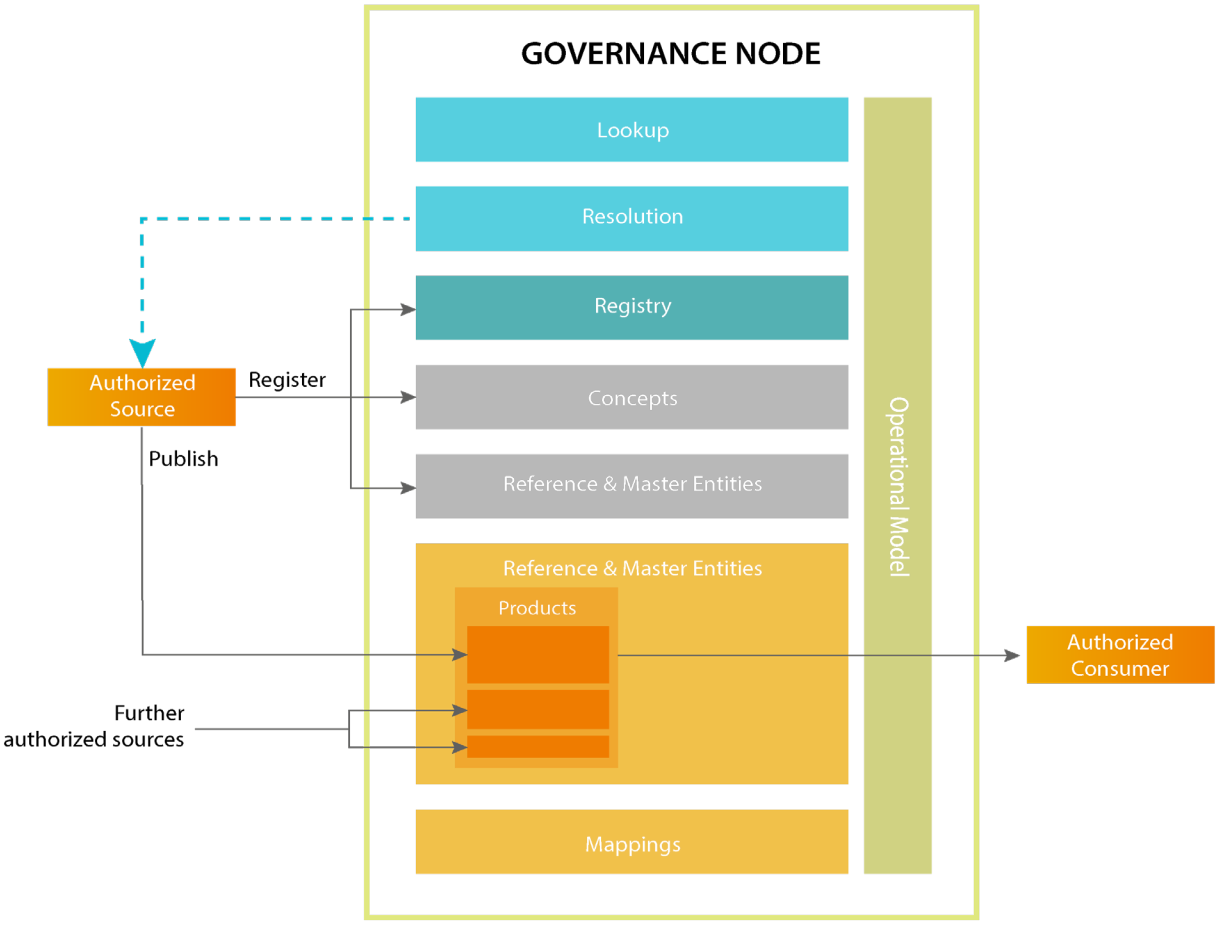


Figure 3: High-level functional architecture for distributed Reference and Master Data Management with the focus on the Governance Node.

From this we can see how distributed management and governance could work. It will require strong standardization and a central catalog on the governance structure and meta-data level.

Some degree of central governance is therefore required to ensure the standards and registry exist and are functioning well. Governance is critical for any (data) platform as described in [1]. The Governance Node can be also the place for community-driven FAIR maturity assessment as described in [3].

