Data Management Guidelines

DMG 13 – Master Data Management (Incorporating Reference Data)

July 2021

Contents

[Purpose 2](#_Toc77772882)

[Relevance 3](#_Toc77772883)

[Definition 6](#_Toc77772884)

[Approaches for Managing Master Data 9](#_Toc77772885)

[Approaches for managing reference data 12](#_Toc77772886)

[Master Data Architecture 16](#_Toc77772887)

[Master Data Design for Systems 25](#_Toc77772888)

[Useful Resources 31](#_Toc77772889)

[Document Administration 32](#_Toc77772890)

[Useful Resources 33](#_Toc77772891)

Purpose

This document contains the following sections;

* Why MDM & Reference Data are important
* Definition of Data Types
* The Standards relevant to MDM
* Approaches for Managing Master Data
* Master Data Architecture
* Master Data Design for Systems
* Useful References
* Document Administration

Relevance

Incomplete, incorrect, outdated, redundant, inconsistent, or misaligned Master Data is likely to have adverse impacts on safety, business process efficiency, business intelligence, and regulatory compliance. Conversely, accurate and up to date Master Data is critical to the economic, customer service and safety performance of National Grid. Master Data is also a key enabler of standardisation and efficiency in business operations and in the tactical and strategic use of the management information.

In National Grid, our technology landscapes have evolved over time and we generally have multiple copies of data stored in varying formats in different systems or spreadsheets with poor control of those copies and often without a general agreement on which is the master authoritative source to be used. The maintenance of this duplicate data is performed separately and in isolation and often in differing parts of the organisation. To prevent discrepancies between the data copies on different applications, it is vital to ensure appropriate mechanisms are introduced to ensure consistency of the Master Data is maintained.

The same or similar Master Data is often used and maintained by different departments and people with different perspectives therefore we need a consistent Master Data design across all those areas. Traditionally we have developed solutions without the use of standard reference tables (lookup tables) which undermines our ability to manage, integrate and exploit our data.

This poor Master Data quality can and does have significant adverse effects on the overall performance of National Grid:

* Bad Master Data causes expensive disruptions in operations and processes, such as:
  + Incorrect planning due to the wrong address leads to re-scheduling, additional follow up work and reduced customer satisfaction.
  + Inconsistent Master Data in work management causes incorrect financial coding of work requiring manual correction.
* Bad Master Data can lead to increased operational risk, such as:
  + Problems with traceability of components and component types
  + Incorrect assessment of asset management requirements based on inconsistent asset and work data across multiple applications
* Bad Master Data undermines business information, such as:
  + Inconsistent business information is a key issue during regulatory reporting processes
  + Manual off-line realigning of key information results in a high cost of data preparation and analysis, reduced timeliness and decreased accuracy.
  + Discussions and doubts about the accuracy of data cause many business discussions to revolve around data arguments and disagreements rather than the business issues.

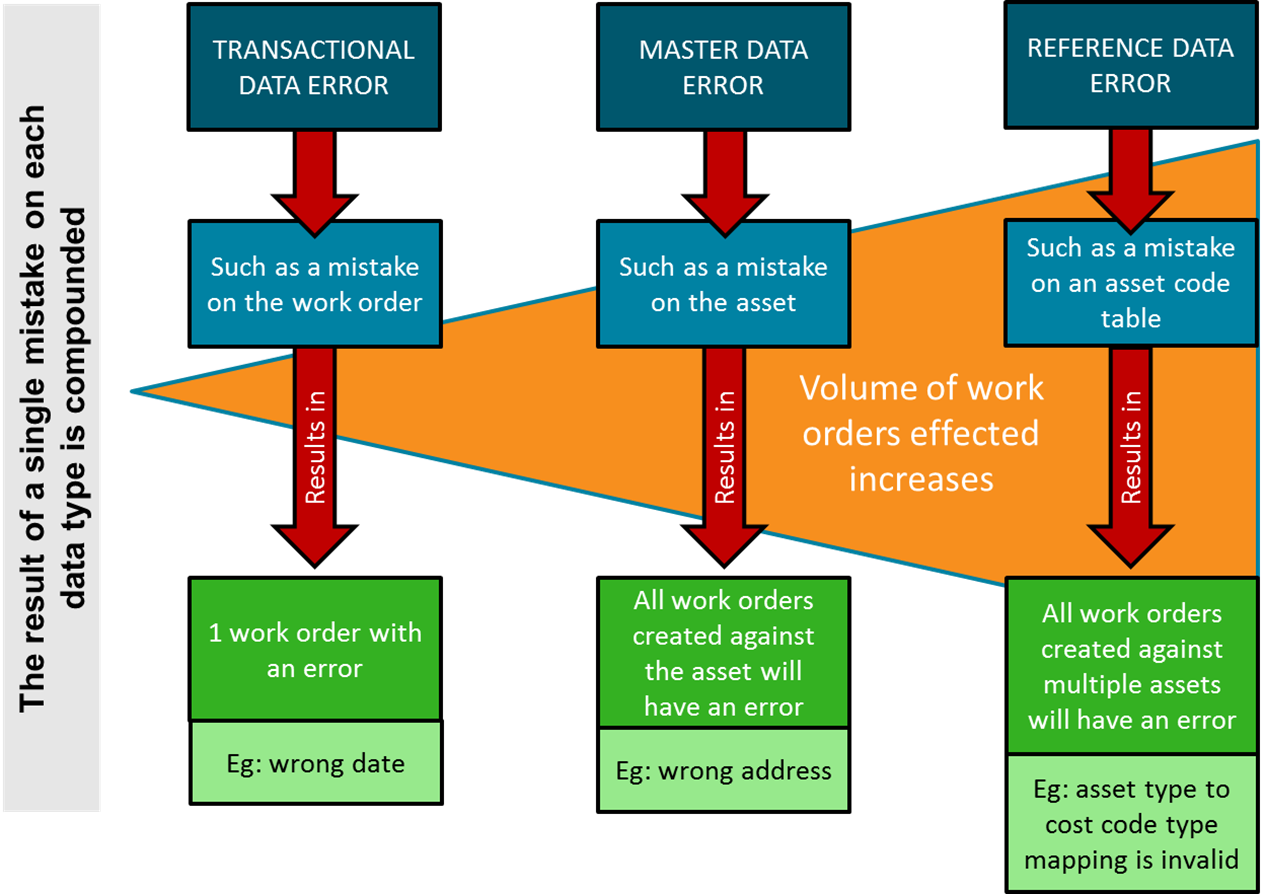
The key reasons for proactively addressing the management of multiple copies of data are:

* There are overheads associated with maintaining multiple copies of data in multiple repositories (whether in centrally controlled systems or in spreadsheet copies). If we can reduce the level of duplication of data sources, we can reduce the maintenance overheads and achieve a more efficient and accurate level of data control.
* Having the option of sourcing reports from alternative data sources creates unnecessary problems. Significant time is consumed when resolving reporting discrepancies and inconsistencies when it is unclear which data sources have been used, and which are the right master sources to use.
* We need to ensure people understand and use the ‘right’ data sources for operational purposes and reporting – thus reducing the opportunity to create reporting discrepancies.
* A critical factor in managing the use of copies of data is to ensure users access slave copies of data at the right time (i.e. recently synchronised with master) – running monthly reports the day before the monthly refresh should be challenged!

