Prescriptive Master Data[1]

A systematic approach to master data excellence

1 Introduction

This white paper introduces the generic concepts and features that are necessary for a master data repository (MDR) to effectively support a data-centric information system. The methodological and organizational aspects of data governance being well documented, this paper focuses on the functional and technological aspects of the matter.

2 A complex challenge

Digitalization and interoperability make MDRs ever more relevant: people, organizations, places, facilities, categories of all kinds and other nomenclatures, all need sharing throughout information systems.

Building a robust and reliable system proves, however, to be extremely challenging.

Complexity factors include:

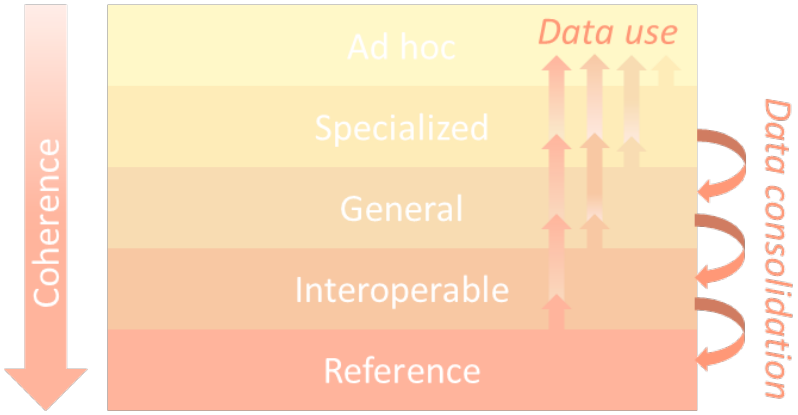
- High volumes, frequent changes;
- Number and diversity of sources;
- Diversity of the data (concepts, integrity, scope, time granularity, detail level...);
- Variability of the data quality and actuality;
- Difficulty to identify entities and convert data to a common ground;
- Data statuses and life cycles, integrity interdependencies, asynchronous processing;
- Conflicting values, exceptions, contextual values;
- Security requirements, including fine access control and compliance to personal data protection regulations.

Complexity tends to increase over time, following the evolution of these factors: systems' complexity augments along with volumes, functional coverage and need for integration, while legislation becomes more and more demanding and security threats get increasingly sophisticated.

3 Data coherence model

Data categories are not equal: some are structuring for the whole organization, other have a very narrow scope; some are durable, others have a relevance limited to a short period of time.

The information system urbanism and the master data technology need to take into account this diversity. In particular, they should allow, and even support effectively the natural stratification of the data as described in the following schema.



- At the top of the data stack lies **ad hoc data**, which may integrate information from other layers but is not technically an input for the rest of the information system; its volatility is therefore not limited. This data is often in human-oriented, productivity-tool format, such as spreadsheets or presentations, emails or simply instant messages, audio or video;
- Just below in the system lies **specialized data**; this data remains local to the part of the system where it is produced, i.e. limited either to a part of the organization (e.g. machine control data) or to some loosely-integrated cross-function (e.g. vCal appointments);
- Farther below lies **general data**: although not integrated throughout the organization, this data is relevant beyond its original scope, and deserves to be routinely[2] brought together, compared or aggregated in some monitoring tool or recurrent analysis; this data needs to be consistent enough to be usable in those consolidations;
- Deeper in the system, strictly inter-operable **data** becomes necessary to coordinate tasks throughout the system: for example, the details of a purchase order needs to be carried over accurately and understood exactly at each step to be processed throughout production, shipping and billing;
- At the fundamental level lies the **reference data**; this data is central to several domains, which requires robustness and stability; when the reference data quality (relevance, accuracy, exhaustiveness) is high enough, it can be used prescriptively, thus structure the processes within the organization, and even beyond (e.g. product catalog).

