IMEKO TC6 International Conference on Metrology and Digital Transformation
September 19 – September 21, 2022, Berlin, Germany

ONTOLOGY-BASED REST-APIS FOR MEASUREMENT TERMINOLOGY: GLOSSARIES AS A SERVICE

Michael Chrubasik [a,*], *Chris T.S. Lorch* [a], *Paul M. Duncan* [a]

[a] Data Science, National Physical Laboratory, Glasgow, UK,
* Corresponding author michael.chrubasik@npl.co.uk

Abstract – Even in today's connected world of measurement, organisations, NMIs and stakeholders across a multitude of disciplines employ their own specialised terminology to convey information relating to measurement, experimentation, and projects. While the terminology used is often governed by existing standards or industry norms, the overlaps between these standards and the breadth and depth of applied language frequently result in confusing terminology landscapes. This introduces significant difficulties for cross-disciplinary and cross-industry collaboration. We present a framework to record semantic information in discipline-specific parlance within a flat glossary-like ontology and make it accessible through a RESTful Application Programming Interface (API). This paper outlines the requirements to enable such operations: from vocabulary generation, transformation to ontology, and final exposure via an API. The methodology was exemplified using a pharmaceutical industry-aligned use case, whereby communication between several pharmaceutical industry partners and academic institutions was facilitated.

Keywords: ontology, RESTful, terminology, reproducibility, collaboration

1. INTRODUCTION

National Measurement Institutions (NMIs), alongside their partners and other third-party associates, cover a wide range of scientific areas. Each member within these areas applies their own specific language, terminology, and jargon in communicating concepts, which leads to difficulties in inter-industry collaboration. Although institutions and companies have internal and project specific glossaries, the collation and distribution of this information to partners is still cumbersome and inefficient.

Difficulties interlinking company-specific languages arise due to the lack of machine-operable accessibility to the relevant standards and corporations' internal lexica. For this reason, the existence of machine-operable and well-established vocabularies has applications in a wide range of industrial, academic, and metrological arenas [3][4]. Without controlled vocabularies that allow for the inclusion of rich semantic information, such as synonyms, sources etc., translation and miscommunication become a significant bottleneck to large collaborative projects. Ensuring interoperability by removing the ambiguity associated with company-specific vocabulary can boost research output through the smooth exchange of data, reduce costs associated with labour intensive vocabulary mapping [6], and can improve product quality through the mitigation of errors [5].

The National Physical Laboratory (NPL), has been exploring the use of knowledge bases for domain-agnostic measurement ontologies [**1**] as a method to capture and inform on measurement data within the organisation. These techniques have also been applied externally, with partners at the Medicines Manufacturing Innovation Centre (MMIC), by leading the construction of ontologies for use in pharmaceutical approval processes. The MMIC is a collaborative centre working across major pharmaceutical manufacturers (AstraZeneca, GSK, Pfizer) as well as academic and technological organisations to provide industry-wide innovation [7]. Whereas previously individual companies developed their own solutions, each party works together in developing both company and sector-wide solutions. It therefore follows that a controlled vocabulary is essential to allow each member to participate in this collaborative process. NPL is working with the MMIC to aid in the creation of a dashboard which would control all quality information required to approve the release of pharmaceutical products from clinical trials for human consumption. As the consequences of releasing substandard drugs are extremely high, this is a highly regulated and complex area.

The pharmaceutical industry – mirroring the measurement sciences in this respect – has a plethora of standards and vocabularies to guide the use of terminology. It is clear however, that this abundance of jargon introduces congestion in communication at the inter-organisational and scientific level. This leads to miscommunication issues and subsequent time costs.

We present a framework to record science-area-specific semantic information and idioms within a flat glossary-like ontology and make that information accessible through a RESTful Application Programming Interface (API), also referred to as a restAPI. The MMIC "Qualified Persons Dashboard" project, a collaborative project involving many pharmaceutical industry partners, organisations, and academic institutions, represents state-of-the-art research and collaboration in the field of digital pharmaceutical manufacturing assurance and compliance. However, as alluded to, communication and collective agreement on terms between parties has been challenging due to terminological ambiguities and tough harmonisation challenges.

