TAXONIC

Reference data as a service

How emerging technologies support the next level of data governance

Material Subject to Creative Commons License:

A Taxonic whitepaper, by Jan Voskuil @Taxonic taxonic.com

TAXONIC

Contents

Introduction

Large amounts of information are exchanged in the form of PDF- documents. Increasingly, text is replaced by data, creating new data governance challenges. A case in point is the pharmaceutical industry. New EU directives force pharmaceuticals to move from text based to data based submissions to obtain market authorization for medicines. This involves data structures with an estimated 1700 data points, using more than 75 reference datasets. These reference datasets ensure every submission uses the same terminology to refer to countries, currencies, substances, diseases, adverse effects and more.

Protecting the quality of reference data is an essential ingredient of data governance, just as important as monitoring the quality of IT-systems. Organizations start to act accordingly. This whitepaper investigates what reference data management is, why it is important, which processes it involves, and how this practice can be optimally supported using emerging technologies.

When we exchange text, our linguistic abilities enable us to unambiguously interpret its contents, with a rigor and subtlety that are, frankly, quite baffling. Even today, there is no scientific consensus on how these linguistic abilities work. Indeed, the capabilities of today's most advanced computer programs to understand text are still lightyears away from ours.

When exchanging data, interpretation of their meaning is paramount. Businesses in the US and Europe are required to report their public financial statements in the form of data: a file with raw numbers and codes, generated by bookkeeping software. How do we know a particular number stands for the business' current assets, or, instead, its liabilities? The use of reference data is a key to the answer. Reference datasets define the permissible values used in data fields and constitute a taxonomy for categorizing raw data into meaningful chunks. Put differently, reference data are used to nail down the semantics of other data. This paper argues that semantic web technology is particularly relevant for reference data management.

One of the core challenges in managing reference data is to handle their structure. Reference data sets are list, each having its own varying number of columns. Many are organized in a tree-like, hierarchical structure — adding even more complexity. Take for instance Medical Subject Headings (MeSH), a comprehensive controlled vocabulary for the purpose of indexing articles and books. It serves as a thesaurus that facilitates information retrieval. See Figure 1. The reason for adding this type of complexity is knowledge management. When you don't know the exact form of the concept you are looking for, a hierarchical order comes in handy. You can drill down to find exactly what you need, and navigate the tree to discover related items. Now suppose you have your own library with medical papers and you want to use MeSH for assigning topics to them. A standard for exchanging reference datasets, including a way of representing hierarchical structure, is necessary to optimally support such use cases. As we will see, semantic web technology provides just the right standards for that.

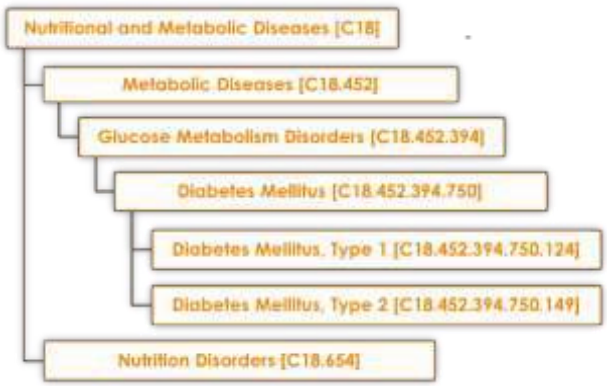

Figure 1 Vocabularies are often hierarchically structured, as shown in this extract of MeSH.

1 The Rise of Reference Data

Reference data management is an area within Master Data Management (MDM) — the discipline that aims at structural improvement of data quality. MDM requires a data governance organization that puts in place policies and procedures. The goal is to provide the end user community with a trusted single version of the truth. Based on this, informed decisions can be taken at every level in the organization.

The driving force behind MDM is the insight that there different types of data, each with its special characteristics, challenges and concerns. Table 1, cited from a paper[1] by the renowned data governance expert Malcolm Chisholm, shows how reference data compares to other kinds of data. Each kind of data has its own particular characteristics and governance and management needs.

The term "metadata" in this table needs some explanation. The term is widely used for a variety of things centered on providing information about other data. In this context, the term is intended to designate the information typically encoded in column headings in spreadsheets, and database table structures. We will return to this type of metadata shortly.

METADATA	Metadata	Describes the structure and semantics of information assets
MASTER DATA	Reference data	Categorizes other data (e.g., currencies, countries, substances)
	Transaction Structure data	Represents entities in transactions (e.g., customers, products)
	Enterprise Structure data	Represents responsibilities in transactions
EVENT DATA	Transaction Activity data	Represents the results of operations being carried out (e.g., sales records, payment records)
	Transaction Audit data	Tracks individual transactions (e.g., transaction log files)

Table 1. Different kinds of data.

The different types of data in the table are ordered by value and risk. Data types more towards the top have increasing value: errors are multiplied in all other datatypes lower in the hierarchy. Data more towards the bottom tend to exhibit larger volumes and higher dynamics. New transactions may occur every second, whereas updates in the list of countries occur once a year or less.

