

From Data to Knowledge: Semantics and Implementations

Judith Goud, Akana, Bennekom, The Netherlands

ABSTRACT

Biopharma organizations are collecting increasingly more data for regulatory purposes, but are also starting to explore data gathering through apps, social media, and other methods for marketing purposes. The transformation of these different types of clinical data to knowledge requires a deep understanding of the data, as well as the drug development process.

It would help tremendously if the same definitions are used across the clinical data lifecycle, allowing for different implementations in the necessary systems. The ISO 11179 registry metamodel provides a good platform for managing concepts and their domains, including their associated implementations.

By binding systems to a central repository with robust common domain definitions and terminologies, and managing the metadata components in order to maximize data reuse, industry can implement more efficient end-to-end clinical lifecycle processes and generate reusable data that feeds the next wave of discovery and innovation efforts.

BIG DATA

According to many studies carried out over the last couple of years, digital data is showing exponential growth.^{1,2} The same trend applies within the Biopharma industry that is collecting increasingly more data for different purposes like regulatory approval and health economics, but also for marketing and research analyses.

Before marketing approval the more complex indications under study warrant collection of more complex data, and regulatory authorities also require more data for data transparency and health economic purposes. Post-approval marketing departments want to know more about the use of the medicinal product to focus their efforts, and regulatory authorities are also interested in ongoing evidence about safety and efficacy of the product. This 'real-world data' is becoming more important to align interests of the various stakeholders typically involved in healthcare.

For the pharma industry this real-world data can ensure continuing innovation and research activities. It can be used to evaluate unmet clinical needs, enhance the identification of patients for clinical trials, and contribute to health technology assessments (HTA) as part of the reimbursement process. After marketing approval real-world data is used to demonstrate the safety of a drug in regular clinical practice and may be essential to align interests of the various stakeholders typically involved in healthcare. In case of conditional approval real-world data collection can be essential to satisfy regulatory requirements.³

For one or multiple reasons mentioned above Biopharma organizations are increasingly exploring big data as well as gathering 'real-world' data through apps, social media and the like.

An important feature of this big or real-world data is that it is largely unstructured, emanating from videos, internet search tracking, customer service phone calls, and other sources of digital data that are typically in a semi-structured or unstructured format.⁴

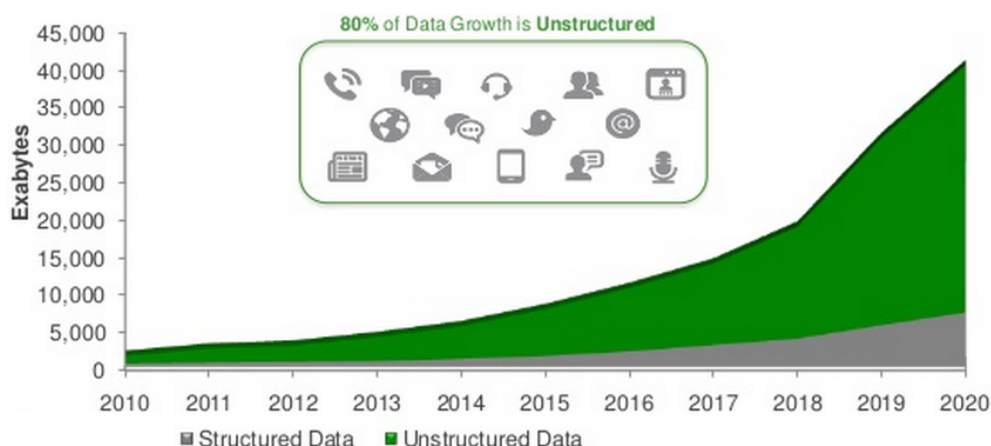


Figure 1 Worldwide Corporate Data Growth (Source: IDC The Digital Universe, Dec 2012)

In 2020 unstructured data is projected to account for approximately 80% of all existent data¹, which makes search and retrieval using conventional storage, database, and business intelligence technologies much more difficult.⁴ According to analysts, the ability to analyze big data will become a key basis of competition and innovation, and they foresee a shortage of talent necessary for organizations to take advantage of big data.¹

BIOPHARMA

Traditionally the Biopharma industry has been collecting vast amounts of data and storing it in a myriad of data gathering and analysis tools such as document management systems, relational databases, datasets, MSOffice, and trial management systems. At the same time many organizations have struggled with sharing information across systems, functions, partners and geographical locations, and using this information effectively. Bringing real world data gathered through apps, social media and internet into the mix complicates the situation further.