This paper outlines the requirements to facilitate such operations: from generation of a vocabulary, over transformation of said vocabulary into a flat glossary-like ontology, to exposition of the ontology via an API. Furthermore, we present our industry-aligned use case via a

web-front-end dictionary that shows direct application of the ontology-based restAPI as well as the wider industrial use case. Finally, we discuss future potential applications of this methodology.

2. METHOD

Our methodology was based on a four-step process. Steps 1 and 2 pertain to the construction of the glossary and are graphically illustrated in Fig 1., while steps 3 and 4 define the API data link and front-end dictionary application, illustrated in Fig 3.

1. Through guided exposition by domain experts, a centralised glossary was agreed upon and established in a collaborative environment.
2. The glossary was validated by the ontology engineers and programmatically transformed into an ontology format.
3. The resulting ontology was exposed to users through a Python-based API that handles queries and delivers query results in HTTP form to be consumed by any third-party service. This exposed ontology now represents a centralised and controlled vocabulary, essentially delivering terms as a service.
4. A basic web-based dictionary application allows parties to query project terms and understand how organisational terms are related to those through a user-friendly interface. Terms can also be identified as organisation-specific or as grounded in existing standards. Aligning their terminologies allows collaborators to readily navigate ambiguities and communication issues.

2.1. Building a glossary

Institutions and organisations operate based on established industry or discipline-specific parlance. As such, creating and maintaining controlled vocabularies is critical for effective communication, particularly when addressing disparities between organisation-specific terminology. This becomes paramount when projects include multiple parties in cross-industry collaboration. Glossaries aid in collaborations as they not only include terms and definitions, but also acronyms, synonyms, term sources, etc. For larger vocabularies, additional questions of governance and stewardship arise concerning responsibilities for upkeep, change and maintenance of the glossary. At this stage a well-organised and planned approach to glossary development was required.

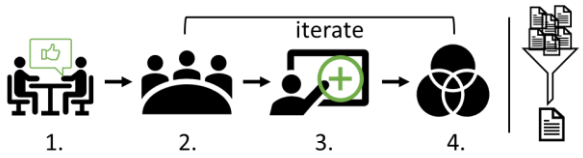

Fig. 1. The glossary-building framework method: (1) All parties agree upon the necessity of a centralised vocabulary, identify central themes and nominate a committee of domain experts. (2) Terminology is decided upon by a committee of domain experts based on the central themes. (3) The core terms are approved and appended to the glossary through an iterative process. (4) End-of-iteration harmonisation is performed and additional information, e.g. synonyms and term sources, are added. Iteration terminates once a final consensus is reached.

For the purpose of our use case with the MMIC, we followed an iterative framework of collaborative creation with our partners as demonstrated in Figure 1. It is critical not to overextend the vocabulary beyond the central themes and to filter out unnecessary additions.

There are many commercially available glossary creation tools, such as the IBM InfoSphere [11], however, due to time restrictions a basic spreadsheet-based approach was chosen, which would make transfer to an ontological format easier in future steps. Using this approach, each unique term was depicted in a row with its descriptors and metadata being described as columns. Glossary creation is a particularly lengthy process and should be initiated early on in a large-scale project. The fluid nature of vocabularies means changes are to be expected, however even early iterations of the glossary can be useful and highlight significant issues in project communication.

A glossary specification document was generated which identified 10 unique columns to describe terms for the vocabulary, as shown in Table 1.

Table 1. Unique column descriptors for terms in the glossary.

Column	Description
ID	A unique identifier for all available terms in the glossary
base term	Agreed base term that will be displayed in the "Qualified Person Dashboard"
base definition	Definition for the agreed base term
abbreviation	Common abbreviation for the term
term source	Source where the term originated from, if known (e.g., standard)
related terms	Other terms directly related to the base term
category	Descriptive term for usage context
stewardship	Descriptor for ownership and responsibility
status	Descriptor for whether term has been agreed upon or is under discussion
comment	Property allowing parties to raise concerns and comments
+ optional organisation specific terms	
term synonym	Collaborator synonym terms
term definition	Synonym term definition
term source	Synonym term source

Applying this specification document and framework allowed us to establish a routine operation for the generation of this glossary. Our early iteration currently encompasses 60 unique terms based on the pharmaceutical drug approval process, largely built on terms that all companies will adhere to within the "Qualified Persons Dashboard". Due to the breadth of pharmaceutical standards involved in this work, harmonisation of identified standards, such as "ISO 11238 (2018) Health informatics"[15], was guided by domain experts.