Implementing Master Data Management (MDM) will enable National Grid to realise significant benefits by:

* **Improving operations**. Accurate asset location information will significantly improve efficiencies and reduce costs associated with some of the key business processes.
* **Increasing customer satisfaction**. Effective Master Data management can provide a unified view of the customer by leveraging customer data across multiple platforms. In addition, Master Data management is a key prerequisite to reaping benefits from innovative use of customer interaction platforms, from a monthly bill to the self-service internet portals.
* **Increasing insight and transparency**. Clean and accurate data improves organisation’s ability to measure and interpret organisational performance and serves as a key enabler for improved business decisions.
* **Ensuring compliance**. Well implemented Master Data management programmes can help utility companies address the critical business concern of regulatory compliance. Master Data management will address a broad range of compliance and governance issues, including Sarbanes-Oxley, IFRS and Global Trade Compliance. Benefits of improved Master Data quality and the implementation of the correct processes in these areas can yield significant savings in audit related costs.

The other type of data that this guideline covers is Reference Data. This data type is frequently over looked when business entities are implementing Master Data governance and change control processes as it is much lower in volume and frequency of change than traditional Master Data, however, it can have significantly more far reaching consequences (as shown by the diagram below) and therefore is at least as important to manage appropriately.

For example: One mistake on a transaction will only affect that transaction. One mistake on a bit of Master Data will affect every transaction that uses the Master Data item. One mistake on a bit of reference data will affect every bit of Master Data that uses it and in turn every bit of transactional data. Thus, the error is compounded.

In summary, successful Master Data Management relies on a high-quality design of the logical data structures, of the use of standard data objects, of the governance and data maintenance processes and of the design of processes and interfaces which keep consistency across multiple applications.

The scope of Master Data management here is to ensure master source data and related copies are successfully synchronised and aligned on a frequency appropriate to their business use. Where we have replicated data and no master source is formalised, the requirement is to establish the best source to deem master and then establish master-slave relationships between copies of data.

It is therefore advisable, during any data change project, to dedicate significant attention to:

* The target system Master Data structures
* Master Data maintenance processes

Definition

In National Grid we define Master Data Management (MDM) as:

*“Master Data Management is the on-going reconciliation and maintenance of Master Data”.*

In order to understand the importance of Master Data, then it needs to be seen in context of other data types. All data can be assigned to one of the following definitions:

|  |  |
| --- | --- |
| Data Classification | Description and Example |
| Master Data | These are the ‘nouns’ of the business – they are the things that will participate in events. These will be referenced in order to perform transactions.  Master Data is that which is used in multiple systems or components of systems i.e. it is duplicated or shared across the system architecture and therefore needs to be managed in such a way that all copies remain synchronised and all components that need to share it remain compatible.  From a DAMA perspective the following definitions apply:   * Master Data is the authoritative, most accurate data available about key business entities – this may be held all together in one system or the master components could reside across multiple systems * Master Data is data about the business entities that provide context for business transactions   Example :   * Asset, Customer, Project, cost centre, site. etc. |
| Reference Data | Reference data is data that defines the set of permissible values to be used by other data fields. Reference data gains in value when it is widely re-used and widely referenced. Typically, it does not change overly much in terms of definition (apart from occasional revisions).  Example :   * Type of asset: such as: transformer, compressor, tower, pipeline, meter, valve, etc. |
| Metadata | Metadata is data that is used to describe data. I.e. data which describes the structure and taxonomy of information. Defining metadata correctly is critical to the ability to store and manage the data at a practical level. Correctly done it can help support data quality by predefining data capture parameters.  Example :   * A field should be 8 characters long, the format is numeric only, it is mandatory (i.e. cannot be left blank). * The document is saved in pdf. |
| Attribute Data | The items of data that describe the fixed characteristics of the entity, generally non or slow changing in value.  Example :   * The commissioning date for a specific piece of equipment * The length of a specific pipeline |
| Measurement Data | The value recorded for a time varying characteristic of an entity  Example:   * A Transformer oil dissolved gas value (taken for a specific transformer oil tank at a specific time) |
| Transaction Data | Transaction data is data describing an event (the change as a result of a transaction) and is usually described with verbs.  Transaction data always has a time dimension, a numerical value and refers to one or more objects (i.e. the reference data).  Understanding data as being ‘transactional data’ will require us to understand the volumes involved and to determine the appropriate storage requirement.  Example :   * Typical transactions are: Work orders, Financial: orders, invoices, payments. |

The two key areas covered by this document concern our management of:

* Master Data: we need to actively manage our current legacy of replicated data and improve controls such that the level of replication is reduced as opportunities arise. If we can encourage more REUSE of Master Data, the overheads in managing replication can be reduced.
* Reference data: we need to establish common reference tables which are shared across multiple systems. If we can establish shared and common reference tables, their REUSE will have real business benefit as described in the section below.

There are less concerns currently with our overall management of ‘measurement data’ and ‘transaction data’ or ‘attribute data’ – if there are issues with these, they will be addressed on a localised basis.

While defining Master Data can be fairly straightforward, in practice there can be some confusion on whether an object should be considered and managed as Master Data or whether it should be treated as another (e.g. transactional) data type. Some general indicators that may help to decide whether the object should be considered and managed as Master Data are listed below:

* **Behaviour** – Master Data interacts with transactional data within the organisation. Most commonly when one describes a business interaction Master Data is captured within a noun while transactional data type is captured by a verb in the sentence (e.g. “Vendor sells a part, and a partner delivers materials to a location”)
* **Cardinality** – in general, as the number of records in data set decreases the need for data elements to be considered and managed as Master Data decreases (e.g. if a company has five customers it does not need to deploy a MDM solution and full governance structure to manage this data). Though it is still Master Data and should have the appropriate processes and controls.
* **Volatility** - Master Data tends to be less volatile than transactional data.
* **Re-use** – Master Data objects are typically re-used in the course of businesses operations. I.e. shared or duplicated across systems.
* **Value and complexity** - The more valuable the data object is to the business and the more complex it is (in terms of attribute number, cross attribute validations etc.) the greater benefit there is when it is to be treated as a Master Data element within an MDM solution.

Approaches for Managing Master Data

Data Management Standard, Principle 5 requires that data has an authoritative source (Master Data). The underlying philosophy of MDM is that the authoritative source should not be duplicated, replicated or distributed into multiple locations, unless it is a non-editable version and the data flows are clearly understood and documented in the Data Library (via the data models, data dictionary and business data glossary, which is directly related to BMS requirement 5.1). Breaking this principle will require maintaining/updating in more than one place with the complexity, and inevitable failure, of maintaining alignments.