Figure 2 Data Sources for generating Effective Information

The challenge for Biopharma organizations is that transformation of these different types of data to knowledge requires a deep understanding of the data, the drug development process and (local) clinical practice. It would help tremendously if all information was available and common definitions were used across the (clinical) lifecycle.⁵

Several initiatives are under way to develop common data definitions from academic institutions⁶, the institutes⁷, and the Biomedical Research Integrated Domain Group (BRIDG)⁸. These clinical data standards provide a basis for organizations to develop their own internal, cross-functional definitions and standards.

However, as these standards and definitions are being developed, they need to be managed in such a way that they can be used consistently across the organization. The ISO 11179 registry metamodel⁹ provides a strong platform to achieve this goal.

ISO 11179

According to Wikipedia "The ISO/IEC 11179 registry metamodel is a result of combining semantic theory with data modelling, and it provides a platform for managing definitions and their associated implementations".

Semantics is defined as 'the study of meaning', independent from structure. The following two sentences provide an example of the same syntax but different meaning:

Time flies like an arrow
Fruit flies like a banana

The same words are used, but they have a different meaning in the context of the sentence (e.g. 'flies' is a verb in the first sentence and a noun in the second example).

Semantic theory was first introduced by linguist and philosopher Chomsky. He characterized the job of linguistics as construction of a system of rules which represent 'what a fluent speaker knows about the semantic structure of his language'. Such a theory must contain rules that allow the decomposition of sentences to understand the meaning. To achieve this, semantic theory consists of two parts: a dictionary, and a 'distinctive set' of 'projection rules'.¹⁰

ISO 11179 translates the 'dictionary' principle as a thesaurus type relation between wider and more narrow (or specific) concepts, e.g. the wide concept "color" has a relation to the more narrow concept "hair color", while a basic principle of data modelling is the combination of an object class and a characteristic, for example, "Person - hair color".

Thus, ISO/IEC 11179 combines a wide 'concept' with an 'object class' to form a more specific 'data element concept'. For example, the high-level concept "color" is combined with the object class "person" to form the data element concept "hair color of person". Note that "hair color" is more specific than "color".

The 'projection rule' principle from semantic theory is implemented as a relation between a concept and its representation, e.g., "sex" and "gender" are the same concept although different terms are used. The different possible representations of a data element concept are described with the use of data elements.

Differences in representation may be a result of the use of synonyms or different value domains in different data sets in a data holding, for example "gender" is a data field of 8 characters in a clinical data management system, while "sex" represents the 3 character variable in a SAS dataset.

ISO 11179 further defines the concepts and data elements through permitted range of values ('value sets'). These are called conceptual domains and value domains. The conceptual domain contains all values available worldwide that can be associated to a concept, whereas the value domain represents the (subset of these) values used for a particular representation (a 'codelist').

An example of a conceptual domain for 'sex of person' is Female, Male, Bisexual, Undifferentiated, and Unknown. For the representation of 'sex' in a SAS dataset the value domain can be M = Male, F = Female, U = Unknown. The letters M, F and U are then the permitted values of sex of person in a particular data set.

The above descriptions can be represented as the following graphical diagram:

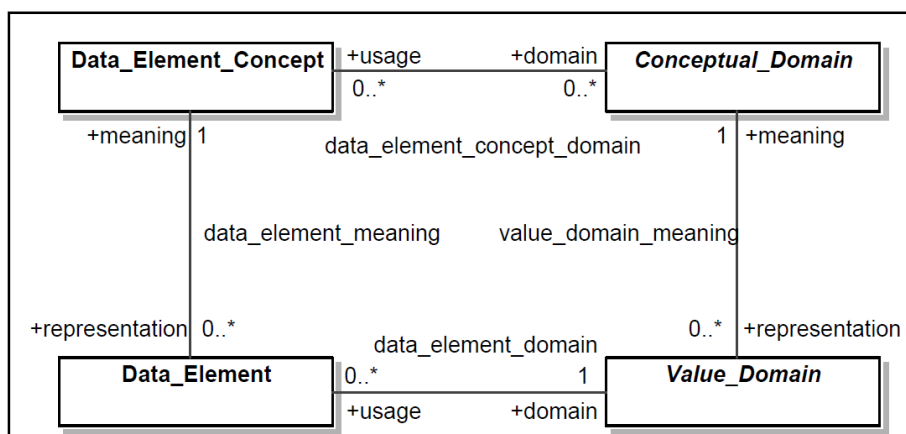


Figure 3 ISO/IEC 11179 High-level Data Description Model

The core object is the data element concept, since it defines a concept and, ideally, describes data independent of its representation in any one system, table, column or organization.

According to ISO/IEC 11179 each Data element in a metadata registry:

- should be registered¹¹
- should be uniquely identified within the register¹²
- should be named according to Naming and Identification Principles¹²
- should be defined by the Formulation of Data Definitions rules¹³
- may be classified in a Classification Scheme¹⁴

Data elements that store 'codes' or enumerated values must also specify the semantics of each of the code values with precise definitions.

Data elements and code values can be precisely defined by capturing data about the data, or 'metadata'.

METADATA

Metadata can be defined as 'structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource'¹⁵. This definition covers two aspects: information about the definition and purpose of an element, plus its usage-specific information.

The definition information should follow industry level standards, such as those defined by FDA and WHO¹⁶ or developed by the International Society for Pharmaceutical Engineering¹⁷, or at least be agreed upon cross-functionally within the organization to facilitate consistency and data exchange.

The usage-specific information describes the data characteristics like the name and version, and how the data element is formatted in different systems and artifacts. For example the vital sign heart rate is a text element named 'Pulse Rate' in the protocol document and a numeric data field of 3 digits with unit beats/min in the clinical database. It also contains

information about references between systems and artifacts and transformations the data element is subject to during the data lifecycle. This type of metadata provides clear lineage between clinical data lifecycle stages and associated systems.

The information about purpose and usage describes at what moment a data element was used and why. This is crucial for true end-to-end data transparency as it shows the relationships from data as defined for a specific collection use case (e.g. clinical trial protocol as the starting point of data generation) through its use during analysis and reporting for a dossier (e.g. describing the drug's safety and efficacy). Further it enables tracking across all usages (e.g. different studies) and impact analysis of changes or new data elements.

METADATA MANAGEMENT

The introduction of clinical data standards and registry standards like ISO 11179 raised the awareness in many companies that these standards need to be managed in a way that facilitates their consistent use¹⁸. Currently the majority of BioPharma companies manage data definitions using documents and spreadsheets. However these tools are designed for other purposes and therefore do not support many requirements for proper content management. This results in issues such as change control/versioning, business rule implementation, impact analysis of changes, and reuse of elements. The outcome is creation of copies (and keeping track of them) and team consultation, involving a significant amount of manual effort, which can lead to human error as processes need to be set up for review, governance, approval and tracking of impact analysis and where data definition changes need to be implemented.

A metadata repository is better suited for these activities and can serve as the single source of truth within an organization. The repository should hold all information about data element structure and meaning, and where and when a data element is/has been used. Ideally it also enables (automated) reuse of data elements and captures of all references and transformations the data element was subject to during the data lifecycle. Figure 4 shows an example how a centralized repository can overcome the challenges and enable effective data standard implementation.