Lookup

Potentially, any RMD entity could be included in any data product that is to be assembled. Therefore, we need an enterprise-wide solution that permits anyone in the enterprise in every business unit to **discover** all registered concepts and then be able to obtain the relevant entities within them. This means that we need to have a central **LOOKUP** service for users to search and find information based on any used term or identifier. The lookup service needs to support state-of-the art search and filter capabilities as well as contextualized ranking to enable users to actually find entities or detect gaps.

For Master Data, the Lookup serves also as an entry to a 360 view on corresponding entities, e.g., products since some information may be only referenced and available at the sources (see section on Resolution).

Registry

From a technical side, having a **REGISTRY** for any reference and master data entity can be considered as a technical pre-condition for distributed RMDM. Registrations happens on source, concept, model and entity level.
An **AUTHORIZED SOURCE** is a system or organization, which is registered on the governance level for one or more data domains and corresponding specific concepts. After registration, the source can register and publish corresponding RMD entities.
In alignment with the FAIR [2] guiding principles, all data and metadata are assigned a globally unique and persistent identifiers by the Registry.

Resolution

In a distributed environment, it may be the case, that not all information is loaded to the Governance Node from an authorized source. The **RESOLUTION** serves as an API for all information loaded into the Governance node and also to allow to redirect or assemble information in an ad-hoc manner from authorized sources. Again, a standard will be needed in order to access RMD from the points at which it is being maintained. There are many architectural options possible for this, but a single interface and resolution standard will greatly help all users of RMD to easily get to it.
These global unique persistent identifiers created by the Registry are made resolvable by the Resolution component. A direct resolution mechanism for particular, known elements can be provided, e.g., through target URLs.

Concepts

To easy findability and facilitate integration, we need to carefully define the types or **CONCEPTS** under which authorized sources will publish RMD entities. Since agreement even on concept level may be difficult, we only require that concepts are well defined and linked to each other, at least partially as a well formed taxonomy with strong subclass relationships. For stronger alignment concepts may be captured in an ontology.
A further element is that everyone in the enterprise must have a clear understanding of **who is working with what RMD concept,** in order to avoid the duplication of effort and have competing implementations for the same concept. This means that only **AUTHORIZED SOURCES** are allowed to publish RMD. For distributed management, any group in the enterprise can choose to become the authority for any concept.
An **AUTHORIZED CONSUMER** is a system or organization which is registered at the governance level and which can subscribe to different concepts to view and retrieve updates of corresponding entities. Registration of Authorized Consumers is important to govern the impact of (breaking) changes to the data models or registered entities by providing very simple data linage.

Data Model Repository and Data Modelling Requirements

Data models are at the heart of data integration. Getting the approach to data modelling right has huge potential for the digital enterprise and the ability to create data products. An enterprise-wide solution will be difficult to use if every Reference Data concept is implemented with a completely different structure to every other Reference or Master Data concept. It will be even more difficult to use if every concept has to undergo source data analysis by any team that needs it. This implies there must be a mechanism for standardization of RMD implementation, in terms of structure.

Whatever technical structure is chosen, whether it be relational databases, graph databases or something else, the underlying logical structures should be standardized as far as possible. However, though we which to have one common data model, experience tells us that

- there is no global or enterprise-wide data model
- different source data models are not always compatible
- standards are difficult to adopt in specific business applications
- model alignment creation still requires expert human knowledge

The challenge is thus to **encourage standardization while remaining open for bottom-up extension** and diversity of model perspectives. The **MODEL REPOSITORY** provides this flexibility with a layered model approach as shown in Figure 4.