2.2. Creating an "ontology"

Ontologies categorise and represent entities and concepts along with their interdependent properties and relations. This work does not qualify as an ontology based on the previous definition, however it establishes a foundation from which a more formal ontology can be established in the future.

Many highly regarded libraries, e.g SKOS (Simple Knowledge Organization System) [12] and DCTERMS (Dublin Core Metadata Element Set) [13], are available in the

knowledge management space that aid in constructing knowledge organisation systems, e.g. vocabularies and taxonomies. Employing these recommendations in aligning and describing metadata and column properties improves access and readability of the vocabulary. Furthermore, transferring to an ontology format enables a less complex presentation of information, simplifying findability and accessibility of the data and in particular, making it machine operable.

These resources allow us to describe unique terms as classes with their relationships and metadata as properties, building semantic links and increasing richness of the available information.

2.3. Creation of a restAPI

A restAPI is an API that adheres to the REST architectural style and enables interaction with RESTful web services. REST is an acronym for Representational State Transfer and embodies a set of architectural constraints [2]. A restAPI obtains resources via commands, utilising established HTTP protocols such as "GET", which retrieves a resource from the API [2].

The concept of restAPIs is widely regarded as the *de facto* standard protocol for web APIs due to their ease of use and widespread availability. Open-source frameworks like Swagger [9] exist to greatly alleviate the time-consuming task of building RESTful interfaces. Swagger is a collection of open-source tools, derived from the OpenAPI specification, that assist in the design, development, documentation, and consumption of restAPIs. The Swagger editor allows one to define their API specification and export the result as a functional server, based on the programming language required.

For our use case, we implemented a Python server, specifically using the Flask Python web framework [14]. Flask offers tools, libraries, and technologies for developing web applications. Once this restAPI server was established, answers to queries could be served.

In linking the API with the web service, one must initially stage the ontology using the *Owlready2* [9] Python library within the controller file containing the extraction format definition. The controller file was automatically generated through the swagger editor tool based on the API specification. Then, employing *Owlready2* such that term rows and columns can be flexibly translated, the controller file was tailored to a specific use-case. The script was made to dynamically adapt to new iterations of the glossary, thus reducing engineering time spent on validation checks.

2.4. restAPI consumption: web application

In technical terms, restAPIs are "consumed" by applications that use their information. To exemplify this functionality of the API a web-application was developed using the Flask Python web framework which allows for quick deployment as well as flexible and dynamic Python-based web content delivery.

Web pages are typically split into two parts: the back-end service generates quasi-dynamic HTTP request based on the nature of the user-requested information and consumes the API via the available *requests* [10] Python library. Requested information is passed back up the chain to be reformatted for its intended use within the front end. The front-end renders information to the user and lets them interact with it. When using Flask, the framework is also responsible for "hosting", i.e. delivering the front-end to the user. A basic representation of this can be found in Figure 2.

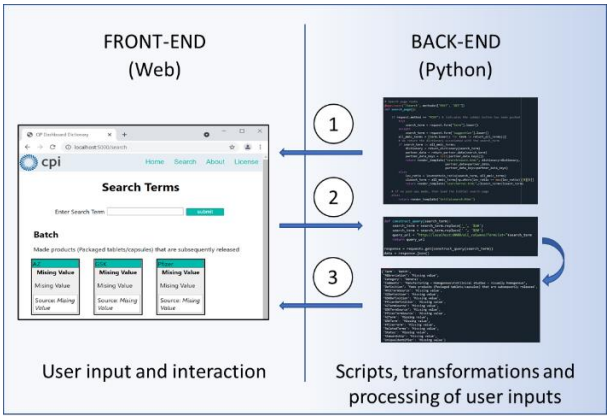

Fig. 2. A basic representation of front- vs back-end: 1. Python-based hosting generating a website and allowing user access. 2. User request for data. 3. Querying of data via Python script and API followed by output of query to website.

3. RESULTS

We present our use case in the form of a dictionary web application which by applying the accessible method discussed in Section 2. This serves as a precursor to a direct glossary implementation within the "Qualified Persons Dashboard". Figure 3 below describes the entirety of our results and the interplay between the framework steps.