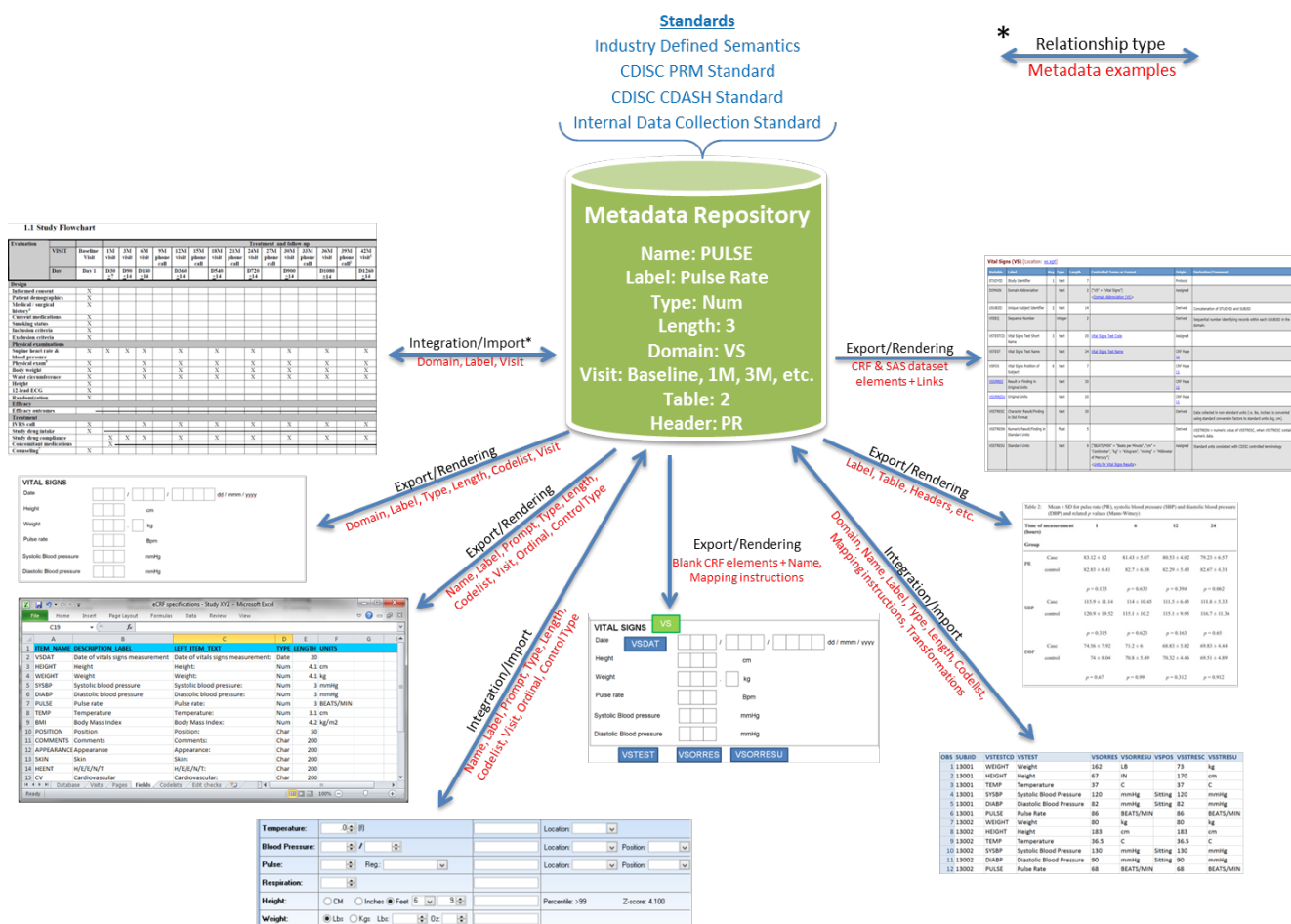


Figure 4 How a Repository Enables Consistent Metadata Management

A metadata repository based on the ISO11179 data description model offers further advantages of differentiation between the definition (conceptual level) and usage-specific (data element level) information. It also provides a solid structure to manage the metadata consistently to facilitate data exchange, as:

- Metadata can be managed at the data element, codelist and permitted value level, which facilitates reuse;
- Systems can be bound to a central repository managing robust common domain definitions and terminology;
- Automated governance workflows can be applied, capturing the history and reasoning for updates ; and
- Industry can generate reusable data that enables more efficient, automated, end-to-end clinical lifecycle processes.

CONCLUSION

To 'bring Big Data to life' sounds easy enough in theory, but when the actual data is collected, especially in large quantities, it can be extremely difficult to access and process the precise data needed. The organization of big data sets and provision of common definitions of data elements can significantly help with understanding the information that feeds the next wave of discovery and innovation efforts.

