THE EDGE

Master Data Management

Specialised for the pharmaceutical industry

Master data management

Specialised for the pharmaceutical industry

This paper describes the design of a central Catalogue system used as a method of managing all the master data used across an organization. The benefits within an overall Drug Discovery IT architecture are highlighted with an emphasise on the positive impact such a system can deliver to the enterprise. An example implementation is discussed demonstrating the feasibility of the approach.

Introduction

As any organization grows, maintaining an accurate working knowledge of its assets becomes increasingly difficult. This is often manifest by an increasing diversification of systems and a lack of consistency in the use of so called *master data*. Master data is defined as sets of data that are synchronized copies of core business entities used in transactional or analytical applications across the organization, and subjected to enterprise governance policies, along with their associated metadata, attributes, definitions, roles, connections and taxonomies.[1] Inconsistency in the master data often leads to an increase in the manual effort required by data managers attempting to maintain data standards. The central catalogue is designed to remove such problems, by maintaining consistency and reducing the cost of future enhancements to the informational infrastructure. In short the *Catalogue* should be a foundational concept within the information management strategy of any organisation ensuring that quality is controlled at source. The *Catalogue* is used to manage not only the system entities themselves but also their applications. In essence the *Catalogue* is a way to organize and centralise the corporate knowledge contained in the master data.

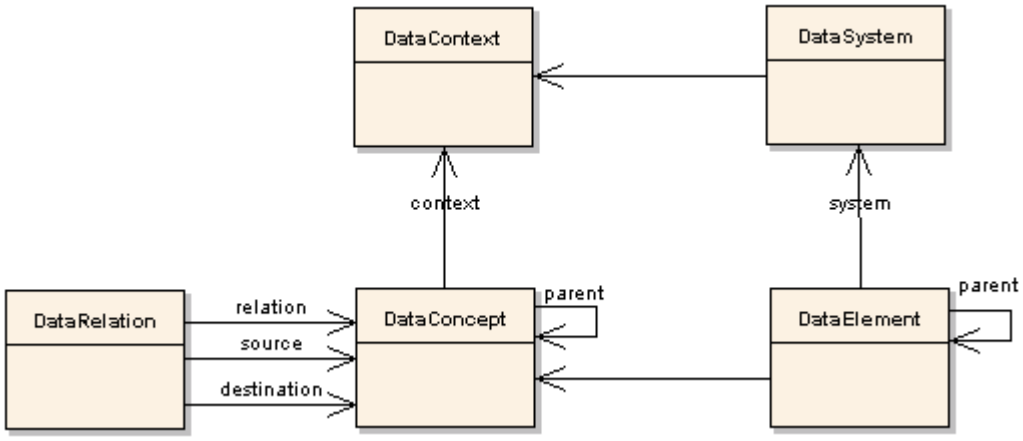
Catalogue Design

The *Catalogue* contains a logical organization made up of a number of data *Contexts* which are in turn split into a number of trees of *Concepts*. This can be understood as a classical conceptual graph, similar to those seen in a number of AI systems and more recently in the semantic web. This is based on a simplified version of ISO/IEC 11179[2], an international standard for representation of metadata from an organisation in a metadata registry.

These *Concepts* are then linked to a number of physical data *Systems* through data *Elements* which actually implement the dictionary lookups. In this way the *Catalogue* keeps a logical separation between concept and realisation, providing a very flexible way of representing a system.

[1] David Loshin, Business intelligence network 2006

[2] Wikipedia http://en.wikipedia.org/wiki/11179



Put more simply, A *Catalogue* organises similar dictionaries in a tree of *Concepts* and provides a single overview for management of all the integrated system dictionaries. The management of an individual dictionary may still be the responsibility of another system, but the dictionary entries it creates can be made available to all applications linked to the central catalogue.