Data use takes place inside or above the layer to which it pertains, at no cost ; consolidating the data from a layer to a deeper one, to the contrary, requires a specific data organization effort, through some combination of the following means:

- Use of heuristics: induction rules, statistical analysis (big data);

2 'routinely' is important here: obviously, it is possible to draw statistics from emails content, appointments or volatile machinery metrics, for example; however such practice is no routine, and its results typically belong to the ad-hoc data layer.

- Addition of value: data stewardship, data quality management.

This asymmetry between data use and et consolidation is not only unavoidable, but also soundly founded in information theory : just as in thermodynamics, data quality (exactness, homogeneity, completeness...) ordered[3] to their use (relevance, availability, actuality...) will not appear spontaneously ; an effort (cross-checking, correlation analysis, inquiries, data governance[4]) is necessary to increase it.

4 The basics of prescriptiveness

The very purpose of reference data is to be used systematically where appropriate, in order to ensure the coherence and effectiveness of data exchanges and corresponding business processes.

Data excellence is a prerequisite to such prescriptive references: prescriptiveness decreed on sloppy data is due to fail at best; applied blindly, it could even downgrade the organization's data relevance and derail its processes.

While big data's statistical approach is sufficient when prescriptiveness is not at stake, when the data not only guides but structures business processes, exactness becomes mandatory.

Such data excellence raises no particular challenge regarding clearly defined concepts with fairly stable instances and values, especially when a unique, coherent source of information is available; such are, for example, ZIP codes or time zones.

To the contrary, reaching the data excellence that will allow achieve prescriptiveness in areas where the complexity is high requires significant investments, as data excellence requires a combination of organization and tooling:

- Without data-centric processes and training, the best tools remain powerless, as they can't create the missing information[5];
- Conversely, without effective data integration mechanisms, without the ability to express the richness of the information[6], technology will fail to reduce the costs of data quality management to the point where the full coherence potential of the organization's data becomes achievable.

More specifically, rigorous modeling and data quality management are necessary to provide proper data accuracy and completeness, hence repository relevance as a prescriber:

- The data model of the repository must be powerful enough to accommodate the complexity of the information, which depends on time, location (e.g. language) and context, as well as the relevant meta-information;
- Powerful tooling is needed to support data integration, stewardship and quality management;
- The access control and journaling model must enforce regulations as well as corporate policy;

3 with minimal entropy

4 e.g. Civil Registration, business register office

5 Data and meta-data

6 Data and meta-data

- The implementation needs to meet the desired performance and security requirements;
- Most importantly, processes need the same level of precise tuning: flawless organization, sufficient resources and appropriate training are required to cover data stewardship, data quality management, access control, compliance and governance.

5 Soft prescriptiveness

When the diversity and the sophistication of the systems using the reference data increase, specific challenges arise:

- Long transactions: the more data sources, the more potential conflicts to address, some through manual stewardship; it is not possible to guarantee that the reference data will be quickly updated; when activities relying on the reference data need an update to be taken into account immediately, they must be granted proper isolation from the legacy data until their transaction is fully processed;
- Timely processing of changes: activities may need to be organized in a way that optimizes the handling of certain changes, e.g. through batch processing; or, they can need to freeze their view of some data until some on-going process related to it has finished. In either case, they may need reference information updates to be delayed for them;
- Contextual truth: in some cases, although correct, the reference data is not suitable for a subset of the information system; this may be the case in particular when a more accurate information is to be shared between some applications but not system-wide.

Giving up prescriptiveness altogether in front of these challenges is the most common practice: extra, loosely-coupled data sets are created to accommodate the respective needs of the various activities, thus sacrificing at least part of the integration benefits of reference data.

However, proper tooling and organized approach allow embrace this whole complexity without losing coherence. That "soft prescriptiveness" is built on the following principles:

- For each specific information type, unique, canonical reference values remain in force; for each data category, a designated authority has the final say about maintaining those values whenever data conflicts arise; canonical values are softly promoted: they are used by default and remain always accessible;
- Contextual values can overload the canonical values for a given perimeter when appropriate, either to provide isolation pending the end of a transaction or durably for some functional reason;
- Requests to update the references can specify that the supplied data is deliberately contextual, or that it should be kept such in case it were not accepted as canonical data by the data integration process;
- Metadata (status, comments...) allows to describe the reasons behind the contextual values and handle their life cycle;

- Canonical data is available to any user or process having access to a corresponding contextual value, allowing for gap analyses at any time.