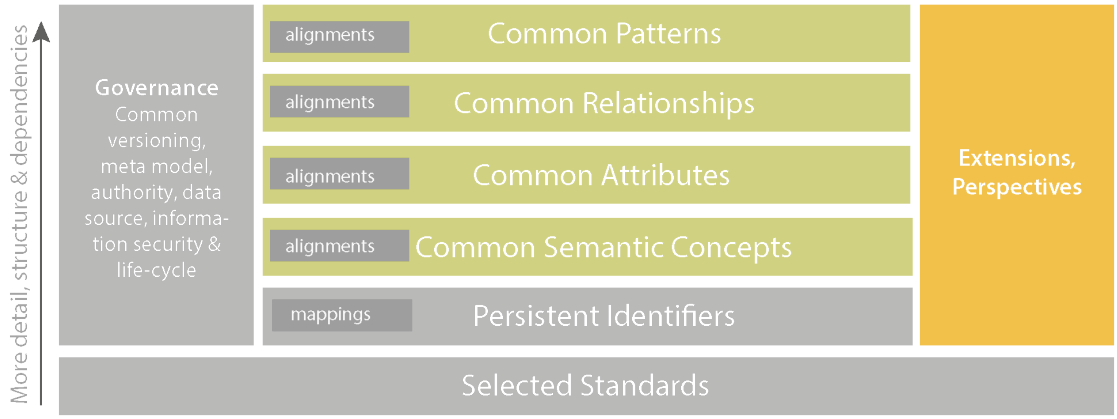


Figure 4: Model stack to enable modular model reuse - grey=MUST, green=SHOULD, yellow=MAY reuse or comply with.

Reuse and standardization are emphasized as part of an integrative, bi-directional approach. As with open source, I can reuse everything and contribute to enhance the existing. To get away from an all-or-nothing situation, we encourage reuse, by providing a layered data architecture approach to building data models with increasing complexity. Note that extension can be made on each level. It is only required to use a shared persistent identification mechanism for all RMD and that minimal governance metadata is aligned. Some of the attributes will be mandatory, and others will be optional. But they will be standardized to some degree depending on the concrete data domain. This means that the layers can be interpreted as agreement layers, which provide the flexibility for distributed RMD producers to align as needed.

With this stack, the main challenge is to get to a critical mass for reuse by making re-use cheaper than re-invention (e.g., making it very easy accessible, high quality, normalization and documentation). Standards in the layered approach can be content standards like IDMP, Allotrope but also format standards such as RDF, SKOS or other.

Reference and Master Data Entities

RMD Entities are identified with a global unique persistent identifier from the Registry, which can be resolved by the resolution component. According to the model stack, for every entity the governance information is available and a reference to a common semantic concept should be given to increase discoverability. Entities of the concept should be harmonized regarding content as far as possible, e.g., the required attributes, filled values used reference lists, units and label style. The core entities of the enterprise needed for data products should be identified and blueprints should be created to assist with integrating any non-standard entity.

Mappings

Mappings are used to align RMD from different sources where the overlap cannot avoided due to historical or business process reasons. It is not always possible to define all-agreed data mappings. What is regarded as the same and what as different can be a question of perspective. Thus data mappings are required to be contextualized so that different perspectives can be captured. Additionally, similarity can be expressed with SKOS or similarity scores.

Operating Model: Assigning Accountabilities in Distributed Management of RMD

Distributed management of RMD does not mean anarchy. Yet, this is what is likely to happen, if we simply say that anyone in the organization is free to manage any set of Master Data or Reference Data without providing governance guidelines. The main reason that anarchy is a likely outcome is that it is not clear what has to be done to manage RMD in general, let alone in a distributed environment. So we can easily imagine that for, say, a particular Reference Data concept, the owner does not really know what tasks have to be done to manage the implementation, or realize that some tasks have dependencies, or understand who will be best person to carry out a particular task. Therefore, the possibility of different people clashing because of lack of clarity about accountabilities is a risk.

The best response to meet these core governance needs is to develop an **operating model** covering roles, responsibilities, policies and standards. An operating model identifies the roles that are needed in a particular context, and what is expected in general from these roles. The roles can be identified and described, as shown in Figure 5.

Concept Owner	Overall accountability for ensuring the implementation of the Reference or Master Data concept has high-quality content and is available
Data Manager / Steward	Accountable for the data provided by a specific authorized source.
Data Architect	Responsible for the structure and alignment between different concepts.
Business Analyst	Gathers requirements for new entities, ensures they are semantically matched to related content. Clarifies questions arising from different perspectives of authorized sources/ providers.
Data Analyst	Investigates how a given Reference Data value or changes on the data models can impact consuming systems, integration pipelines or reporting.
Platform Owner	Manages the IT aspects of the governance node and interfaces.
Data Entry Operator	Updates Reference and Master Data
Data Validator	Ensures that data has been updated correctly (content-wise)