The benefits of having a central *Catalogue* of dictionaries go beyond providing a convenient source of dictionary look-ups. The *Catalogue* provides a consolidation point for the corporate knowledge contained within the master data used across an organisation. This has the benefit of separating the master data from the mechanics of data capture, removing the dependency between data capture systems and the structure of information owned by the organisation. Centralisation of all master data enables a consistent and aligned set of terms and concepts to be used across all facets of the organisation. Data quality is improved by realisation of a consistent set of dictionaries at the point of capture, rather than attempting to map divergent dictionaries together during data consolidation in for example a data warehouse. This delivers great benefits to the overall data architecture and life cycle. Liberating the master data from the data capture systems is yet another architectural principal which allows an organisation to be in more control of its data and the architecture of its information systems.

In this paper we describe the key features of a Catalogue and review a real world implementation of a Catalogue as a case study.

Key Terms

The following section describes the key terms used in this paper.

Catalogue

A *Catalogue* is an organised collection of data look-ups or dictionaries that provide a consolidation point for corporate master data.

Data Context

A *Catalogue* should support a number of contextual name spaces, or *Contexts*. A *Context* is a classification of concepts. Each *Context* contains a hierarchy of *Concepts* or groups of similar dictionaries. There is often a separate *Context* for each domain serviced by the Catalogue, for example research, development and clinical. Different *Contexts* may represent the different disciplines involved in research, such as chemistry and biology.

Data Concept

A Data Concept is used to classify similar dictionaries. Typical *Concept*s include Receptors, Cell Lines, Species, Compounds and Samples. Data concepts help to provide a natural data categorisation introducing structure to the master data.

Data Concept Specialisation

Concepts can be arranged in a hierarchy with each child *Concept* considered a specialisation of its parent. In this way, the top level concept is described as the *Context*. An example of a high level *Concept* with child concepts is the relationship between the abstract *Concept* Substance and its concrete specialisations Compound, Peptide, and Protein. These specialisations, or child *Concepts*, are all types of Substance. Building this structure into the data helps to enforce data consistency and adds value to the master data ensuring it can be suitably exploited within analytical applications.

Data Element

Data dictionaries are realised as *Data Elements* associated with a *Concept*. There can be multiple dictionaries entries or *Elements* within a *Concept*. For example, the *Concept* Species may have look-ups for Rodents, Apes, and Common, with the dictionary entries being stored in different *Data Systems*. This separation between concept and realisation provides a great deal of flexibility and allows logically consistent concepts to be realised from heterogeneous data systems.

Data System

Data Systems provide the source of data for the *Data Elements* or dictionaries contained within the *Catalogue* of *Concepts*. The *Catalogue* should be able to interface with different types of system and technology. *Data systems* are typically a variety of databases, but can also include heterogeneous systems accessed using web services. The *Catalogue* does not necessarily need to implement the dictionaries but can be used to link specialist systems designed specifically for this purpose. Obvious examples include the compound concept that is typically realised within a specialist chemical registration system. In essence this forms a look-up of simple textual compound identifiers. The business logic used to generate that list of identifiers is complex and realised externally from the *Catalogue*.

Design considerations

In this chapter we shall discuss some of the design considerations which should be taken into account before embarking on a *Catalogue* project.

Consolidation

Before deploying a catalogue into productive use, it is important to decommission any redundant dictionary registration systems. For example, if two biological data management applications both provide a Species dictionary, the ability to register a new Species should be restricted to a single application. This may be achieved in one of the existing data management systems, or better realised within the catalogue itself. Isolation of master data management reduces the dependency between data capture applications and the corporate master data, facilitating more architectural choice. The existing data management systems used for transactional data capture must then make use of the Species concept (or look-up) directly from the central catalogue. Future changes to the master data are then immediately consistent as all applications source data from a single source often referred to as a truth source.

Conceptual Dissonance

Conceptual dissonance describes the state in which there is a conflict between two uses of a term within different contexts (often disciplines) within an organisation. This can relate to different terms which share the same meaning or the same term which has different meanings depending on the context in which it is used. This is a common issue especially within the highly technical and educated environment of pharmaceutical research and development. To counter this phenomenon any good Catalogue must offer some level of aliasing to avoid this conflict. For example the term dose is typically used in screening to refer to a concentration measured in µM, however during *in vivo* studies dose almost always refers to the dose of compound related to the subjects body weight and is measured in mg/kg. Another example are the terms Receptor and Target, two terms with equivalent meaning.