Soft prescriptiveness does not only grant flexibility to the business processes: it is an irreplaceable data quality tool, thanks to its ability to easily collect and organize structured information revealing:

- Alternate, potentially more accurate or up-to-date data values;
- Data model issues, such as concepts confusions or wrong cardinalities;
- Business process discrepancies.

The flexibility of soft prescriptiveness requires rigorous management in order to prevent abuses. In particular:

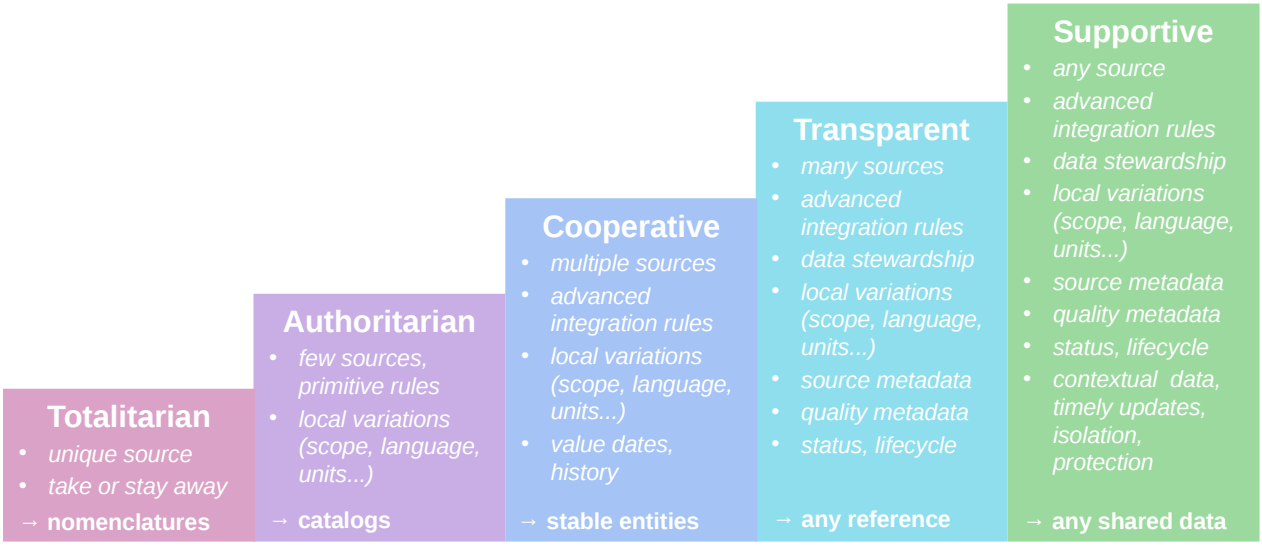
- Concepts that are ontologically distinct (e.g. department vs cost center, official name vs usual name...) need to be identified and clearly separated into distinct data containers, rather than addressed with contextual data;
- Concepts that have some similarity with others in the master data, but are in fact specific to an activity, should not be part of the master references in the first place.

Quality processes should review contextual data either for cleansing or to identify patterns revealing conceptual defects or organizational discrepancies.

6 Master data technical maturity model

Master data maturity models usually address the data governance but lack specifics regarding concrete, mere technical implications.

The maturity model represented in the diagram below, to the contrary, focuses on the technical achievements that, with the proper organization, will make that governance possible and practical for each master data category.



The available repository technology in an organization may be too limited to handle the most demanding data categories, due to their complexity (sources, variability, quality, usage).

In other words, the repositories' technical capabilities define the maximum potential extent and usage of the master data. The data governance ambitions and the repositories' technical road map need therefore to be carefully aligned.