Figure 5: Possible Roles for Distributed Reference and Master Data Management

Distributed RMDM cannot be a volunteering exercise with personal dependencies. We will need to capture what tasks have to be done, or at least most commonly have to be done. A RACI matrix is a good way to do this. An advantage of a RACI matrix is that it shows how the different roles interact, in addition to identifying the specific tasks. Tasks for governing external reference data to be include e.g.,

- Select external authority for Reference Data concept
- Identify any contractual or legal implications for selected external Reference Data concept
- Set up any contract for external Reference Data concept
- Obtain initial version of external Reference Data entities
- Do source data analysis of external Reference Data entities

Another aspect of an operating model is escalation pathways. These are needed for when a particular area of the enterprise has some kind of difficulty with Reference or Master Data that they cannot solve by themselves. Typically, a central resource such as a Data Governance Office is needed that has the necessary expertise or can organize it. At a minimum, this central resource will be able to receive and process escalations that come to it. If the central resource does not itself have the necessary expertise, it might find it in the network of individuals involved in Distributed Reference and Master Data Management. A central resource is also needed to monitor and evaluate the progress of the Distributed RMD program. Reporting on the progress of the initiative and organizing efforts to improve processes are important.

A distributed Reference Data Management program can be seen as a network with a central node with most of the capabilities and activities being carried out in the outer nodes, which can interact. There is a central node that organizes the network and supports the outer nodes rather than directing them.

If we zoom out however the "central resource" is part of a larger distributed governance system as shown in Figure 6. The diagram illustrates that also for governance it is important to include a distributed architecture that allows shared governance between different data offices within and across large organizations.

By deliberately designing the organizational architecture for Distributed Reference and Master Data Management in this way, we will greatly increase the chances of a smoothly running program.

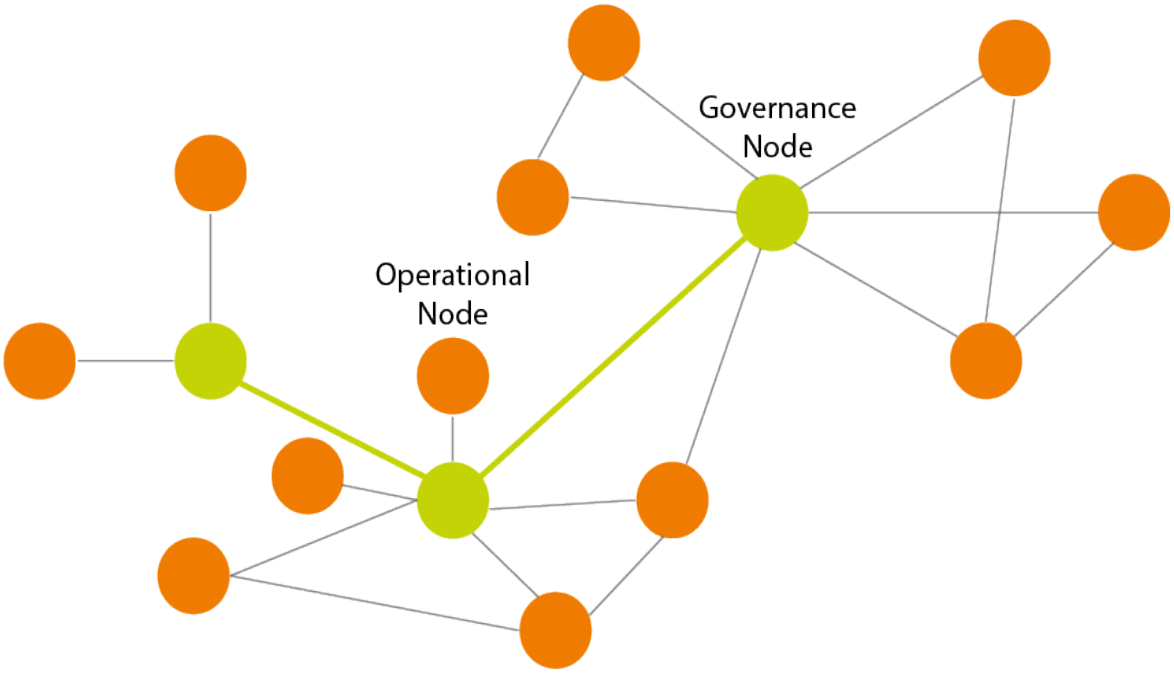


Figure 6: Schematic representation of a distributed reference and master data management operating model with local governance centers as part of a distributed governance network.