If there is an insurmountable reason why data updates have to be editable across multiple copies, then the authoritative source must still be established so that when the duplicate becomes out of sync it can be refreshed from the master. (This is directly related to BMS requirement 5.2). The master must always be updated first, followed immediately by the copies. The copies should never be updated first and risk being more accurate than the master. If this happens the business will no longer know which source to trust and both data sets become useless. A situation where critical data is editable in two systems should be added to the local risk register to ensure the situation is monitored and controlled. As shown above, an error in Master Data is quickly compounded.

In summary, we need to ensure we are clear regarding:

* Which is the ‘master’ copy of the data i.e. the source which is actively maintained
* Which, if any, are the ‘slave’ copies i.e. the sources which should simply be refreshed from the master copy
* How to manage the master-slave relationships across the data copies

The sections below go into more detail of some of these considerations.

## Understand What Data we have

We must firstly understand what data we have in what data repositories and which data is potentially a duplication of other data. A Systems Inventory allows us to formally record the range of data repositories used in a business area and add additional information regarding whether they are centrally supported, or user maintained.

## Establish which systems are agreed master sources

We need to establish which of the identified systems hold what types of data e.g. which hold asset data, which hold work related data, which hold finance data. Once we understand what type of data is held where, we can start to understand which systems hold copies of the same data.

At a surface level, two systems may appear to hold the same particular type of data, but it will be necessary to interrogate key personnel to understand if there are subtle differences between them e.g. one system holds planned assets and another system holds live assets OR one system holds employee pay related data and another system holds employee training data. Detailed interrogation is required to truly confirm if the datasets within systems represent the same business data requirement. It is also worth point out the following points during interrogation;

* The data in the system of origin doesn’t automatically mean it is the authoritative source (though sometimes it will be).
* An authoritative source isn’t always a centralised data warehouse
* The data in the system of origin might not be clean, quality checked or ready for consumption.

When establishing the Authoritative source it is worth considering the following:

* Business data requirements and data movement
* How the data will be shared; internally and externally (subject to Data Sharing Agreement between consumer and provider to accommodate the obligations?)
* How Data quality will be managed
* Cost of data integration
* Data architecture
* Ownership and Authority
* Controlling change

Business rules…etc.

## Establish improvement plans to reduce the level of data replication

A business area will establish improvement plans to reduce the level of data replication if this is contributing to data management risks and concerns. A simple step in this approach could be to reduce the level of user generated extracts and excel copies of data. This applies to multiple reports generated from the various sources that arise from duplication of the Master Data from its Master source/system. The Data used in the reports must be traceable to the Master source or to be distinguishable from the master source/system. (This is directly related to BMS requirement 5.3).

## Establish relevant synchronisation processes

Having identified the level of true replication of data sources, we need to ensure we have appropriate processes in place to maintain copies of data i.e. the copies of data are synchronised on an appropriate frequency to the agreed Master Data source. The refresh mechanism is ideally managed on an automated basis rather than rely on manual refreshes which are prone to manual error.

Features of the synchronisation process will include:

* Confirmed identification of systems/datasets with duplicated data requiring synchronisation.
* When master systems have new instances added (e.g. a new asset is installed), a synchronisation process creates the equivalent record in the other relevant systems. The synchronisation process may also have triggers updating related dependent data.
* When the master system has its Master Data changed or deleted, the synchronisation process applies that update to the slave systems within the appropriate refresh cycle.
* Appropriate audit trails are created to confirm last synchronisation date and time.
* Controls are created and acted on to ensure the copies of data from Master Data sources to slave sources complete correctly.

In the event of a Disaster Recovery (DR), the Master Data recovery state will be reflected in the slave systems as part of the DR process; when slave systems require to be recovered, we are clear on how this aligns with the scheduled refresh cycle.

## Establish clear roles and responsibilities

If data is to be duplicated, then roles and responsibilities should be established for fully managing the multiple copies and the processes associated with them.

* We need to be clear who has accountability for approving the refresh frequency of copies of data.
* We need to confirm who can approve a change in the refresh frequency.
* A role is required to approve the use of data copies in Management Information, rather than accessing Master Data.
* Approval points should be established to approve the creation of more copies of data in the design of new solutions with appropriate justification submitted to the business governance processes.

## Formalise documentation of data repositories

We need to ensure we utilise the systems data inventory and data dictionary to formally record where we have copies of data, and which system/dataset is approved and agreed to act as Master Data. We must ensure we properly understand master-slave relationships between copies of datasets. We would expect this documentation to formalise the required refresh frequencies.

## Cross referencing of related systems

Appropriate cross referencing must exist to facilitate the synchronisation and updates of the systems involved in replicated data i.e. use of primary and foreign keys to identify the relevant records to be updated on slave systems.

## Manage appropriate use of slave systems

It would be normal to report from Master Data sources; in the scenario where reports must be sourced from slave systems/data, we need to ensure the reporting is scheduled to run at the right time i.e. after recent synchronisation.

Approaches for managing reference data

Longer term, we may wish to maintain certain reference data at enterprise or entity wide levels but at the start of our data maturity journey we can create operational efficiencies and drive more collaborative data management within a business area by establishing its appropriate suite of reference tables.

The target needs to be to aim to establish common and consistent reference data across all the systems used within a business area. At the outset we may have the following ‘current state’:

* Reference data not used within a system – data validations operate from a specific list of values at best or values hard coded into a solution; in these scenarios we are heavily dependent on user knowledge and experience for managing change or interpreting data
* Reference tables setup for an individual system independent to reference tables for another system
* User controlled IT (spreadsheets etc.) operating from a hard-coded list of allowed values
* Detailed scenarios highlight the benefit of a business area having common reference data or a reference table

The following applies across all its supported and non-supported systems:

| **Objective** | **Description** |
| --- | --- |
| Support consistent reporting from multiple systems | When integrating data from multiple systems, it becomes more important to have consistent use of reference data at business level rather than try to deal with system specific reference codes e.g. system 1 uses S123 to mean manufacturer = Siemens and system 2 uses SIMS. How do we build reports which bridge both systems? Consistent use of reference data codes across systems aids interrogation of the combined datasets. |
| Avoid hard coding values into systems or reports | If we embed actual codes and values for reference data (e.g. manufacturer codes) into our systems configuration or report definitions it becomes more difficult to make changes or assess impact of change |
| Avoid use of free text fields | It is often seen as a shortcut to make use of user definable fields in systems for a new requirement. For example, if a free text field is used for capturing manufacturer identity, we run the risk of creating inconsistent entries and enquiries will give inaccurate results |
| Verify data at source | The existence of a reference data table allows data to be validated at point of capture and rechecked elsewhere in an end to end business process. Adding new codes to a reference data table allows data to be added only once but which can be used in data quality validation in multiple instances and systems. |
| Drive a reduction in data quality errors and establish more efficient delivery of change | If we start to increase our use of reference data we can start to more easily manage changes or additions to the reference table once only rather than run the risk of needing to complete maintenance in multiple points of hard coded verification. |
| Avoid overly large value tables | If value tables are too large the users quickly find short cuts. They get fed up with scrolling down to find the right answer. It is well known that values at the top of a large value table are more frequently used than those at the bottom. In addition, large value tables often end up with duplicates because those managing the master updates can’t see the wood for the trees and add a value already there that they missed. |
| Avoid the use of overly complex compound codes | This applies equally to Master Data and references data.  Building logic into codes can have a business value, but this should be kept simple and remain for a single purpose.  For example it might seem easy from a data out-put (eg reporting) to have one field where the 1st two characters mean one thing, the 2nd four mean something else and the last five mean something different again. It feels like you can tell so much information from this one code. However, it means that each time a new section is added it multiplies the total number of codes required by the number of options in the new section. This makes value tables unmanageable (see above). In addition, combining too much information in one code restricts the number of digits available for each section. These usually means that over time meanings will have to change as letter/number combinations run out. For example: before 1990 X=Y, but after 1990 X=Z.  If these 3 items of information are kept in 3 fields they can be combined and filtered in reports and data out-puts to enable the data to be cut in any way the business needs it. This also allows future flexibility so the data can be re-cut to meet requirements the business is not yet aware of. |