The claims in this paper will be illustrated based on the regulatory affairs domain within the pharmaceutical industry. While reference data management is not tied to any specific domain, this allows us to clarify concepts based on interconnected examples. Therefore, we start off on a short description of the domain and the challenges it has to deal with.

1.1 Regulatory Affairs and the IDMP Directive

To get a medicine on the market in any country requires one to go through an extensive procedure in which a well-documented request is submitted to the health authorities. Before the early two thousands, this was done on paper. A submission would typically be made up of some 200 volumes of about a thousand pages each. Starting in 2003, these documents were often digitized as PDF-files and submitted by way of a compact disc, and nowadays the internet is the preferred channel. It is important to note that while this is obviously a step forward, the information being exchanged between the market authorization holder (that is, the company trying to obtain permission to sell the medicine) and the health authorities still is information in the form of written text, essentially digitized paper documents.

[1] Malcolm Chisholm (2015) The Foundations of Successful Reference Data Management.

The European Commission, in cooperation with the US Secretary of Health and Human Service, has decided that a significant number of key describing characteristics of an authorization application shall henceforth be submitted in the form of raw data. In line with this policy, the European health authority, EMA, has started an initiative to standardize reference data sets for expressing substances, products, organizations, and so-called "referentials". These standards have been formalized and published by the ISO. Together, they are referred to as IDMP: Identification of Medicinal Products. As of June 2016, the directive will be effectuated in a stepwise manner.

The sheer amount of raw data comprised in an IDMP-compliant submission is daunting. The number of defined relations and attributes exceed 400. Many object classes will recur multiple times in a submission: a given medicine may involve multiple substances, each substance may be produced by multiple suppliers, and each of these objects will have specific attribute and relation values. Some estimate that the number of data points in a typical submission will exceed 1700.[2]

This has impact. At a recent conference, one speaker quipped that pharmaceuticals can have only one priority now that the skies are falling in Europe. However, it is not only European submissions that are affected. The US has stated that it will effectuate a similar directive, making the same IDMP format obligatory in the US two years after the EU. Other countries in the world will follow sooner or later.

1.2 Explaining the Trend: the Importance of Data

The objectives behind IDMP are to increase the quality of life for all of us. Ensuring the quality of authorized medicines requires working with immense amounts of information. Computers can do that more effectively than humans, but they require data, not text.

As an example, consider the situation where health authorities discover that there is a serious problem with a specific substance produced by a specific supplier. This substance may be used in many medicines marketed by many market authorization holders. Now the question is: exactly which medicines are affected and need to be taken off the market immediately? The answer to this question is currently buried in gargantuan amounts of PDF-documents. No search engine will ever provide a reliable answer. The only way to resolve this structurally is expressing the information in the form of raw data, and let a computer process these. It is precisely this set of data that the IDMP directive will produce.

Texts are good at expressing subtle nuances and capturing a train of thought, but since text requires a person to interpret it, this does not scale against the need to make decisions based on large amounts of information. This results in pressure to express information in the form of data, so that computers can take over the work. The developments in the pharmaceutical industry are therefore matched by similar developments in finance, retail, manufacturing, agriculture, automotive and many other domains. This trend necessitates a new approach to managing data quality. Reference data management is a key ingredient of this.

[2] Lior Keet (2016). "Facilitating IDMP Compliance While Aligning With Corporate Data Integration Strategies". Paper presented at the DIA Regulatory Submissions, Information and Document Management Forum conference, 8-10 February 2016, North Bethesda, USA

2 The Nature of Reference Data

2.1 Reference Data in a Nutshell

Reference data sets are very commonly used. Sometimes they are called code lists, value lists, controlled vocabularies, business vocabularies, look-up tables, taxonomies, thesauri or coding systems. They define permissible values in certain data fields, thus providing information needed to make other data meaningful and interpretable in an unambiguous way.

Say, a product can be safely stored until 3 days after opening. To make sure everyone involved has the same understanding of which unit of measure is meant, the different options for specifying it are taken from a shared list. This list is the reference dataset for units of measures and will contain such items as days, months, years et cetera. Each entry in the list will contain a code to be used in records ("D"), a descriptive label associated to it, perhaps in different languages ("Days" in English, "Jours" in French), along with other information that provides users with understanding as to the meaning of the entry.

This enables information exchange, and is also a prerequisite for effective records keeping over time. If you find in the records of your organization that many years ago products were sold to a customer in a country with country code GDR, it is the reference dataset that enables you to retrace this to the German Democratic Republic, even though the country is not existent anymore. Without a managed set of reference data sets, historical records soon become incomprehensible.

Since the purpose of reference data is to enable shared understanding, it stands to reason that many reference datasets are defined by an external party, often a standards body such as ISO. In addition, however, every organization will have reference datasets defined internally: the list of cost centers and journal headings, the list of function profiles, the list of production locations, of customer types, of product lines, et cetera.

To manage reference data properly, a data governance unit must be formed. A central role within this group is the data steward. He or she is the one who is operationally responsible for the required management and administration tasks. We discuss these tasks below in Chapter 3.