Aliasing provides a method of managing these terms without constraining the different scientific disciplines to artificial and unfamiliar terms. This often overcomes one of the key objections to introduction of good master data management practises, negotiation of terminology within large heterogeneous organisations.

Centralisation

Centralisation is an important technique used to gain control of the master data within an organisation. Consolidation of master data helps to present a single consistent viewpoint, but centralisation provides a guaranteed single system which effectively controls definition and implementation of the master data. The motivation to effectively manage master data is often most evident in those responsible for data analysis (or data mining). These activities are most severely impacted by a poor master data strategy.

A common approach used by a number of pharmaceutical companies is to centralise the definition of Assays, centralising the definition of parameters and their inter-relationships for biological testing. In practice data capture systems may be required to extend the definition of the Assay with quality parameters to facilitate data capture, however, the centralisation of Assay definitions is often used to define the core data captured during biological research. Typically these systems are document based and miss the opportunity facilitate automatic translation of the assay definition into a formal system for capture reducing their impact in minimising the effort associated with data management.

Minimise Transformation

The catalogue is used to realise a master data strategy controlling definition of dictionaries used in transactional data capture systems. A consistent master data strategy ensures that the dimensionality of data is enforced at the point of capture ensuring that data can be effectively exploited by analytical systems. This in turn simplifies the data life cycle from transactional systems into long term storage in a data warehouse. A well implemented catalogue will reduce the need for complex data transformations and reduce the complexity of data loading processes. Traditionally warehouse projects are often over complicated as a result of the lack of master data management leading to elaborate data transformation requirements to introduce consistency after capture. The positive benefit of so called “Master Data Management” systems on data warehouse projects has been anticipated by the industry for sometime[3]. Industry opinion suggests that data quality issues should be fixed at the source rather during consolidation.

---

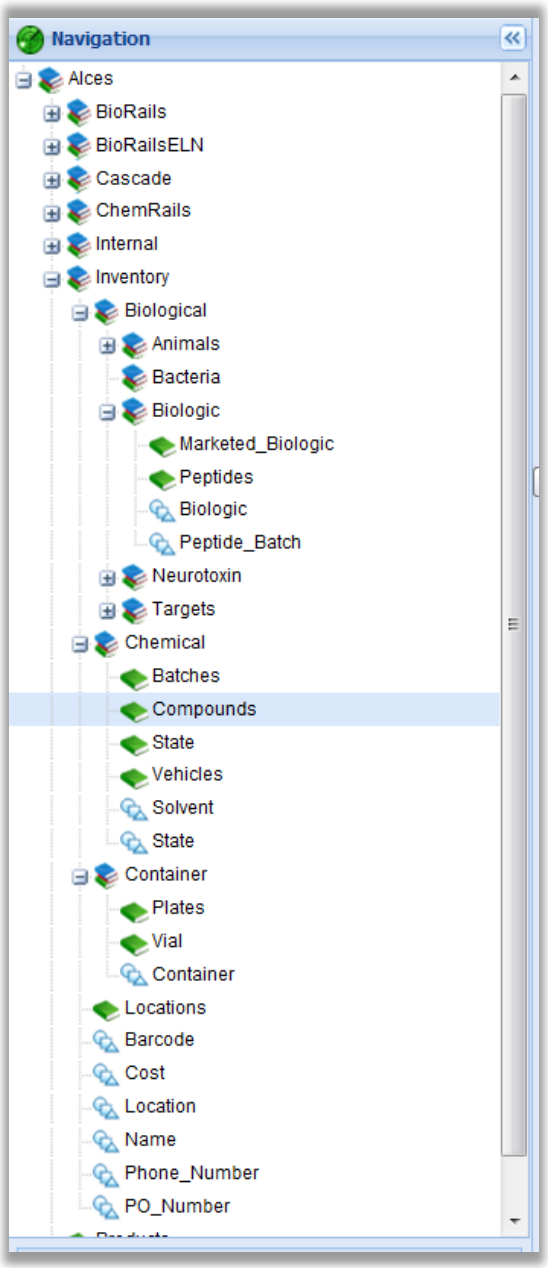
[3] The Data Warehouse Lifecycle Toolkit by Ralph Kimball et al

Case Study

The following section discusses a real world implementation of the *Catalogue*, demonstrating its feasibility.