## An approach for creation of business area specific reference data

The approach to retrospectively building an appropriate reference data foundation may seem daunting and therefore we must address this one step at a time. A suggested approach is listed below:

1. Catalogue existing reference data usage for potential broader use across systems used within a business area
2. Select a frequently used set of reference data as initial priority where it is known there is inconsistency between systems e.g. asset manufacturer identifier
3. Identify those systems (e.g. from the Systems Inventory) requiring the same reference data.
4. For each system

* Confirm existence of relevant reference tables
* Compare and contrast content of reference tables to derive a consolidated set of requirements – if necessary create translations between sets of tables
* Where the reference table existed previously replace with the consolidated table or the subset relevant to the system – if the codes are changed, existing data may need to be migrated to using the new codes.
* Where the reference table did not previously exist, make the table available to the system so that future solution designs and modifications can start to use the updated reference table

1. Ensure projects making change to the affected systems maximise the opportunity to use the reference tables when they next make change. This should be part of the business led technical design review process.
2. For brand new solutions, the business design review body will need to establish the use of reference tables as one of its quality assessment criteria.
3. Ensure the reference tables requirements are recorded in the business data glossary.
4. Ensure the reference table implementations are recorded in the system specific data dictionaries.

## Controls required to maintain reference data within a business area

Key controls required to maintain the completeness and usage of reference data include:

1. Governance and accountability: a business owner is required to be accountable for reference data (or subsets of reference data tables) definition and content, approve the issue of new codes and any changes to descriptions.
2. Data models/flows should be established for all Reference data entities so it is always clear what systems and processes use them. That way any request for change can be properly assessed.
3. Stakeholders must be consulted and informed on any additions or changes – the stakeholders may be located outside the immediate business area.
4. Processes will be established for managing change in content and definition. Service level agreements may be required to ensure an appropriate turnaround for ‘new codes’ as delays may mean local workarounds are adopted, undermining the formalise reference data creation.
5. Governance via Design Review Boards will ensure new or modified solutions are designed to access centrally maintained reference tables. Reviews will ensure opportunities for retrospective build of reference data tables are pursued.
6. Reference data availability will be clearly visible to business users to broaden usage and support for reference data.
7. Reference data business requirements are recorded in the Business Data Glossary.

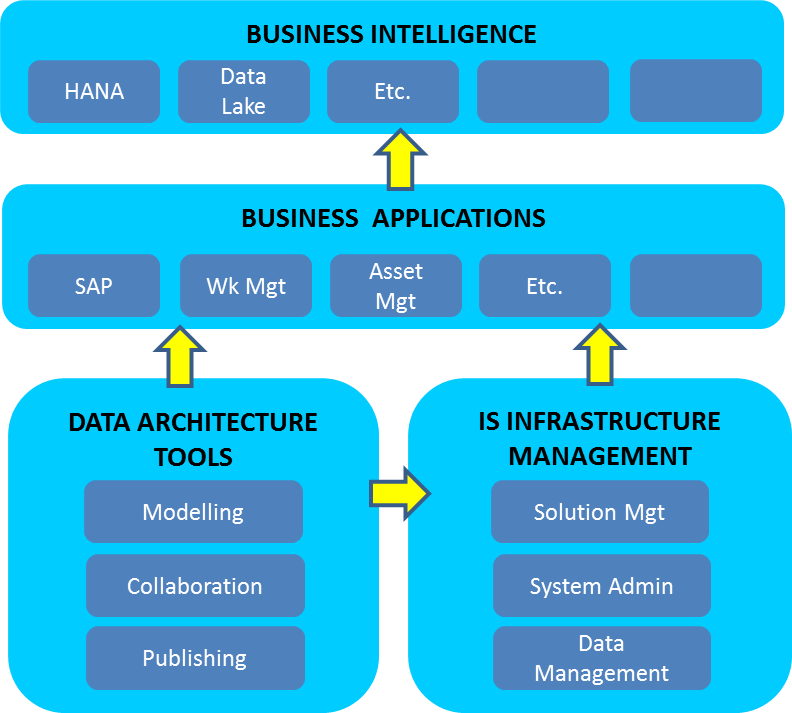
Reference data implemented within systems is recorded in the system specific Data Dictionaries.

Master Data Architecture

A Master Data strategy defines the main applications and interfaces that will manage the Master Data (including Reference Data), the metadata required to design applications as well as the technical control data needed to correctly configure and manage business applications.

The following diagram depicts these main areas of the Master Data architecture.

**Master Data Architecture Landscape**



The business intelligence applications are one of the main final recipients of the Master Data generated and managed by the supporting components. The quality of the reporting generated is fully dependent on the quality of the Master Data.

The business applications support the actual business processes and store the Master Data as well as the transaction data.

The data architecture tools support the design of the data models and process flows which are the basis for the business application design as well as the governance and management of data.

The IS infrastructure management handles the development, deployment and maintenance of business applications. This area is not really in the scope of Master Data Architecture but is mentioned in this document at a high level for completeness.

These four components should satisfy both the differing user requirements as well as the ability to interact with one another seamlessly. This results in a balance that needs to be achieved between “best of breed” on the one hand and a high level of standardisation on the other hand.

By developing enterprise wide standards for each of these areas and implementing them in a systematic step by step approach, National Grid could steadily converge to an application landscape which is more cost effective to roll out and maintain and which provides the business with a high degree of consistency in the operational and strategic information managed in those systems.

There are three ways that Master Data Management can be implemented:

1. To manage the Master Data manually across all systems and processes that use it. This is significantly cheaper and easier to implement. But in the long term is inflexible and requires significant resources and manual processes to maintain. It is much more liable to let data quality issues through.
2. To manage the Master Data centrally via a Master Data System (MDS). This can be expensive to implement upfront, but would, in the long term, have lower resource costs and data quality issues. An MDS can be set up locally, globally or somewhere in between; within or across data domains.
3. The 3rd option would be a hybrid of the first two and dependent on what was possible and practical.

The first point above hasn’t been covered in this guideline as that would be down to specific, local governance – providing the principles of MDM from section 4 & 5 were followed. Therefore, the rest of this section covers the second point.