2.2 Crosswalks

Figure 2. Different reference datasets use different identifiers for the substance called simvastatin. Source: Wikipedia.

Different IT-systems in different parts of the organization may use different reference datasets for the same subject. For identifying substances, for instance, at least ten are in use world-wide, each taking a slightly different perspective on the subject — see figure 2. This makes it necessary to create an overview of which reference datasets form what is called a terminology group. Translating from one

Figure 2. Different reference datasets use different identifiers for the substance called simvastatin. Source: Wikipedia.

reference dataset to another within the same terminology group becomes necessary, using a mapping table called a crosswalk. Crosswalks thus enable meaningful exchange of information between systems using different terminologies for the same thing. The topic of terminology groups and crosswalks merits a separate paper. For our present purposes, it suffices to observe that crosswalks push the need for quality and reliability of reference datasets even further.

2.3 Risks and Benefits

The importance of reference data management increases with the number of reference datasets in use. It is estimated by some that roughly 30% of all tables in an average database are reference data tables. An IDMP-compliant submission contains values taken from more than 75 designated reference datasets. Each is the designated vocabulary within a larger terminology group: the other vocabularies in each group cannot be used for IDMP purposes, making a translation necessary based on a crosswalk.

With a growing number of IT-systems that exchange information, and the number of reference datasets increasing, it is unavoidable that problems stemming from inconsistent versions or unseen errors will at some point spiral out of control, with significant impact, potentially even jeopardizing business continuity.

In this connection, it is relevant to reflect for a moment on quality processes for information systems in general. In the regulatory affairs domain, standards are extremely high in this regard. IT-systems are required to be certified, software vendors are subjected to audits. Surprisingly, however, the quality of reference data is unmanaged. When done properly, reference data management offers corresponding benefits: control and agility in data governance, and managed quality of data which leads to predictable results of business processes driven by these data.

2.4 Reference Data and the Future of Existing IT-Systems

To enable the management processes that warrant reference data quality, IT- systems need to fulfil an important requirement: they need to make it possible to control the versioning of the reference datasets they use. Put differently, the system must offer controls by which one can make it use a new version without needing to change the software. It must externalize reference datasets.

Some software products manage reference data as part of the system, instead of as an externally controlled resource. This is a serious design flaw that makes the reference data versioning puzzle practically unsolvable. The whole idea of having reference datasets defined by external bodies is precisely the separation of concerns between end user functionality (which is the responsibility of the software vendor) and data governance (which is the responsibility of the business sponsor). Data originating from different silos within the organization, and also from outside, need to be consolidated and processed as a consistent whole. To guarantee consistency, reference data sets must be managed centrally.

The business sponsor has to be in control of which systems use which versions. Therefore, only IT-systems that allow administration of reference data from outside are sustainable in the long run.

3 Supporting Reference Data Management

3.1 Linked Data and the Semantic Web

When World Wide Web Consortium (W3C) director Tim Berners-Lee invented the Web in 1989, he envisaged not just the web of documents that we know today, but rather a web of data: the Semantic Web. Linked Data is a technology standardized by W3C that has emerged from more than two decades of work on this. The single most important challenge in creating the Semantic Web is: when we place raw data on the Web, how can make these data useful and interpretable? The answer to this fundamentally changes the way we work with data.

In the Linked Data approach, the answer is two-fold. First, we give all conceptual things — called resources — a globally unique name, more specifically, a URI. A fictitious example would be the URI https://www.sis.gov.uk/ to refer to James Bond. This guarantees that the names we use remain recognizable and interpretable, wherever on the web we use them. Contrast this with a traditional database or spreadsheet. If you take a customer number out of a spreadsheet, say the locally defined customer number 11241, it loses its meaning.

The second part of the answer is that we structure data in the form of a graph. The nodes and edges are resources signified by URIs. A graph can be represented by a bag of triples of the form: <subject, predicate, object>. Consider Figure 3. Crucially, this graph contains some data, and also some information that pertains to the data model or schema — what we called metadata in Table 1. Thus, the fact that Moneypenny works for MI6 is a fact about our world (hence, data), whereas the statement that "works for" is a relation between a person and an organization is a statement about the language in which facts are expressed (hence, metadata). The abbreviated URI's starting with "mi6:" are made up, but the others exist. They are minted by standards bodies, have rigorous semantics and can be looked up on the Web by typing them in the address bar of your browser.