BioRails Catalogue

BioRails is a biological data management platform developed by The Edge. Its design was based on extensive experience of managing these issues within the pharmaceutical research environment. One of the key design elements of BioRails is the Catalogue which forms a fundamental component of its Administration module. The Catalogue is used to define all dictionaries used within the system. Dictionaries are organised into a number of Concepts and realised as data elements implemented either internally within BioRails itself or externally within other data systems. The primary role of the Catalogue is as a source of dictionary based parameters for the capture of structured data during biological research. The BioRails Catalogue is used as the basis for the following case study



Context, Concepts and Implementations

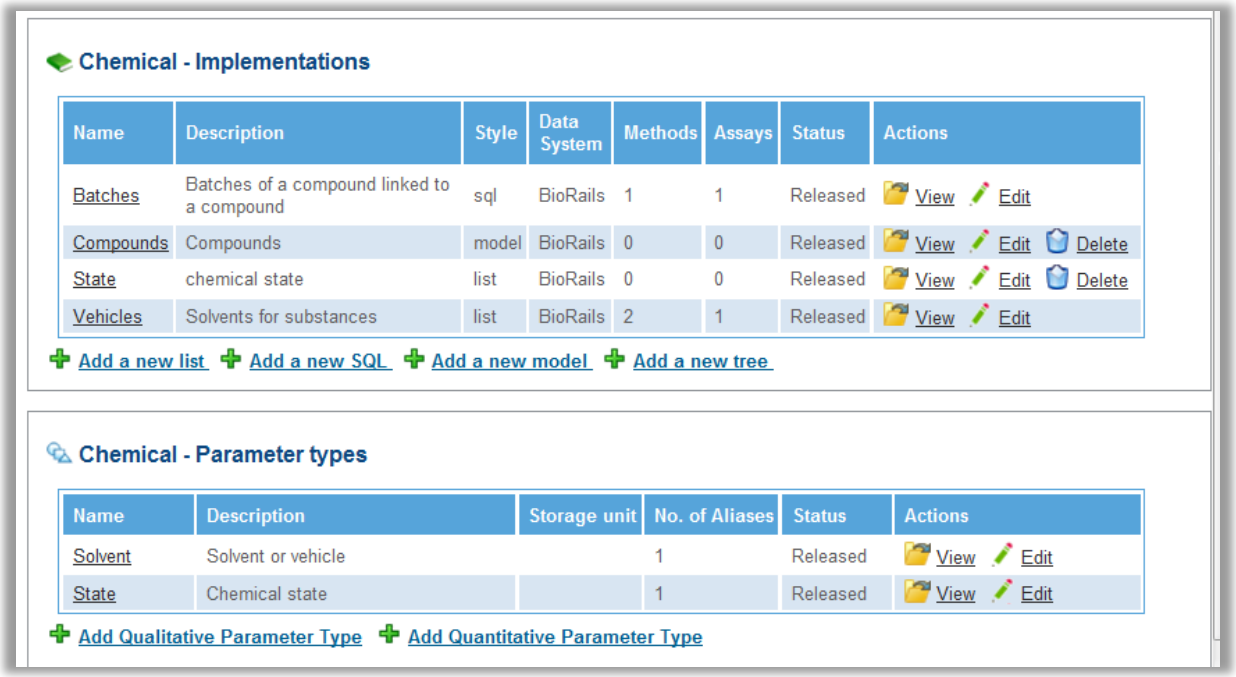
The hierarchy of concepts from the root context to the realised dictionary are visualised using the “dictionaries tree”. Parent concepts are displayed as folders within the tree illustrating that they contain child concepts or specialisations. For example the concept *Compounds* is shown as a child of the parent concept *Chemical*. The *Compounds* concept is realised as an implementation on the 'leaf' of the tree. In turn the *Chemical* concept is a specialisation of the concept *Inventory*.