A Master Data architecture would provide National Grid with the tools and approach consisting of:

* An enterprise Master Data management system (EMDS) for enterprise wide data standards
* A standard toolset for data governance, data management, design, modelling, publishing and collaboration
* The first set of common data elements which will start to drive standardisation in reporting and in application configurations

This type of vision will not be achieved in a big bang approach in which all data, governance and application aspects are re-designed and re-implemented in one effort. The approach follows the following principles:

1. Whenever new applications and data work is performed, it should be ensured that the MDM work is future proof and scalable where appropriate. i.e. it fits into the global architecture and data model and is covered by the governance model
2. Ensuring that the overall architecture is initially designed as scalable doesn’t mean all components must immediately be built and would usually be done incrementally.
3. Develop the global standards in a top down approach, defining and standardising the most important elements first
4. Ensure that the architecture can cater for the integration of applications which are not (yet) compliant with the global standards. This is generally achieved by providing the tools to manage the necessary mappings for the non-compliant elements in those applications. This flexibility is key to the ability to quickly integrate acquisitions at least at a reporting level

## Business Intelligence / Reporting

Good MDM practices are key to being able to provide high quality, reliable and universally understood Business Intelligence (BI). A BI tool will need the following:

* To know which the authoritative source is to take the data from. (full Master Data lineage should be known).
* To know that the data has not been updated in a different place.
* If data is to be combined from multiple sources, then it will need to be formatted the same and contain common keys to link them.
* Clear definitions and understanding should be held for every data field (particularly the reference data) used in a report so users have absolute clarity of what they are looking at and fields cannot be misinterpreted.
* No Master Data should ever be updated in a report. This renders it untrustworthy even if it now appears to tell the right story. The Master Data must always be updated at source and the report re-run if errors are found.
* Master Data shouldn’t be hard coded in to the report if possible, but fields extracted, and filters applied as needed. Otherwise as soon as the Master Data is updated the report won’t work as expected. Data may be excluded which will likely be undetectable.

## Business Applications

The following is a list of the key requirements that the business applications have of a technology-based Master Data management system:

* Provide cost effective and efficient operational Master Data maintenance functionality at a global, local, entity specific and/or cross entity level.
* Perform the maximum level of validation at the point of first entry, where the original information is available.
* Create a support and management process to manage the data governance and provide the business with the data tools to manage data consistency, quality and completeness.
* Provide Master Data consistency across all main business application platforms.
* Provide advanced data cleansing, de-duplication and syndication tools; taking into account cost and return, also volume of data.
* Provide tools through which subscription services for data enrichment can be centrally managed and integrated.
* Support a shared service centre approach in which back office functions for multiple lines of business are provided as a central service.

## Data Architecture Tools

The data architecture tools are tools that will be used by the data governance organisation (people with a special focus on data quality and data maintenance) as well as by the information systems organisation which designs and manages the applications by which this information is stored and used.

The following is a list of the key requirements that the data architecture tools have of a technology-based Master Data management system:

* Provide tools to support build and manage a complete logical view of the data elements required for an optimal execution of reporting processes.
* Provide tools to support, document and manage the way in which data elements are implemented in multiple information systems.
* Consolidate access to all metadata through one central repository.
* Leverage the existing metadata in standard software applications.
* Support change control and versioning of metadata.
* Provide tight integration between the process models and the (master) data models.
* Provide tight integration between the business definitions and the corresponding IT representations (table/column definitions).

A common set of metadata is a key pre-requisite to any attempt to standardise and control the Master Data used in National Grid. The other deliverables would be:

* Build and manage a catalogue of the logical data structures as well as the many physical implementations in both the legacy and the future environments.
* Consolidate all metadata for easy access in one place.
* Leverage the existing metadata in standard software applications.
* Support change control and versioning on metadata.
* Provide tight integration between the process models and the (master) data models.

## Infrastructure Management

The scope of this section is to highlight the important Master Data considerations that should be incorporated into an Infrastructure Management Strategy. The IT strategy should:

* Provide tools to support change control and versioning of configuration data.
* Support a separation of development from test and production environment.
* Strictly control model and software changes to productive environments.
* Minimise the number of different technologies used where possible.
* Choose components which support service-oriented architectures. Automated components should be a key factor, to drive timely replication.
* Provide controlled identity management.
* Support change control and versioning on Master Data.
* Support a separation of development from test and production.
* Strictly control changes to productive environments.
* Standardise on one standard technology wherever possible.
* Choose components which are based on the same standards and platforms, aligned with the IS platform strategy.
* Prevent data loss.
* Keep operational systems lean for performance and flexibility reasons.
* Manage the Master Data such that data that is no longer required or out of date is no longer available.
* Standardise message transfers and make them centrally manageable.
* Provide for easy access (Single sign-on) and easy identity management.

## Master Data Sharing Architecture

The Master Data Management (MDM) hub is a database with the software to manage the Master Data that is stored in the database and keep it synchronised with the transactional systems that use the Master Data. There are many different types of architectures where the answer to managing MDM is a system and it is important to identify which one is the most appropriate, however physical ‘hub’ architectures are the most commonly used.

There are 4 types of ‘hub’ architectures[[1]](#footnote-1)– the key difference between the four hub types is the number of fields that are stored centrally:

**A repository hub stores all fields**

The complete collection of Master Data for an enterprise is stored in a single database, including all the attributes required by all the applications that use the Master Data.  The applications that consume, create, or maintain Master Data are all modified to use the Master Data in the hub, instead of the Master Data previously maintained in the application database, making the Master Data hub the system of entry as well as the system of record.

There are no synchronisation or latency issues with updates getting propagated to multiple copies of the Master Data.  There are no update conflicts caused by updates to more than one copy of the Master Data.  In general, a single copy of the Master Data is significantly easier to manage and will generally be of a higher quality than multiple copies with all the potential synchronisation and mapping issues. However, deciding on a common data model for all applications will be a difficult task both politically and technically. You will also need to transform and load all the current databases into the hub, removing duplicates in the process. And you will need to figure out how to handle history – you are changing your databases to use a new key for all your Master Data so in many cases you will need to create the same kind of key mapping that the other two MDM styles require to be able to access history records.

**A Registry Hub (also called Federated) only stores the identifiers**

This is the opposite of the repository approach, as each source system remains in control of its own data and remains the system of entry, so none of the Master Data records are stored in the MDM hub.  All source system data records are mapped in the Master Data registry, making the Master Data registry the system of record (a virtual Master Data system).  Data maps show the relationship between the keys of the different source systems (i.e. one row in a table for each Master Data entity and columns for the keys of the application systems).  For example, if there are records for a particular customer in the CRM, Purchasing, and Customer Service databases, the MDM hub would contain a mapping of the keys for these three records to a common key.  The benefit being that because each application maintains its own data, the changes to application code to implement this model are usually minimal, and current application users generally do not need to be aware of the MDM system.  The downside is that every query against MDM data is a distributed query across all the entries for the desired data in all the application databases. Plus, adding an application to the MDM hub means adding columns to the key-matching table, which is not a big issue, but it may also mean changing queries to include the new source of information.  Finally, while it helps you find duplicates, it does not help you in cleaning them up (i.e. if a person has many records with different phone numbers, there is not a way to determine the one to use).