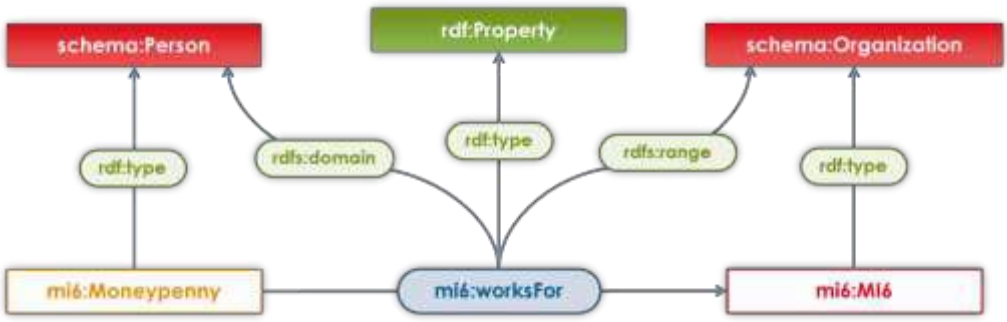

Figure 3. A simple example graph

We could publish the triples representing this graph in one file, but also separate them out into two distinct files, as in Figure 4. We could then publish the file called "model.ttl" on the web, so that others can use the same ontology in their datasets, while keeping the actual data in the file "data.ttl" secret, as behooves a secret intelligence agency.

Note that much of the ontology used to define the semantics of our little dataset is not even defined in the file “model.ttl”: it is defined in the other ontologies mentioned — RDF, RDFS, and schema.org. RDF and RDFS are foundational standards made available by W3C, while schema.org is a general purpose ontology created and maintained by a cooperation of several Web giants, including Google and Facebook. By itself, our data model only defines a single property: worksFor.

Figure 4The example graph can be represented using triples in different files.

One idea behind reusing existing RDF data models is that it leads to better data models for less money. Data modeling is highly specialized work, and small errors have large impact — metadata are on top of data value hierarchy in Table 1. An additional benefit of reusing ontologies is an unprecedented level of interoperability. Different parties can now share the same ontology to express their data with, obviating the need for conversions, ETL-processes and other plumbing.

This approach is more revolutionary than it seems. There simply is no way in which we could put the columns of a spreadsheet (or database) in one file, and the data in another. The moment we separate spreadsheet data from the spreadsheet structure, the spreadsheet data lose their context-dependent meaning and, thus, become useless. Consequently, there is also no way in which we could publish models on the web, and use several of these models simultaneously to describe the semantics of our spreadsheet or database data.

This feature of Linked Data, the ability to treat data and metadata alike, to combine or separate data and model, and to reuse and mix data and data models from different sources, has enabled us to start realizing Tim Berners-Lee’s vision of the Semantic Web. And it is of eminent importance to reference data management.

3.2 SKOS

Simple Knowledge Organization System, or SKOS for short, is a concise ontology describing the way reference datasets — such as coding systems, term lists, glossaries, controlled vocabularies, thesauri, taxonomies, library indices and the like — are structured. It is a W3C-standard ratified as a recommendation in 2009 and broadly used. In the Netherlands, SKOS is adopted on the national ‘comply- or-explain’-list, effectively making it an obligatory standard for public sector organizations publishing reference datasets.

Such knowledge organization systems often include some notion of hierarchy. SKOS is designed to capture this informal structure. Take for instance the library thesaurus of the WODC, the research institute of the Dutch Ministry of Security and Justice, which is used to

structure the content of its extensive Web library. We find that both “juvenile offender” and “adult detainee” have “detainee” as broader term. In this case, the relation seems to imply a subclass-relation of some sort, as in “an adult detainee is a (kind of) detainee”. Elsewhere we find that “penal institution” has “criminal law” as broader term, so that the relationship implies more of a thematic kind of relatedness. We cannot meaningfully say that “a penal institution is a (kind of) criminal law”.

Thus, the semantics of the hierarchy in thesauri, controlled vocabularies and other reference datasets does not coincide with a formal “subclass of”-relation. Because

it is exactly this notion of hierarchy that SKOS captures, it does not allow for inferences to be rigorously made as truly formal logics-based ontologies do. Therefore, SKOS is sometimes characterized as defining a “soft semantics” for knowledge organization. The centerpiece of SKOS is the notion of “Concept”, which is succinctly defined as “An idea or notion; a unit of thought.” This resource’s URI is, abbreviated, skos:Concept. A number of predicates are defined that express semantic relations between such concepts, such as skos:broader. Some other predicates have strings as value and indicate what the preferred term and alternative terms are. A simple example of how SKOS works is given in Figure 5.

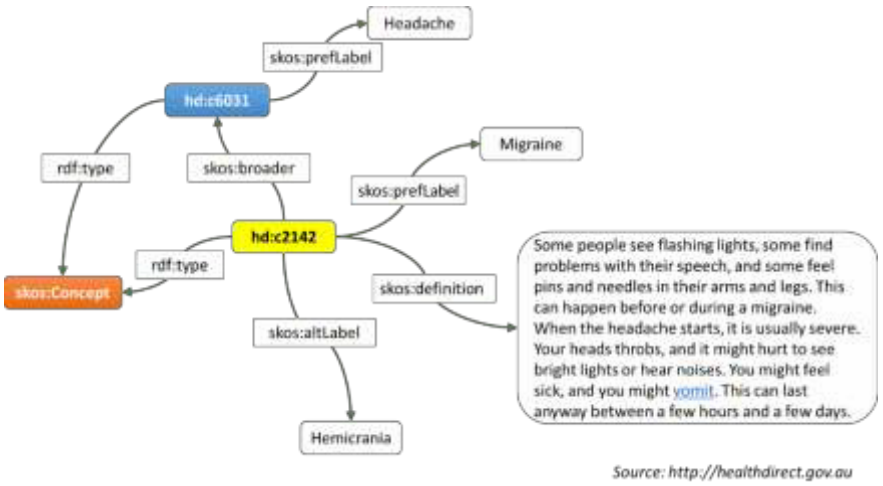

Figure 5Two concepts, their attributes and their relations, organized using the SKOS standard.