The context *Inventory* contains other catergories of inventory concepts such as *Biological*, *Chemical* and *Container*. Biological inventory concepts are futher categorised into concepts such as *Animals*, *Bateria*, *Biologic*, *Neurotoxin* and *Targets*.

The dictionaries tree can be used to explore all the concepts available within BioRails.

Data Elements

The *Chemical* concept and has a number of Data Elements (implementations of the concept) associated with it.



Chemical - Implementations

Name	Description	Style	Data System	Methods	Assays	Status	Actions
Batches	Batches of a compound linked to a compound	sql	BioRails	1	1	Released	View Edit
Compounds	Compounds	model	BioRails	0	0	Released	View Edit Delete
State	chemical state	list	BioRails	0	0	Released	View Edit Delete
Vehicles	Solvents for substances	list	BioRails	2	1	Released	View Edit

Add a new list Add a new SQL Add a new model Add a new tree

Chemical - Parameter types

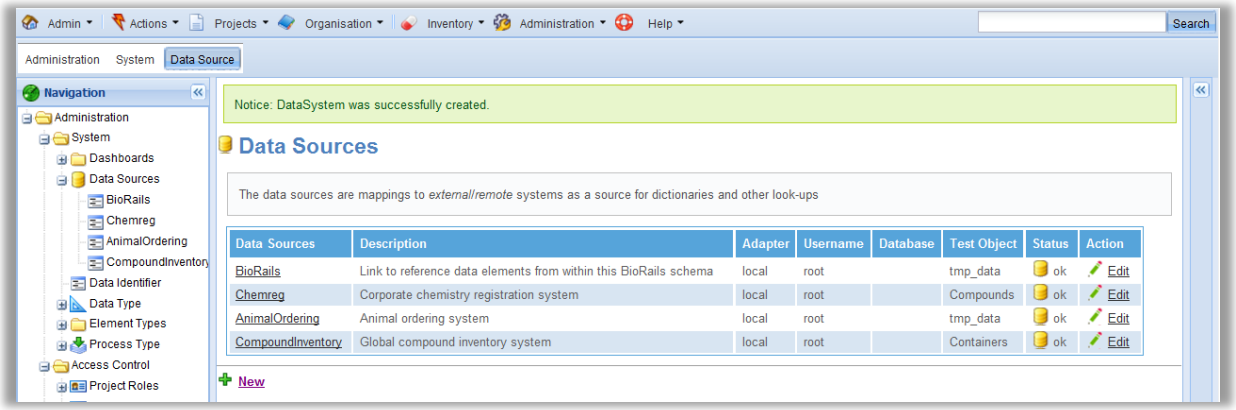
Name	Description	Storage unit	No. of Aliases	Status	Actions
Solvent	Solvent or vehicle		1	Released	View Edit
State	Chemical state		1	Released	View Edit

Add Qualitative Parameter Type Add Quantitative Parameter Type

In the example above there are four implementations of the *Chemicals* concept. The first, *Batches* represents batches (units of manufacture) of the compound. This concept is realised externally to BioRails and linked through an SQL statement from an external database. The implementation of the concept *Compounds* is a lookup mapped using a BioRails model to an external system. The BioRails model connects the simple identifier look-up in BioRails with the external compound registration system, translating a complex system into a simple identifier look-up. In contrast the concepts *Vehicles* and *State* are realised as an internal list within BioRails. Hierachical concepts such as taxonomy can also be created using a tree of concepts.

Data System

The BioRails catalogue can be used to federate and categorise concepts that are implemented in external systems for Animal ordering, Compound registration and inventory . The data souce status is used to indicate if the database is online. This enables the system to operate with remote databases which may be taken offline for routine administration tasks.



Parameter types

The BioRails Catalogue includes an added feature for presenting concepts and their implementations as dictionary based parameter types. If a parameter type is added to a concept all the lookups in the concept, including its children, are made available as a parameter type for use in the definition of protocols for capturing data.

For example the implementations for *Solvent* and *State* from above are available as parameter types which can be used for recording dictionary based information.