Knowledge Eco-System

Registered RMD entities provide the foundation a data eco-system because there we have the agreed upon elements – contributed from different sources. With the basis of the registry of entities and at least partially aligned models, a significant barrier is overcome. On top of this, more complex information and knowledge can be expressed by combining different domains in graphs – forming a knowledge eco-system (see Figure 7). When people understand each other's data regarding the basic elements they can add their specific knowledge. It is like people have the terms and grammar: now they can talk. Four aspects are important here:

1. As in the common reality, we not always agree and different viewpoints can be valid. Often, right or wrong can be decided only in the context of a particular perspectives. Cross-domain linking depends on the specific application scenarios.
2. Cross-domain links express domain knowledge (e.g., gene-disease associations), some of which may be proprietary not be shared freely.
3. Context gets important: e.g., probability and provenance
4. Modularization is key to enable further re-use and alignment on overlapping knowledge areas.

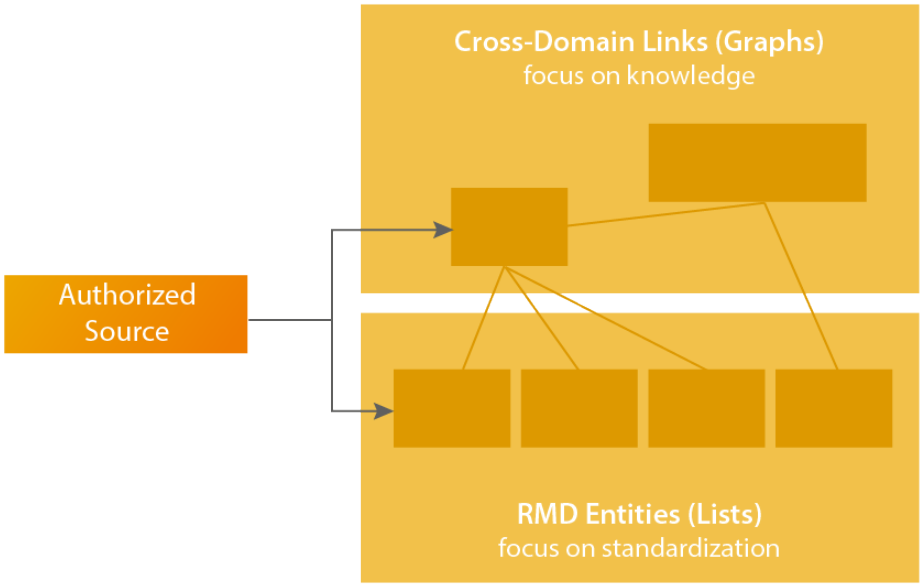


Figure 7: Knowledge can be effectively exchanged when it is based on common Reference and Master Data entities.

Now think about your organization.

The Path Forward.

Which issues do you see related to reference and master data management?

