2025 IMEKO TC-6 International Conference on
Metrology and Digital Transformation - **M4DConf 2025**
Benevento, Italy, September 3-5, 2025

An Overview of Metrology Knowledge Storage: Taxonomies, Ontologies and Constrained Vocabularies

Clifford Brown[1], Julia Neumann[2]

[1] *PTB, Berlin, clifford.brown@ptb.de*
[2] *PTB Berlin, julia.neumann@ptb.de*

***Abstract* – Within metrology new and innovative digital ways of working are being implemented that can offer great benefits to the future science, technology and administration of the area. How metrological knowledge is stored and hence processed will play an important role in maximizing the benefits of digitalisation. In this work a high level overview of current and future ways of storing scientific knowledge, the advantages and disadvantages are discussed. Finally, a potential optimal route to scientific knowledge storage for metrology is proposed.**

I. INTRODUCTION

The administrative processes and procedures that support Metrological Science and Technology are now evolving into a digital ways of working [1]. A core example of this is provided by the development of Digital Calibration Certificates as a means to replace paper based ('analogue') certificates (SmartCom [2]). Although the evolution of digital methods will not immediately provide new science for metrology, the future need to collect and collate more meta data related to the science of measurement processes and procedures will provide the possibility to interconnect scientific information in ways that have not been possible before. Using AI and ML methods to analyse this information can lead to new scientific discoveries for metrology.

It is important to state that there are significant immediate benefits of digital ways of working in metrology including: improved speed and quality of communication and transcription of data; the ability to store additional meta-data beyond what is currently stored and the machine readability for metrology communications.

An important aspect of collecting information (of any kind) is how to store it in a digital way that retains all the benefits and qualities of the information for now and into the future. Unfortunately, data, and for metrology scientific data, manages to stay 'alive' for significant lengths of time [3]. Even digital data struggles to stay active indefinitely. Without the relevant meta-data that provides a clear description of the data being stored alongside the core scientific data, data tends to become unusable even if it is findable. The FAIR principles [4] provide a number of important and clearly defined ways to keep data relevant, extending the life of data, retaining its value into the future..

II. TRADITIONAL VERSUS MODERN METHODS FOR STORING SCIENTIFIC INFORMATION

In general, a Computerised Information System (CIS) of any kind (be it a traditional localised system e.g. Excel, to a new AI system generating radically new information) is fundamentally made up of two parts: data used (either held locally or obtained from some form of distributed storage system); and the process used to manipulate that data in some way to generate some form of new, additional information which is typically displayed to the User of stored somewhere for future use.

In creating a new, quality assured computerised system there is a tendency to assume that the input data is a already given, defined, quality assured and fixed, and that the major work to be performed in generation of the new CIS revolves around the software development of the process to manipulate the data. There is no question that the processing part is very important and does require significant effort and testing to make sure it functions correctly. However, the problem tends to be, that the data aspect is put to one side until after the process part has been created. Only relatively small quantities of data tends to be used in testing a CIS. But, the quality and hence the capability of the dataset as a whole, is a significant factor in the creation of a successful and quality CIS [Ref]. This type of issue tends to be more relevant when a traditional, functional decomposition of the core problem is considered.

Another way to develop CIS systems is to apply the object-oriented (OO) paradigm to the core problem [5], and is known as Object-Oriented Analysis and Design

(OOAD). The Unified Software Development Process (USDP [6]) is an example of an OOAD software engineering practice that encompasses the Agile style of CIS development. In this technique the data as well as the processing part is considered together in developing the system. The core problem is broken down into a hierarchical set of interacting objects, where each object has its own data and processing capabilities . The OOAD provides a data model that can automatically be converted into a relational database structure, or taxonomy (Ref wiki taxonomy [7]). A taxonomy for a structure that is well defined follow a number of self referential rules [8]. One of the most important rules states that any piece of information should only be stored in one location, and this ensures that the data held is self consistent and a good taxonomical design makes it more robust against accidental corruption.

III. TAXONOMIES

As mentioned above the OOAD methodology performed correctly provides an object oriented data model that can automatically be converted into a relational database structure, or taxonomy. A relational database structure or schema consists of a number of two dimensional tables linked together using linking columns of identical information between the tables. The data information held in a row of a table typically describes the data attributes of a single object instantiation of a class from an OO model. An individual table with a relational schema or structure describes the full details of a particular core concept within the functionality or purpose of the system. For example, in metrology, a measurement would be an important concept. The relevant data would probably include: the magnitude or size of the measurement as a decimal value, and the unit associated with the measurement either as a string or maybe a link to another table containing a list of units. (Figure ?)