Realisation

Integration

The BioRails catalogue has been designed to provide a modular implementation of the catalogue concept. This means it can be used to centralise master data independently of the target transactional systems. This is important within organisations with established transactional systems. For example commercial data management systems can be adapted to make use of the centralised catalogue as a source of master data. In this respect the test management features of BioRails make use of the BioRails catalogue module in exactly the same way. The model and SQL based implementations can be used to present a single centralised view of master data from the catalogue even though their implementations are maintained in distinct external systems. Modern systems often make use of service orientated architectures to deliver master data to the transactional systems. The BioRails Catalogue provides an extensive set of web services for accessing catalogue data. Its design realises the most modern integrative techniques, and is in stark contrast to the unilateral integration approaches taken by almost all vendors in this space. All too many software suppliers are motivated only by the desire to be the single vendor used by their customer. The nature of the problems being dealt with almost always transcend the capabilities of single vendors and such integration strategies quickly become barriers to useful deployments.

Other implementations

As we have described the catalogue design is based on public standards and other commercial and non-commercial implementations are available. Examples include enterprise vocabulary systems such as developed by the National Cancer Institute (NCI). A number of commercial solutions are available for master data management from suppliers such as SAP, Initiate and TIBCO. All the commercial solutions target classical customer-product-asset master data systems, and adaptation into the complex world of pharmaceutical research may prove challenging but should be assessed.

Conclusion

Master data management is at the heart of any strong data management strategy, whether the data is captured in Laboratory Information Management systems (LIMS), Electronic Laboratory Notebooks (ELN) or simply in Excel. The value of information can only realised if it can be found and exploited for decision making. Structured, high quality data starts with consistent master data. We have demonstrated not only a standard for representing such master data but also a real implementation in the life science industry (the BioRails Catalogue).

The Catalogue is used to integrate BioRails within an enterprise architecture without the need to replicate data or generate custom connection modules. Externally realised concepts can be quickly deployed within the test management features of BioRails without traditional fragile integration methods. The Catalogue module within BioRails can be used as an independent implementation separate from BioRails providing the basis for a centralised master data system. The advanced concepts and implementations can be accessed from within a heterogeneous IT architecture using a web service integration strategy.

The BioRails catalogue can be used as the basis for a strong implementation of master data, leveraging the most modern techniques in master data management. In general we would recommend using the BioRails catalogue as the basis for a specific implementation of the catalogue concept tailored to the precise business needs of your organisation. This allows us to extend its capabilities in line with the specific requirements for deployment. A number of enhancements are being considered from experience gained through deployment within *in vivo* and translational research environments including support for Web Servers as a data source (data system).

The value of a strong master data implementation is heavily dependant on high quality analysis of the organisation combined with a flexible solution for its implementation. This must strive to separate the master data from the operational systems responsible for data capture whilst facilitating utilisation of that same master data from a variety of environments.

A number of commercial solutions are available for master data management, critical assessment of their utility within a pharmaceutical research environment is recommended before deployment. Use of mainstream commercial solutions could deliver advantages over a customer solution provided they can meet the demands required from this specialist environment.

References

Many of ideas in this paper and for the BioRails Catalogue have been taken from OWL and ISO/IEC 11179 which aim to provide a universal metadata repository.

See the following references for background material:

ISO/IEC 11179 http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/PubliclyAvailableStandards.htm
NCI CADSR http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr/ISO11179
OWL Ontologies http://www.w3.org/2004/OWL/
NCI EVS project http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/vocabulary
Unified Medical Language System http://www.nlm.nih.gov/research/umls/
Gene Ontology http://www.geneontology.org/GO.doc.shtml
TaxonTree http://www.cs.umd.edu/hcil/biodiversity/

About the authors

Robert Shell, Ted Hawkins and Andrew Lemon are founders and principal consultants at The Edge a multidiscipline consulting company based in Guildford, UK.

Futher information

If you would like more information or to discuss a project:

The Edge Software Consultancy Ltd
77 Walnut Tree Close
Guildford
Surrey
SU1 4UH
UK
Email mdm@edge-ka.com
Tel +44 2380 411098
www.edge-ka.com