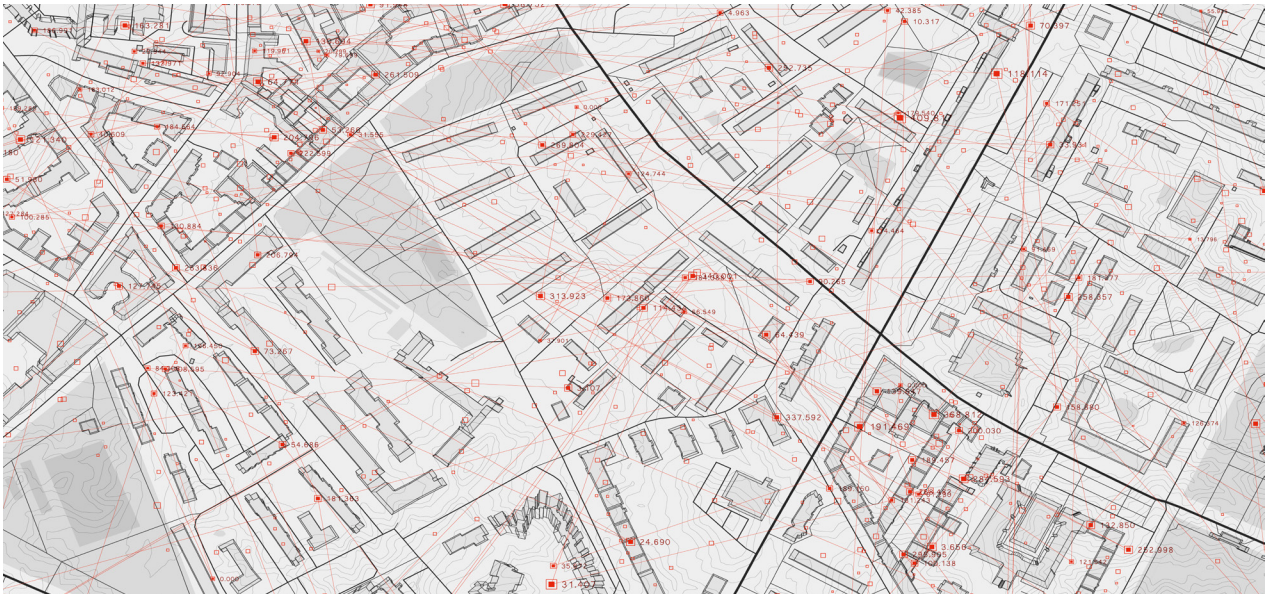
Which data is managed already in a distributed manner?

Which data governance mechanisms are already in place?

How do you align governance initiatives from different departments or functional units?

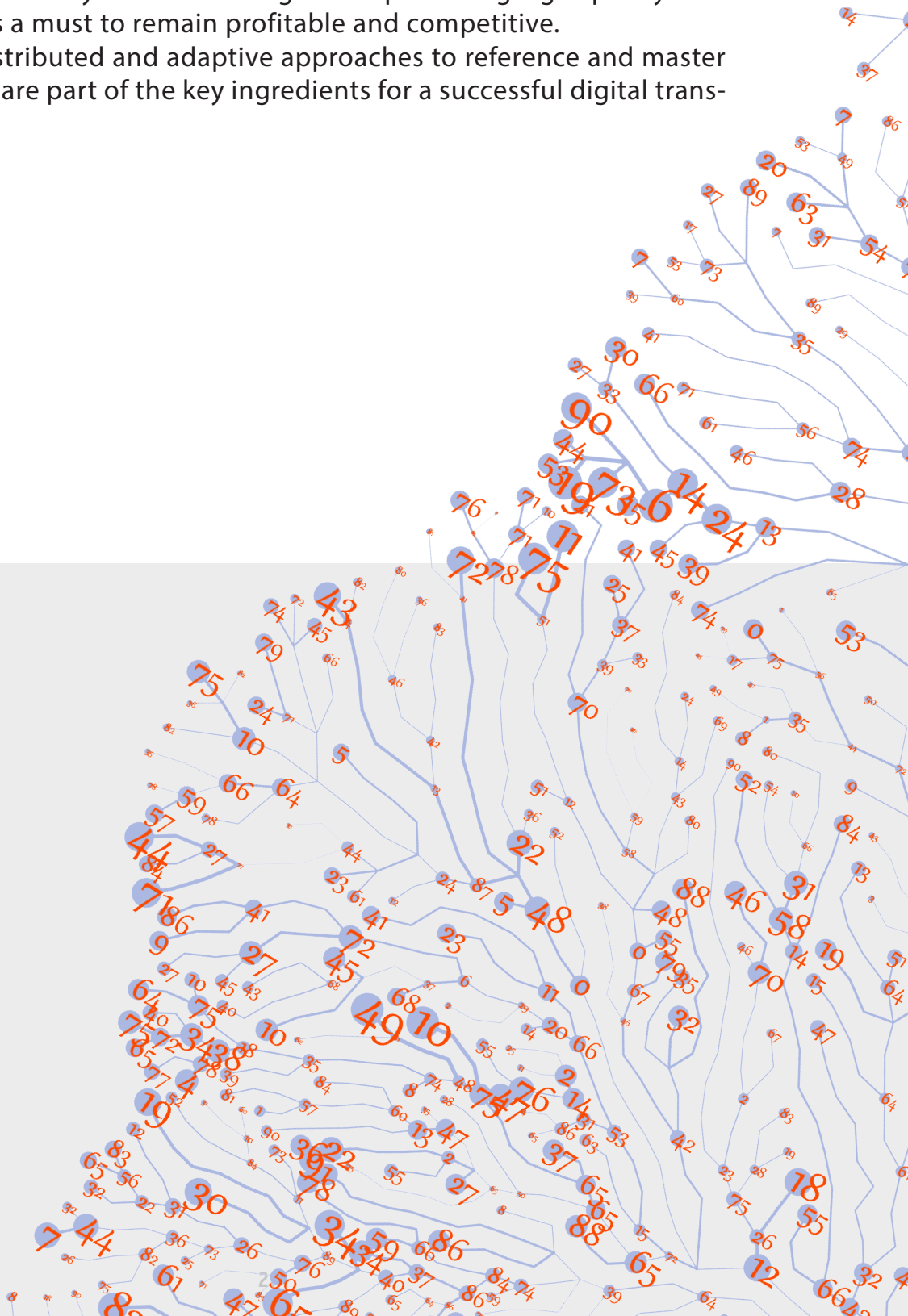
Which use-cases could be used to test the approach?