Functional breakdown methodologies can also lead to taxonomies describing the underlying data but the process of determining the structure requires an additional process to be performed and the output will need to be made consistent with the functional decomposition. Frequently the needs of the underlying data and the functional decomposition contain contradictions requiring a round of modifications to be made that tend to lead to results based on undesirable compromise solutions.

Taxonomies are ubiquitous structures for storing scientific information and are the most common method used for storing information. For example, when collecting data from a fixed in time experiment that, by definition, is limited in concept, not requiring additional, new types of information to be stored. Referential databases can be queried using Structured Query Language (SQL). SQL is a logic-based language using the names of the main concepts in a system whose information is being stored in a referential database.

IV.

V. TAXONOMY ISSUES

Apart from the process dependent issues of determining a good taxonomical structure that defines the scientific information for a system in a meaningful and effective way, inevitably any taxonomy will be limited regardless of how it is generated. Even well-defined taxonomies are invariably limited in time. Science is an ever-evolving subject with new ideas and associated data being uncovered continuously. Occasionally data models can be extended in a simple fashion to include new types of data for a system, but invariably a major revision of a taxonomy is required to include new types of information. Also, the code behind the taxonomy will need to be updated to work with the new data taxonomy. This tends to be a time and resource expensive procedure and is only undertaken when there is scientific justification and associated funds to do so. In addition, most likely, there will also be the need to provide some form of backwards compatibility with the old data structure.

VI. ONTOLOGIES

Ontologies are very different to a taxonomies. It stores information in an unstructured way (Ref my paper last year). In theory, you can store any information in a single ontology; even information that is totally disconnected. An ontology is infinitely flexible in terms of the information it can store. This has advantages and disadvantages.

Ontologies work using the mathematics of First Order Logic (Ref. Horrocks [9]) which allows for simple statements such as:

“Measurement 1” “has Magnitude” “0.579”

“Measurement 1” “has Unit” “kilogram

to be stored in an instance of an ontology.

This triad, or ‘triplet’ of information, “subject”, “predicate”, “object” where “subject” and “object” are effectively nouns describing “things” in system of interest, and the “predicate” is a “verb” or action that connects the two objects. Using this reasoning, it is completely possible to store all the information that exists in an example taxonomy in an equivalent ontology which is literally and simply a list of triples. However, where ontologies exceed the capability of taxonomies is in the fact that it is self-describing and hence completely extendable. There is no limit on how to extend the ontology. So, if new

information needs to be stored, the pattern for the extension is stored in the ontology, as a triple, and then the new type of information is also added, as triples.

This ability to store new information with minimal effort would appear to be the perfect solution for storing ever-changing scientific information. However, there are problems associated with the ontological solution.

VII. ONTOLOGY ISSUES – CONCEPT DRIFT

The determination of a taxonomy to describe the data associated with an information system leads to a group of linked table structures. These table structures typically represent core concepts with the problem space. They are necessary and sufficient to describe the problem under consideration; they represent implementation of the important nouns in a requirements document for the system.

Ontologies do not, in principle, need to use concepts to operate. As mentioned above information in the form of triplets can be added to an ontology simply by adding them to the full list of triplets. However, the resulting ontology will be disconnected, random facts and will not be able to generate any interesting results other than the basic information it stores. Structured storage like taxonomies, which use core concepts, join information together within the linked table structures, and can, using SQL, generate interesting information beyond the basic facts input. So, the main benefit of Ontologies, i.e. their capability to store any type of information, is also the source of its main weakness.

The middle ground is to impose the use of concepts within an ontology. By using the same methods of discovering the concepts for a taxonomy from the problem requirements, the concept (or class) meta information can be stored within the ontology. The concepts, or classes, of information can then be used, (almost like a pastry cutter to create multiple biscuits for cooking), to create actual instances of pertinent objects. In the example shown earlier, the concept of Measurement could be created with two attributes "Magnitude" and "Unit", and Measurement1 would be ('is a') an instantiation of the Measurement class.

An ontology can therefore provide the ability to evolve as new information becomes available and can embody concepts to group connected information together into higher level meaningful entities, or objects. Probably the most appropriate storage system for scientific knowledge. However, this potentially ideal storage system is compromised by the reality and complexity of how scientific knowledge evolves. Scientific knowledge is filled with contradictions. Even some of the most successful theories e.g. relativity and quantum mechanics in particle physics do not complement each other cohesively.