This graph tells us that there is a Concept with the preferred label “Mi- graine”, and which is also called “Hemicrania.” It is related to the broader concept labeled “Headache”. SKOS also makes available a number of predicates for relating concepts across different datasets, such as “exactMatch” and “closeMatch”. This makes

it possible to map concepts in different reference datasets in a consistent way and thus to create crosswalks using a rigorously standardized representation form — a form that can be exchanged without hassle.

3.3 Business Processes

Let us now see how Linked Data technologies can help us executing essential reference data management tasks. The consensus is that in the context of reference data, the data steward at least eight major tasks to perform:

- Profile organizations. Create a database of originators of reference data, their contact information and other details.
- Profile reference datasets. For each reference dataset, what it is for, how it is structured, which parts are relevant, how often it is updated, and so on.
- Execute semantic analysis. What is its exact structure, and what do the different pieces of information mean?
- Document semantic analysis. The results of the semantic analyses must be made available to the end user community, preferably in a way that makes optimal sense from actual work processes.
- Import reference datasets. The reference datasets are imported in a central repository where they are administered and overseen in one spot.
- Assign accountabilities. Who decides on life cycle events? Who needs to be consulted and to be informed?
- Track changes. Reference data are curated and business rules are checked. After approval, the new version becomes the production version.
- Distribute reference data. The production version must be made available to IT-systems using whichever technology the IT-system supports.

Linked Data offers significant benefits in supporting these processes. Let us highlight a few aspects, starting with executing semantic analysis. As a running example, we take the list of ATC-codes. This reference dataset is used for the classification of active ingredients of drugs. It divides drugs into hierarchical groups at five distinct levels. The code of each substance reveals at which level it is placed. For instance, A01AA03 (“olaflur”) is a 5th level drug group, while A10 (“drugs used in diabetes”) is a 2nd level group. See Figure 6. Given the hierarchical relations in the dataset, SKOS is an excellent starting point in analyzing its semantics. The primary entity in the dataset is Substance: every entry in the ATC-list characterizes exactly one of these.

	A	B	C	D	E	
1	ATC code	ATC level name	DDD	U	Adm.R	Note
2	A	ALIMENTARY TRACT AND METABOLISM				
3	A01	STOMATOLOGICAL PREPARATIONS				
4	A01A	STOMATOLOGICAL PREPARATIONS				
5	A01AA	Caries prophylactic agents				
6	A01AA01	sodium fluoride	1.1	mg	O	0.5 mg fluoride

Figure 6. An extract of the Excel-sheet with ATC-codes, as obtained from WHO.

So, the first thing we do in our analysis is stating that ATC-data are about Substances. We give this notion a URI, ex:Substance, and make it a subclass of skos:Concept. By saying that Substance is a "subClassOf" Concept, it follows that Substances have the same hierarchical relations as Concepts. This neatly captures the child-parent relations in the ATC hierarchy.

To make this concrete, Figure 7 shows the result of the data steward's analysis[3]. In the left pane, we see a class hierarchy and for each class, the properties associated to it. In this hierarchy the data steward has imported the SKOS ontology. It only has three classes, Collection, Concept and ConceptScheme. We can see that the data steward has created a subclass of Concept called Substance and defined five properties for it, conforming to the structure of the ATC reference dataset. The right pane shows the details of the class highlighted in the right pane, in this case, skos:Concept. In the right upper corner we see this resource's URI, which is minted by W3C. Note that all this information is published by W3C on the Web, and taken from there.