What would be a good scenario to try a distributed approach in the described way?



Conclusion

The digital transformation is experienced as a competitive race in the pharmaceutical industry. Significantly increasing pressure from regulators as well as a tremendous need to speed up innovation capabilities time-to-market requires to work differently with data. Being able to producing high-quality data products at scale is a must to remain profitable and competitive.
We believe that distributed and adaptive approaches to reference and master data management are part of the key ingredients for a successful digital transformation.



About the Authors

Heiner Oberkampf is Head of Data Governance at OSTHUS and a co-founder of Accurids Inc. His approach to technology has a strong focus on customers, business value and cross-organizational collaboration. Working as an intermediate between business and technical community Heiner helps clients to translate their business needs and goals into an information and data governance strategy. He can be contacted under heiner.oberkampf@osthus.com.

Christian Senger is working as Senior Data Integration Scientist at Bayer. He has a strong background in data architecture, storage, and management in Life Sciences, particularly Bioinformatics, Medical Informatics, and Pharma - always acting as intermediary between science and technology. He is dedicated to sustainable provision and governance of high-quality reference data and data models in an understandable, reproduceable, reusable, and interoperable way – for most efficient research and product development.
He can be contacted under christian.senger@bayer.com.

Malcolm Chisholm has over 25 years experience in data management, and has worked in a variety of sectors, including finance, pharmaceuticals, insurance, manufacturing, government, defense and intelligence, and retail. He is a consultant specializing in data governance, data quality, data privacy, master/reference data management, metadata engineering, data architecture and business rules management/execution. Malcolm is a well-known presenter at conferences in the US and Europe, writes columns in trade journals, and has authored the books: Managing Reference Data in Enterprise Databases; How to Build a Business Rules Engine; and Definitions in Information Management. He holds the prestigious DAMA International Professional Achievement Award for contributions to Master Data Management, and can be contacted at mchisholm@datamillennium.com.

References

[1] Platform Revolution: How Networked Markets Are Transforming the Economy - And How to Make Them Work for You - Book by Geoffrey G Parker, Marshall Van Alstyne, and Sangeet Paul Choudary

[2] The FAIR Guiding Principles for scientific data management and stewardship - Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[3] Evaluating FAIR Maturity Through a Scalable, Automated, Community-Governed Framework - Mark D Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, Tobias Kuhn, Philippe Rocca-Serra, Mercè Crosas, Erik Schultes bioRxiv 649202; doi: https://doi.org/10.1101/649202

[4] FDA's new pharmaceutical quality initiative: Knowledge-aided assessment & structured applications - Lawrence X. Yu, Andre Raw, Larisa Wu, Christina Capacci-Daniel, Ying Zhang, Susan Rosencrance; International Journal of Pharmaceutics: X, Volume 1, 2019, ISSN 2590-1567, https://doi.org/10.1016/j.ijpx.2019.100010