Even something as well defined as the systems of units for measurements suffers from knowledge uncertainty. The QuDT ontology system of units [10] developed originally by NASA out of a satellite failure due to the use of inconsistent units, arguably the most developed Unit Ontology has developed in complexity to the point that its use requires significant research into the unit concepts it contains. This research tends to uncover multiple versions of complex compound units which might be considered to be the same concept but contain subtle differences in their definitions. In this way concepts become blurred due to multiple interpretations resulting in confusion and users sometimes deciding (using the 'Open World Assumption' Ref. [11]) to create even further versions of a concept to ensure it is what they need and understand. This, concept drift tends to bloat ontologies almost to the point of becoming unusable. However, this situation reflects the normal reality of the evolution of scientific knowledge, this variation of ideas is necessary, but can lead to uncontrolled complexity.

VIII. HYBRID METHOD: TAXONOMY, ONTOLOGY AND CONSTRAINED VOCABULARY

Dealing with the complexity of concepts that are used within a problem space as described above can lead to ontologies becoming bloated to the point where their use becomes difficult. The 'Open World Assumption', (OWA, Ref.[11]) where in principle 'anyone can say anything about anything', is an idealistic concept that was developed in order to allow ontologies to evolve and so contain ever changing new information. How to manage this complexity whilst still allowing for the flexibility that ontologies provide is the challenge. Constrained Vocabularies (Ref [12]) provide one way to do this.

In a constrained vocabulary the core concepts of a problem space which are well-defined but fixed are referred to as a 'Closed World Assumption' (CWA Ref. [11]). In this approach the information required to understand and process a problem space is assumed to be well known, self-consistent and complete. There is no need to have further knowledge concepts added to the system and in this respect is more like a taxonomy. However, this approach still leaves open the opportunity to store and treat this information semantically in an ontology: allowing the flexibility to interconnect knowledge in new ways. The closed nature of constrained vocabularies controls and manages the complexity for problem spaces, but does not allow new concepts to be automatically added to the system.

An intermediate state between the CWA and OWA is the Partial Closed World Assumption (PCWA). In this case, the problem space is divided into two states: for situations where knowledge is considered well understood and complete, the CWA is used; for other cases the OWA is applied allowing new knowledge concepts to be added. In this way the PCWA approach provides a way to control complexity but still allow for concept innovation in areas of developing science and technology. For metrology, the PCWA could therefore provide a sensible but still flexible way to store knowledge. The international SI system of units in its D-SI format [13], is a fundamental cornerstone of metrology, and could be treated using a CWA.

IX. CONCLUSIONS

A number of strategies for storing scientific knowledge including: taxonomies, ontologies and constrained vocabularies have been discussed. Taxonomies provide a well-controlled, closed world assumption (CWA) environment in which to store knowledge from scientific and technical areas that are well understood, but form an expensive option when the concepts need to be modified and or enhanced. Ontologies provide a very flexible route to storing the ever changing knowledge behind scientific discovery, but can suffer from concept drift and uncontrolled complexity growth due to the associated Open World Assumption (OWA). The Partial Closed World Assumption (PCWA) provides a hybrid option: closing and controlling areas (CWA) of science evolution that are considered well understood, but allowing other areas to evolve under the OWA. The PCWA approach could provide the optimal route for storage of scientific and technical knowledge.

REFERENCES

[1] Digital Ways of Working, https://www.bipm.org/en/-/2025-03-12-metrology-in-the-digital-age.

[2] SmartCom, https://www.ptb.de/empir2018/smartcom/project/,

[3] Historical Data, …

[4] FAIR Principles, https://www.go-fair.org/fair-principles/

[5] Object-Oriented Paradigm, https://en.wikipedia.org/wiki/Objectoriented_programming

[6] Unified Software Development Process, https://en.wikipedia.org/wiki/Unified_process

[7] Taxonomy, https://en.wikipedia.org/wiki/Taxonomy

[8] "An Introduction to Database Design" C. J. Date ISBN 978-0321197849

[9] Horrocks, First Order Logic, https://www.cin.ufpe.br/~in1006/HorrocksDLIntro.pdf

[10] QuDT, https://qudt.org/

[11] Closed World Assumption, https://en.wikipedia.org/wiki/Closed-world_assumption

[12] Constrained vocabularies, …

[13] D-SI Format, https://si-digital-framework.org/