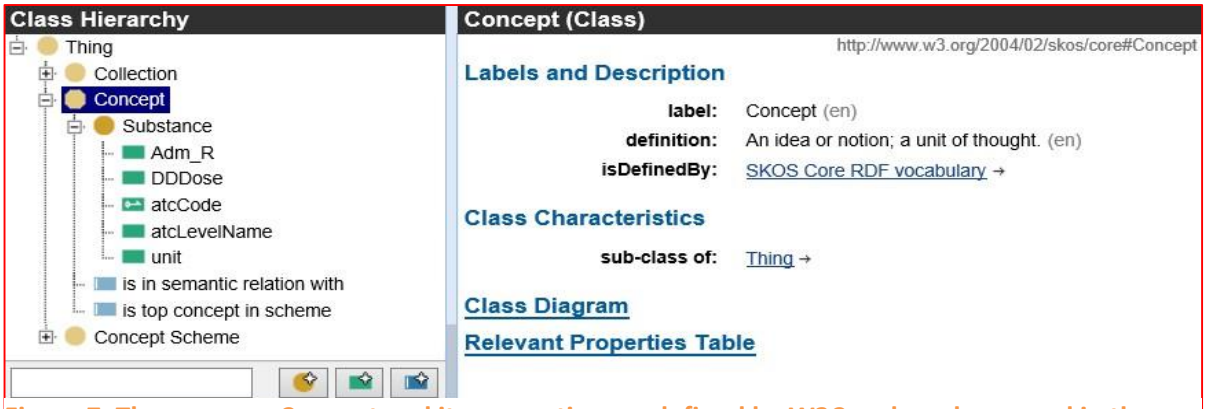

Figure 7. The resource Concept and its properties are defined by W3C and can be reused in the data model describing the structure ATC codes.

[3] The screenshots in this paragraph have been created using TopBraid Reference Data Manager™. Permission to use these has been kindly granted by TopQuadrant Inc., which is gratefully acknowledged.

Highlighting one of the properties we created for Substance, DDD, gives us the screenshot in Figure 8. It has a label, and a definition. In this field the data steward has just pasted the definition found in the ATC documentation. Assuming that the Linked Data server is available in the enterprise, this information is just a hyperlink away from the place where an end user needs to read or input the DDD for a specific drug. As noted previously, this is important for end users, who need access to this kind of semantic information to avoid errors and time consuming search activities.

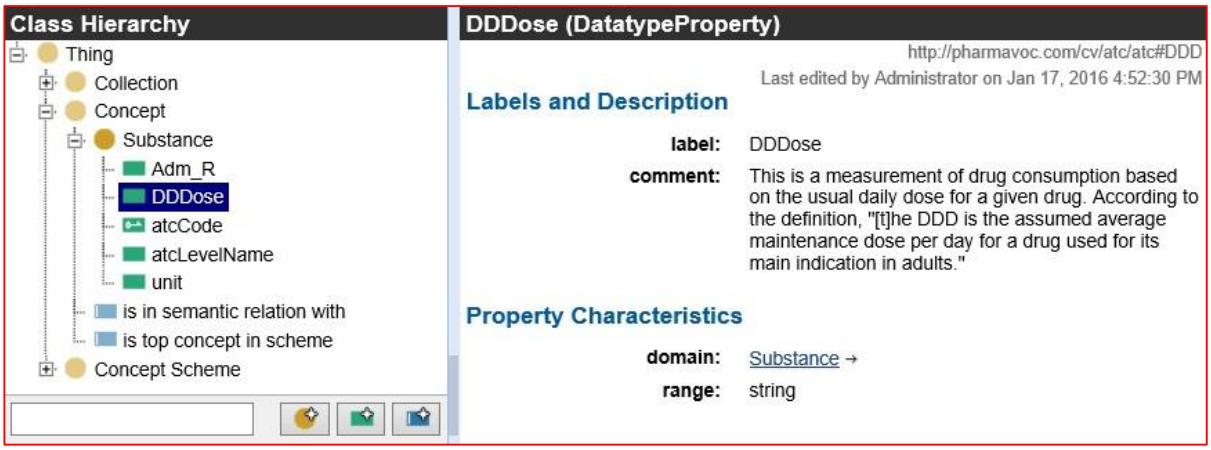

Figure 8. The metadata of each ATC-property, like DDDose, can be managed easily using Linked Data technology.

Now that we have defined the semantics of ATC-codes, the next thing we need to do is import the actual ATC codes into a central repository. The ATC reference data set can be obtained in the form of an XML file or in spreadsheet format, as shown above in Figure 6. An importer tool based on Linked Data technology is easy to provide.

All we need to do is to tell the importer which of the columns in the spreadsheet corresponds to which property in the ontology, and how to deduce for each entry the correct hierarchy level by looking at the code. The importer will then create a new resource for each entry in the list, give it a URI, get the property values right, and create the hierarchical relations. The result is in Figure 9.

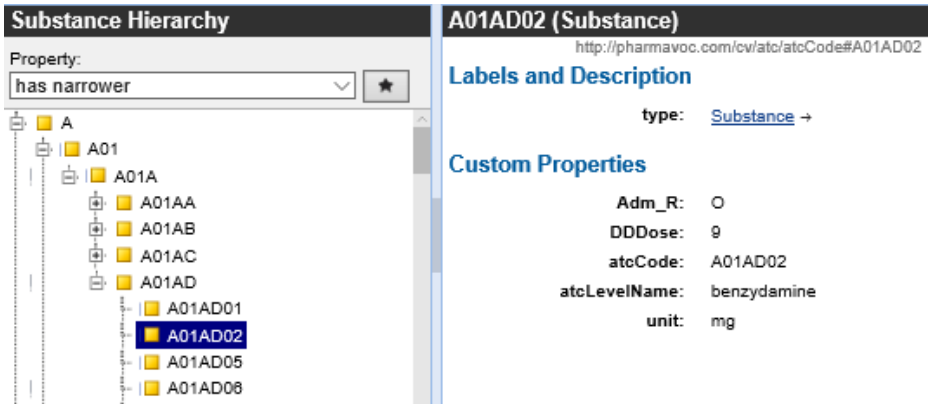

Figure 9. The actual ATC-data converted to Linked Data.