**A hybrid hub stores the core fields**

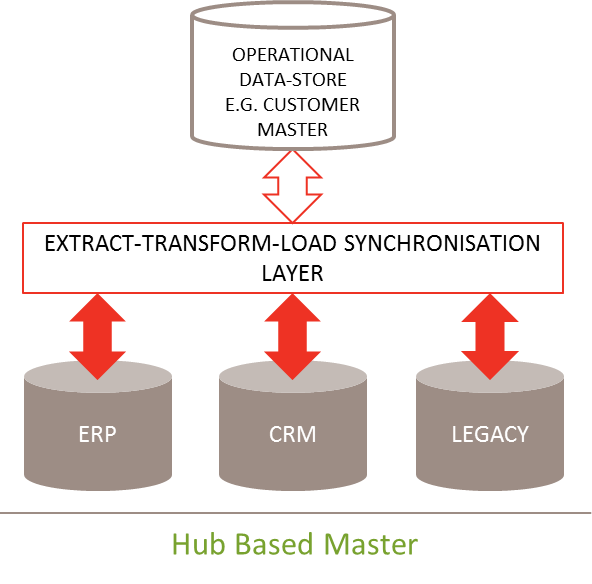
A hybrid includes features of both the repository and registry models, It leaves the Master Data records in the application databases and maintains keys in the MDM hub, as the registry model does. But it also replicates the most important attributes for each master entity in the MDM hub, so that a significant number of MDM queries can be satisfied directly from the hub database, and only queries that reference less-common attributes have to reference the application database. Thus the advantage of the Hybrid approach is that it provides a single, authoritative source for shared Master Data without the necessity of changing all your applications to use it. The most significant disadvantage of the Hybrid style is that keeping the MDM hub copy of the data synchronised with all the source systems can be a complex process.  If you allow all the source systems to change Master Data, you will have a continuous data integration problem caused by incompatible changes coming from different systems.  You can reduce this problem by requiring changes to the Master Data to be made only to the copy in the MDM hub but this may be difficult to implement and enforce.  Also, MDM synchronisation is more complex than data replication because the data may have to be transformed both when it is loading into the hub and when it is sent from the MDM hub back to the source systems because the data models of the source systems may all be different.

**A virtualised hub doesn’t store any fields**

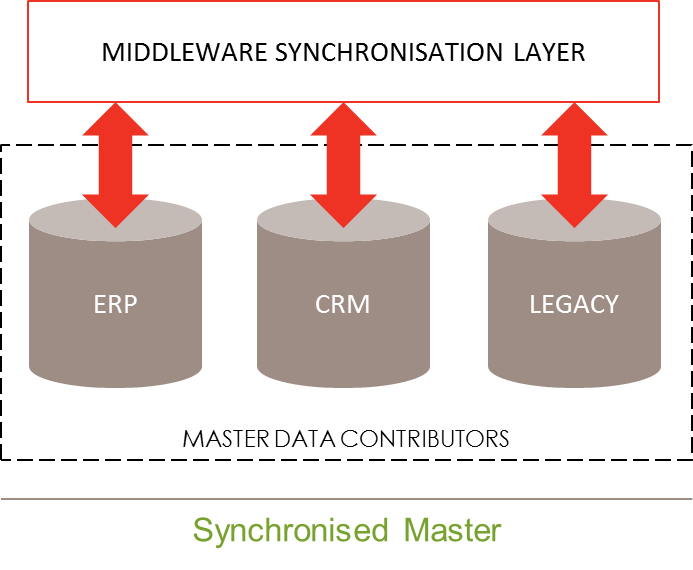
This relies on processes and governance to maintain synchronisation. This is obviously much easier and quicker to implement with very little overhead. However, it doesn’t provide any systematisation for maintaining synchronicity across data sets. The weak points of this system will be in the processes and people at each step. Unlike the other systems it has a low initial outlay, but the BAU costs could be significant.

Below are some typical system architectures for possible MDM solutions:

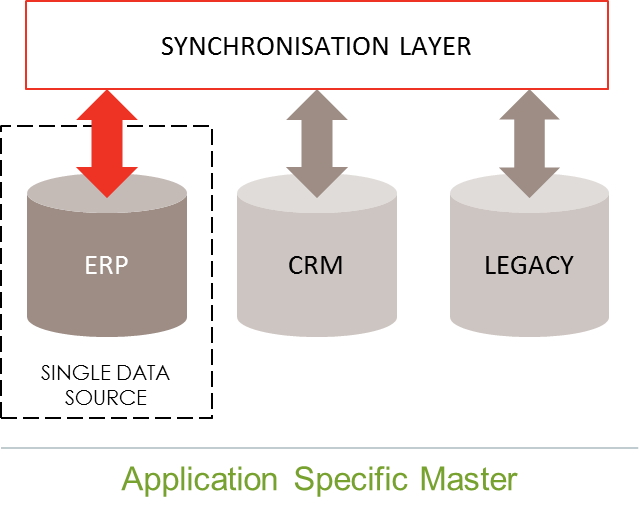
A **Hub structure** overlaps operational and analytical environments. It supports the concept of an Enterprise Data Warehouse. It compromises multiple systems acting as data providers and is appropriate for low data latency and velocity operations.



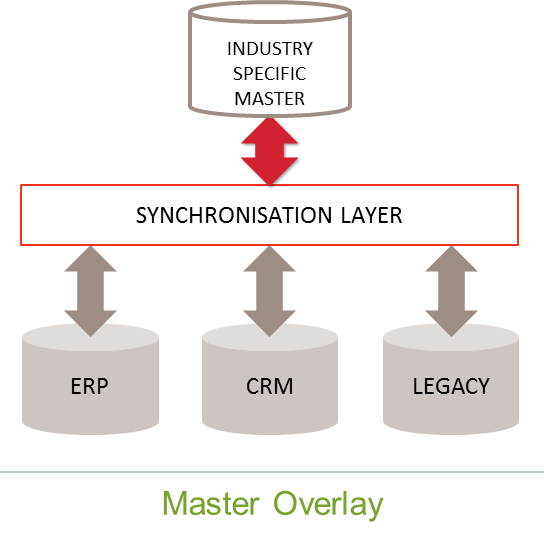
**Synchronised Master** structure comprises multiple operational systems acting as Master Data contributors with real-time information availability. It is well suited to enterprises where data is stored across multiple source systems and to low data velocity operations.