The ISO 11179 Data Description model offers a platform for managing the definitions including their associated implementations across systems and organizations. Tying the concepts to domains and industry terminologies creates a robust conceptual layer that can be associated with operational layer(s) of usage (system) specific data elements and value domains.

Managing metadata in an ISO 11179-based repository will support knowledge development by:

- Creating, maintaining, governing and using data definitions consistently and associating them with specific implementations,
- Maximizing reuse of existing artifacts and data assets, and
- Knowing data retrieved from/exchanged with other systems (like social media) is being used or interpreted correctly.

REFERENCES

1. *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, J Gantz and D Reinsel, IDC View, December 2012
2. *Big data: The next frontier for innovation, competition, and productivity*, J Manyika, M Chui, B Brown, J Bughin, R Dobbs, C Roxburgh, and A Hung Byers, Report| McKinsey Global Institute, May 2011
3. *The importance of real-world data to the pharma industry*, G Bonnelye and A Miniuks, PMLiVE, 26th June 2015
4. *The big data big bang*, D Neef, Nov 24, 2014
5. *Information Sharing Core Principles*, NHS, Dec 2012
6. *Health Sciences Library*, University of Washington, <http://guides.lib.uw.edu/hsl/data/findclin>
7. *CBIIT Project: Semantic Infrastructure 2.0 Roadmap*, US National Cancer Institute
8. *BRIDG model*, <http://bridgmodel.org/>
9. *ISO/IEC 11179-1:2004 Metadata registries (MDR) - Part 1: Framework*, http://www.iso.org/iso/catalogue_detail.htm?csnumber=35343
10. *Syntactic Structures*, N Chomsky, 1957
11. *ISO/IEC 11179-6:2005 Metadata registries (MDR) - Part 6: Registration*, [http://standards.iso.org/ittf/PubliclyAvailableStandards/c035348_ISO_IEC_11179-6_2005\(E\).zip](http://standards.iso.org/ittf/PubliclyAvailableStandards/c035348_ISO_IEC_11179-6_2005(E).zip)
12. *ISO/IEC 11179-5:2005 Metadata registries (MDR) - Part 5: Naming and identification principles*, [http://standards.iso.org/ittf/PubliclyAvailableStandards/c035347_ISO_IEC_11179-5_2005\(E\).zip](http://standards.iso.org/ittf/PubliclyAvailableStandards/c035347_ISO_IEC_11179-5_2005(E).zip)
13. *ISO/IEC 11179-4:2004 Metadata registries (MDR) - Part 4: Formulation of data definitions*, [http://standards.iso.org/ittf/PubliclyAvailableStandards/c035346_ISO_IEC_11179-4_2004\(E\).zip](http://standards.iso.org/ittf/PubliclyAvailableStandards/c035346_ISO_IEC_11179-4_2004(E).zip)
14. *ISO/IEC 11179-2:2005 Metadata registries (MDR) - Part 2: Classification*, [http://standards.iso.org/ittf/PubliclyAvailableStandards/c035345_ISO_IEC_11179-2_2005\(E\).zip](http://standards.iso.org/ittf/PubliclyAvailableStandards/c035345_ISO_IEC_11179-2_2005(E).zip)
15. *Understanding Metadata*, National Information Standards Organization, 2004
16. *Terminology & Definitions in Pharmaceuticals*, <http://www.pharmaguideline.com/2011/09/terminology-definitions-in.html>
17. *ISPE Glossary of Pharmaceutical and Biotechnology Terminology*, <http://www.ispe.org/glossary>
18. *Metadata standards and metadata registries: An overview*, B Bargmeyer (Environmental Protection Agency), and D Gillman (Bureau of Labor Statistics), Washington, DC 20212

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Judith Goud
Sr. Solution Architect BioPharma - Europe
Akana
6-9 The Square
Stockley Park, UB11 1FW
United Kingdom
Work Phone: +31 (0)6 1241 7026
Email: judith.goud@akana.com
Web: <http://www.akana.com>

Julie Smiley
Director of Product Management - BioPharma
Akana
12100 Wilshire Blvd., Suite 1800
Los Angeles, CA 90025
United States
Work Phone: +1 210 710 2901
Email: julie.smiley@akana.com
Web: <http://www.akana.com>

Brand and product names are trademarks of their respective companies.