The left pane displays the hierarchical structure of the ATC reference dataset, reusing the SKOS ontology. We have highlighted the substance with code A01AD02. The right pane displays the details of this substance loaded from the spreadsheet. In addition, it tells us that the type of this resource is Substance. The hyperlink takes us to the definition of this resource in the ontology we just created. The top right corner displays the URI that was generated for this substance.

Increasingly, standards bodies will publish their reference datasets using SKOS. This is already happening in the library world and in bio informatics. This will obviate the need for conversion of Excel sheets and XML-files and further enhance interoperability. It is important, though, to realize that even when this vision becomes a reality, reference data management as a discipline still needs to be executed to warrant data quality.

3.4 Benefits of Linked Data

The benefits that solutions based on Linked Data technologies can bring to reference data management are manifold. Perhaps the most striking benefit is the ease with which the data models of different reference datasets can be captured by reusing and extending existing ontologies such as SKOS. This makes it possible to pin down the semantics of reference data and create crosswalks in a rigorously standardized way. Solutions based on Linked Data are expressly designed from the ground up to deal with the diversity of data models underpinning reference datasets.

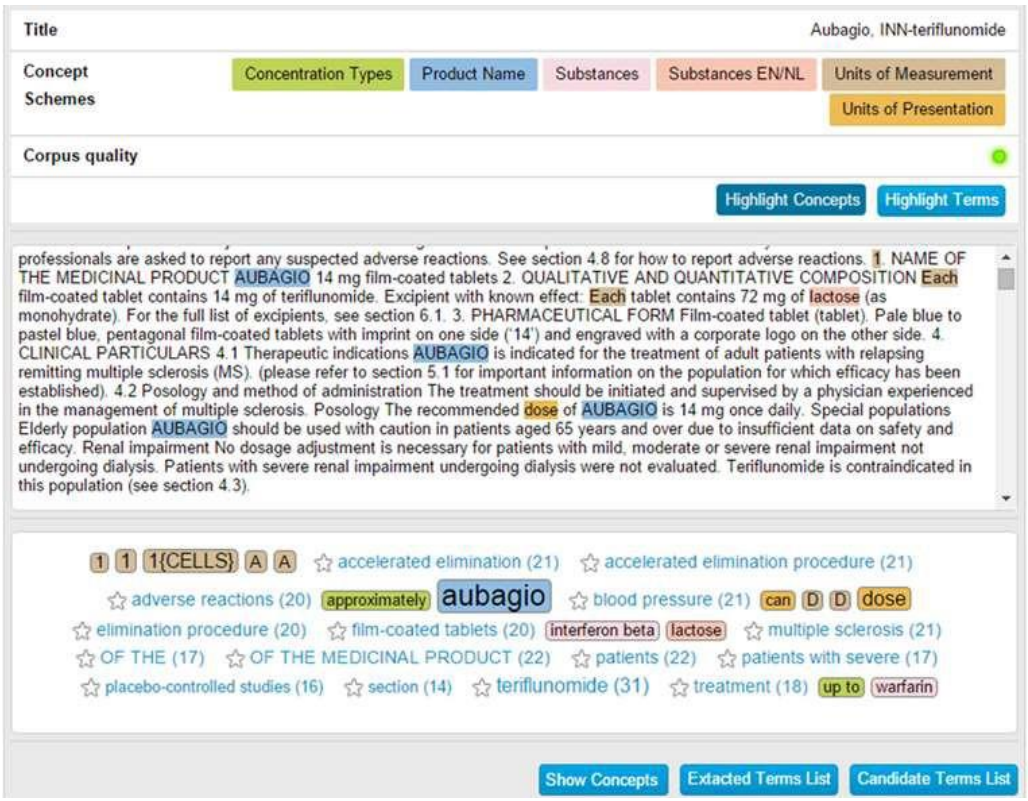

Figure 10. A screenshot of PoolParty Extractor™ recognizing terms in different controlled vocabularies (highlighted in matching colors). Printed with permission from SWC, which is gratefully acknowledged.

Figure 10. A screenshot of PoolParty Extractor™ recognizing terms in different controlled vocabularies (highlighted in matching colors). Printed with permission from SWC, which is gratefully acknowledged.

The benefits of having a Linked Data reference data repository in the enterprise go beyond reference data management itself. Such a repository serves as a trove of deep knowledge encoded in disparate vocabularies. Because they are all present in one spot in a form that ensures technical and semantic interoperability, one can actively query these vocabularies: "Give me all concepts from all vocabularies where "carcinoma" occurs in the preferred or one of the alternative labels".