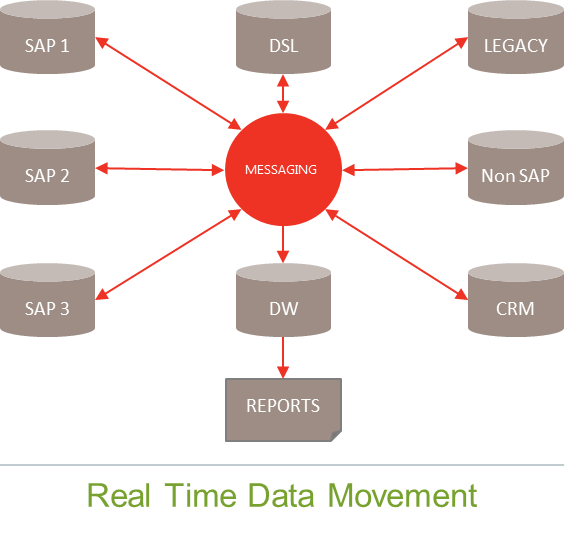
**Application Specific Master** is an operational system as Master Data provider. It is well suited to enterprises where data is primarily stored in a single source system. It can support many enterprise vendors.



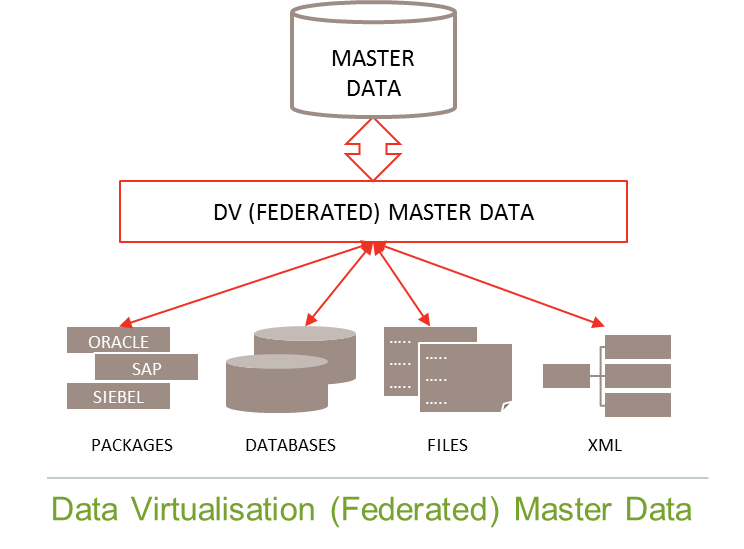
A **Master Overlay architecture** is stand-alone with application-neutral Master Data and industry-specific data models. It is well suited to vertical industries in aligning front-office and back-office systems in real time.



**Real Time Data Movement** is where data is moved between the various systems using a messaging integration hub or ESB. Data movement is done in real time and can be one or two way as required.



**Data Virtualisation (Federated)** Master Data uses a virtual MDM hub created via Data Virtualisation (DV) layer. The DV layer is informed from the Logical Data Model. Powerful data integration is used to access multiple disparate systems. This has the advantage that it is rapid to implement.



Master Data Design for Systems

When designing data structures within a single system such as an ERP (enterprise resource planning) or an EAM (enterprise asset management) system creating the right Master Data relationships is critical to is future success and longevity.

For example if a data structure is created that fully meets today’s requirements using a tactical rather than a strategic solution design, it will provide a quick, cheap solution that will completely meet the requirements. However, as soon as there is any change to the requirements (could be driven any number of factors such as: legislation, business process change, technology change etc), changing the solution, or creating a new one, becomes significantly more difficult. For example, instead of just creating a solution for the new requirements, it may have to encompass changes to the original one. Or, it may mean that the new solution has to be much more complex to accommodate the original one. Thus a tactical solution will inevitable drive further tactical solutions and thus compound multiple, manual work arounds.

Therefore, when new data structures are required, they must always be considered strategically and holistically.

Below is a set of considerations or principles to employ:

## Consistency

The Master Data should be structured consistently. Underlying all the other principles a consistent, logical structure is required for Master Data Entities in order to:

* facilitate the improvement of data quality;
* allow for the identification of business ownership (combinations and/or separation);
* identify performance improvement opportunities and feedback to modify and refine policies;
* allow for the provision of accurate and meaningful management information and accurate regulatory reporting consistently across and within all Business Entities
* provide a single source of information to achieve like-for-like comparison and trustworthy management information;

## Future Proof

In order for Master Data structures to meet the requirements of a diverse and changing business they should be independent of and therefore “Future Proof” to changes in:

* + Organisational structure,
  + External (eg legislation/regulatory) driven requirements
  + Geographical area and
  + Policy/Process changes.

The design should be aware that there is always an inherent danger in adopting a solution created for one purpose for a different one. It may work initially, but over time requirements may diverge to the point where work arounds are needed to bring them back into line. There is also an onus on users always being aware of both needs when reviewing or updating either the data or functionality.

The design should be able to accommodate National Grid’s diverse requirements (including those that are current, those that are expected and those that are unexpected) yet have rigid rules to prevent easy data errors. It also cannot be so complicated that the resources required to maintain it outweigh the benefits of holding the information

Therefore, the guiding rule behind a Master Data structure is that Master Data entities are grouped according to their primary function and relationship with other Master Data entities to support their primary purpose. If this is done logically and data is stored at its lowest common denominator, then it can be cut and diced in any way the business requires it in the future. For example: If the regulator requires data entities to be reported in certain groups, it might seem easy to put those groups as an attribute. However, as soon as the regulator changes its requirements, all that data and its history will need to be changed. If however the information is derived from real attributes of the entities in the BI, then it can be re-derived in the report whenever the requirements change. This can also be back dated just as easily to provide history and both reports can be run together to provide comparisons.

## Asset Management specific requirements

This section has been added as National Grid is an asset intensive asset management company. As a result, we have signed up to and specified an Asset Management Standard as part of the BMS. Good Master Data structures and Master Data management specifically around our asset data (both fixed/financial assets and the physical ones) is absolutely critical to achieving this.

The fundamental requirement of asset management is to be able to access Whole Asset Lifecycle Costs. i.e. for each Physical Asset we must be able to track its cost history and relate this to the Financial Asset so that the Whole Lifecycle Costs can be tracked, including installation costs, time & materials, maintenance & inspection, depreciation and ultimately refurbishment, replacement and disposal.

If this is done in a single ERP system, then the Master Data links between fixed and physical asset are already provided. However, if this is done via an MDM system or a reporting tool such as a data lake then alignment of Master Data and reference data across systems is essential.

The alignment of asset Master Data is usually done in an Asset Hierarchy containing a parent/child relationship. Although this example is referring to an asset hierarchy, other Master Data hierarchies for non-asset data entities can and do exist. Many of these principles will still apply.

In summary, an Asset Hierarchy is primarily used to:

* locate where a particular item of equipment is to be found
* track its maintenance, fault/emergency and cost history;
* to define its purpose and its relationship to other items of equipment.

It also supports the ability to:

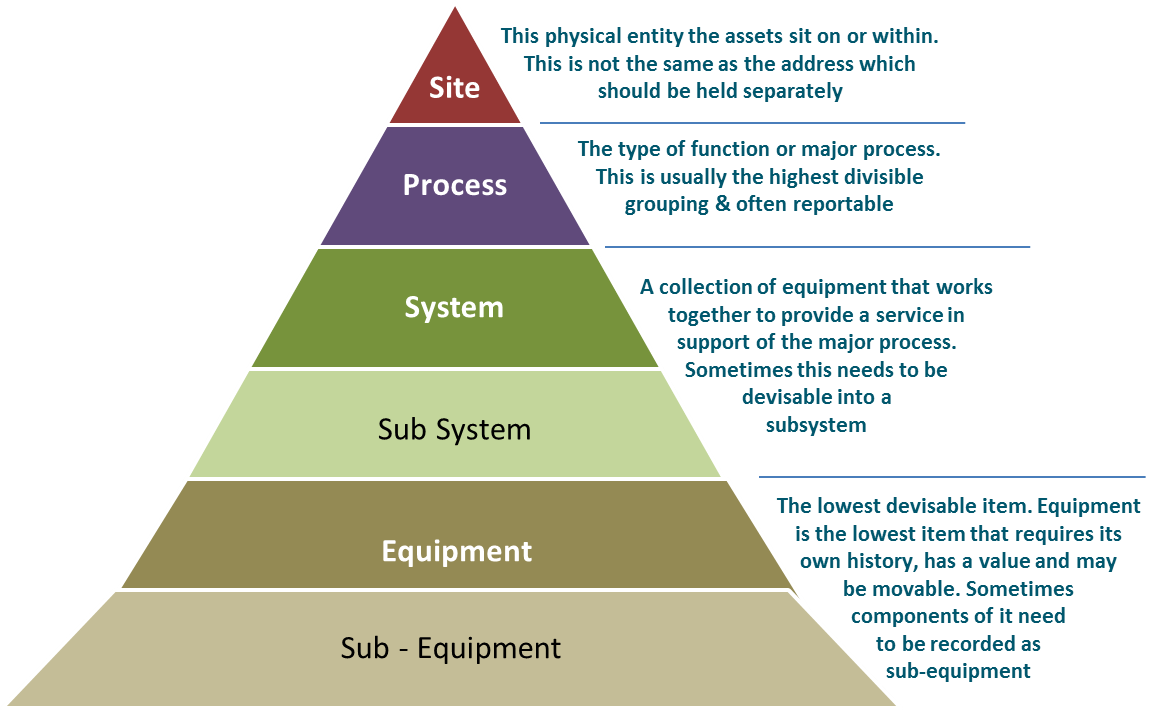
* Enable asset life cycle management using complete and full information feeding on risk and condition to support predictions and innovations.
* Enable the tracking of maintenance costs (OPEX), and costs of new installations (CAPEX) for better investment / asset life decisions and deferred capital expenditure.

Asset Hierarchies should not include levels in the hierarchy relating to Geographies, Organisational Structures, or Business Processes. Most ERP systems have other functionality that is much more suitable for all these demarcations. By appropriate configuration of the system organisation structures and classifications/attributes the reports, changes and searches can be handled with greater efficiency than by building this into the hierarchy

Therefore, the guiding rule behind an asset hierarchy structure is that the assets are grouped according to their primary function and relationship with other assets on the physical site to support the purpose(s) of that site. And by doing so they support the principle of “Future Proof” i.e. they are not subject to regular changes. Work should be targeted at levels and asset types in the hierarchy in a consistent manner where policy, regulation and safety allows it i.e. the same asset type should be maintained at the same hierarchy level, where possible and appropriate, so that scheduling, completion and reporting compliance information can be measured consistently. Where an asset is the same across processes or business entity, then it should be identified as such in the classification data.

It is important to keep levels the same across all the Master Data as this allows for comparative reporting across all lines of business and one set of standard data tables. Although flexibility is required some rules need to be applied to ensure the right levels of consistency are maintained.

Below is a sample of good practice asset hierarchy structure. It is slightly different from the one often quoted from ISO 14224[[2]](#footnote-2) as that was created for a different purpose and a different type of business.



## Information Storage & Retrieval

The Master Data structure must support the requirement to provide meaningful Management Information/Business Intelligence. Information stored about a Master Data entity should be related at all times to that entity.

## A Well Governed Design

Thought needs to be given at the design stage of a new system to how the Master Data structure and data quality could be maintained post implementation. The system configuration needs to take into account the known requirements, the expected requirements and even be able to accommodate new unknown requirements without major redesign.

The functionality and possible enhancements/configurations of an ERP will dictate how the configuration can help guide standards in data quality. However, in order to improve and maintain data quality and system integrity robust governance procedures must also be in place, managed and adhered to.

For example:

* Clear and detailed design documents for reference. Governance would need to ensure the integrity of the structure was maintained and monitored
* Training documentation should not deviate from the design or offer flexibility that encourages errors
* A clear process for system governance and change control must be in place from the point the design is agreed and throughout Blueprint, Detailed Design, Migration, Implementation and thereafter.
* The post Transformation business will need to take into account how team structure can directly affect data quality. For example: separate system administration teams in each Line of Business (where systems are shared) would allow for deviation from one set of standard data tables and put at risk the Cross Lines of Business reporting.
* Very rigid control processes need to be in place when granting access to the system; to what level that access is granted; and how access is removed.
* Governance would need to define who are the most appropriate groups of people to have update access, and how extensive should that access be.

Useful Resources

## External Reference Documents

* DAMA Dictionary of Data Management - 2nd Edition 2011
* DAMA-DMBOK (Data Management Body of Knowledge) - First Edition 2010
* ISO 8000 – The global standard for Data Quality and Enterprise Master Data
* ISO 5500 – The international standard covering management of physical assets

Document Administration

|  |  |
| --- | --- |
| Owner | Dan Robertson |
| Author/s first edition | J Harrison |
| Author/s second edition | Katie Brunton/Lisa Russell |
| Author/s current edition | Linda Zimba |
| Date first created | 14 November 2016 |
| Date of second edition | 30 November 2018 |
| Date of current edition | 1 July 2021 |
| Date last amended | 1 July 2021 |

## Reviewers

|  |  |  |
| --- | --- | --- |
| Reviewer’s Name | Role | Date & version |
|  |  |  |
| Daniel Senter | Head of Data Governance | 21 July 2021 |
| Dan Robertson | Group Head of Data & Analytics | 03 Aug 2021 v3.0 |
| Nishit Ajwaliya | Director, Information Architecture |  |
|  |  |  |

## Change History

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author | Description of Changes |
|  |  |  |  |
| V2.0 | 30 Nov 2018 | Katie Brunton  Lisa Russell | Additions inserted from US comments |
| V3.0 | 1 July 2021 | Dan Senter / Linda Zimba | New branding / template and removal / update of out of date content |
|  |  |  |  |
|  |  |  |  |

Useful Resources

National Grid plc

National Grid House,

Warwick Technology Park,

Gallows Hill, Warwick.

CV34 6DA United Kingdom

Registered in England and Wales

No. 4031152

nationalgrid.com

1. Definitions of each taken from James Serra’s and Roger Walter’s blogs. James and Roger are solution architects at Microsoft [↑](#footnote-ref-1)
2. ISO 14224 – The international Standard for Petroleum and natural gas industries - Collection and exchange of reliability and maintenance data for equipment. [↑](#footnote-ref-2)