The repository can easily grow into a semantic hub that functions as the integration backbone of your enterprise. Commercial, large-scale, enterprise-grade platforms driven by Linked Data standards are now available that are designed from the ground up to provide exactly that enterprise service. Because of the high degree of standardization, there is no danger of vendor lock-in: solutions can easily combine modules from different vendors.

Such a semantic hub not only supports low-cost, high-quality integration between legacy systems. It also provides a foundation from which text management solutions can be driven. This includes tools for entity extraction, which construct

datasets from text, in an automated or semi-automated way. See Figure 9. Another set of solutions that can be driven from such a reference data repository are products for structured authoring — that is, text processing solutions that help authors to use reference data in running text. The rigorously standardized approach that ontologies like SKOS can offer make for unprecedented levels of flexibility when connecting these.

These topics warrant more discussion but are outside our scope. Suffice it to say that the strategic outlook that a Linked Data approach to reference data management offers is in line with many contemporary developments.

4 Central Takeaways

The central takeaways of this whitepaper are:

- The importance of reference data increases as data replace text
- Existing systems must allow for upgrading the reference datasets they use without making a new release of the software necessary
- Reference data must be governed and managed centrally
- Reference data management comprises several business processes
- These require proper tooling, based on a central repository
- SKOS is a W3C Linked Data standard that is particularly relevant for reference data management for four reasons:
    - As a W3C Web-standard, SKOS is rigorously formalized for optimal interoperability and maximal reuse
    - It captures the basic notions of how an individual vocabulary item is described (preferred label, alternative labels, notation, definition, multilingualism, et cetera)
    - It captures semantic relations that often occur between vocabulary items inside a reference dataset, such as the all-important hierarchical “broader” relation
    - It captures semantic relations between vocabulary items in different reference datasets, such as “closeMatch” and “exactMatch”, which are necessary for defining crosswalks that map different reference datasets to each other

5 Acknowledgments

An early version of this paper was presented as part of a tutorial at the DIA Regulatory Submissions, Information and Document Management Forum conference, 8-10 February 2016, North Bethesda, USA. The materials presented there, of which the present paper is a part, are the result of pleasant and intensive cooperation between H. van Bruggen and M. Stam of eCTDconsultancy, and the author. I thank the DIA crowd for offering the opportunity to present these ideas, the many discussions we had. I am indebted to Venkatraman Balasubramanian (Cabeus), Lior Keet (Highpoint Solutions) and Vahé Ghahraman (Alexion Pharmaceuticals) for thoughtful comments on earlier versions of this paper.

Some of the material in this paper was developed as part of a Webinar presented by Antoine Isaac (Vrije Universiteit Amsterdam), Martin Kaltenböck & Andreas Blumauer (Semantic Web Company) and the present author, entitled as Vocabularies as a Service.

I am indebted for ideas, commentaries and inspiration to many. In the first place, I would like to thank Ralph Hodgson and Irene Polikof of TopQuadrant, and Andreas Blumauer and dr. Christian Blaschke of Semantic Web Company, not only for making their technology and APIs available for the research underlying this paper, but also for actively engaging in discussing ideas, providing support, insight and inspiration and solving problems.

Finally, I am indebted for ideas, feedback, discussion and comments to Richard Nagelmaeker (Blue Sky), John Walker (Semaku), Antoine Isaac (Vrije Universiteit Amsterdam), Jos van den Oever (MinBZK/Koop, Logius), Laura Daniele (TNO), Daniel Woning (Isala Hospital) and Linda van den Brink (Geonovum)

Of course, any remaining errors are my own.

About Taxonic

We are Taxonic, an international and independent IT consulting firm. Our focus? Digital transformation and Knowledge Management, which we apply within two areas: Linked Data and Pega (Dynamic Case Management).

As a knowledge partner, we share our expertise, provide tailored advice, and are hands-on in projects from analysis to implementation. Our experts work with the latest technologies. Within the team, we have a broad range of experience; from advising on technology selection and integrating solutions into the technical environment to connecting the system landscape and complete implementation projects.

We offer a full end-to-end service that allows us to guide digital transformations from design to implementation, ongoing development, and maintenance. Additionally, we are actively involved in developing our own software products. We create innovative software solutions based on Linked Data and semantic technology.

Working with interdisciplinary teams enables us to tackle a wide spectrum of complex projects. Taxonic encourages the world to adapt to new technologies, and this is our contribution to society.

About Jan Voskuil

With a background as a linguist, Jan understands the concept structure and the importance of Linked Data and SKOS like no other. He has, therefore, been a strong advocate for the development of Linked Data in the Netherlands driven by his content-driven motivation. Jan enjoys reading and has an interest in geology, which is not an unusual combination for an information architect.

Jan is a "thought leader" with several publications to his name on Semantic technology and Linked Data. He has also put SKOS on the map in the Netherlands by advocating for its inclusion in the 'comply or explain' list.

Veldzigt 2
3454 PW De Meern
The Netherlands

T +31 (0)88 829 66 00
E info@taxonic.com
W www.taxonic.com

bank NL45RABO0161959660
VAT NL851339803B01
CoC 54529190