Throughout development of the restAPI, the usefulness of a more general dictionary-style web application became apparent not only as a use case for the API but also as a general tool for future use. The dictionary application was chosen and designed to enable satisfactory querying of the terms within the glossary and highlight ambiguities, inter-organisational synonyms and potential problems that could arise from related technical jargon.

Implemented through the Flask framework, a small number of functional web pages were established that describe the project and allow for user input to the querying and exploration of the terms contained within the glossary.

Porting of the glossary into the ontology was achieved via a "data connector" tailored to our specific use case vocabulary. While the "data connector" script was specific to the vocabulary employed in this project, the methodology is universally applicable and renders itself useful to similar efforts in the measurement and metrology space currently under way at NPL, such as alluded to in our previous publication on Domain Agnostic Measurement Ontologies [1]. Provenance of terms and versioning issues were encountered at this stage, suggesting a web-ontology tool, such as Web Protégé, might be more suited to the task.

In this iteration of work, users are limited to "GET" HTTP requests only. Updating and changing terms within the controlled vocabulary was a privilege only afforded to the maintainer and was achieved through the "data connector" script. By using this approach, we were able to achieve the level of security required within the use case, and whilst a login feature could later be implemented, it was beyond the scope of this project.

With all required files and connectors established, the Flask API server was shown to serve any third-party application with access to the API and provide users with answers to the available "GET" queries. This is a significant improvement to typical document-based vocabularies. The exposition of the contained terms through the restAPI permits direct programmatic querying of term information as well as enabling machine-readable and operable interaction with the contents of the ontology.

There is a large breadth of exploitation avenues to such available glossaries through direct implementation in applications, such as the "Qualified Persons Dashboard" but also text editors, natural language processors and many more. Furthermore, the base terminology layer represents a great foundation for higher order ontology creation as the controlled vocabulary has already been established.

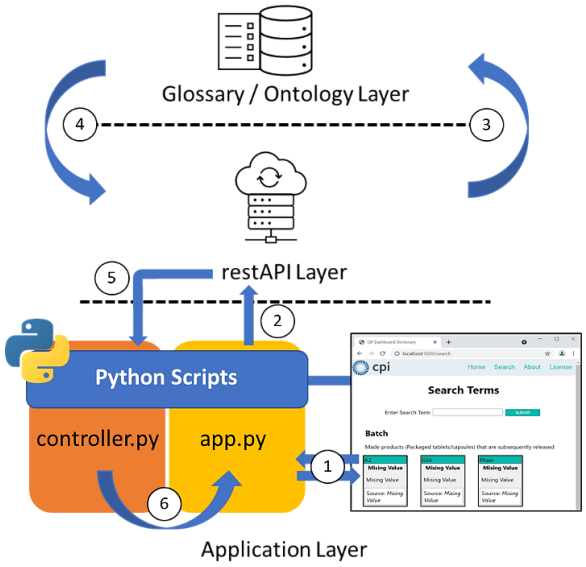

Fig. 3. Framework layer interplay. (1) A user input is sent from the web service front-end to the back-end. (2) Based on the query, an HTTP request is generated within the app.py flask file, which consumes the API. (3) The restAPI retrieves the information from the request and passes it back up the chain (4)-(5) as outlined in the API specification. (6) The requested information is finally passed from the controller file to the app where it undergoes a final reformatting before being rendered in the user front end.

4. CONCLUSIONS

We have presented a framework to record science-area-specific semantic information and jargon within a flat glossary-like ontology to then make that information accessible through a RESTful application programming interface.

The proposed ontology-based restAPI powered centralised vocabulary has the potential to satisfy all project partners and collaborators by providing the most appropriate term via a substitution and/or query interface. This has been exemplified by our industrial use case, which demonstrates the methods potential in providing an avenue into frictionless interoperability.

While the results above focus on a specific use case from pharmaceutical industry, similar scenarios can occur in any large scientific collaboration. In the advent of digital collaborations and product generation, harmonisation efforts will only increase. Abstracting the presented findings to measurement sciences and the application in interdisciplinary and inter-NMI projects, machine-operable "glossaries and terminologies as a service" could be highly beneficial to increase accessibility to technical and specific jargon. It may also benefit the collaboration of such stakeholders through simplification of jargon-alignment, in particular when projects involve external and international organisations.

ACKNOWLEDGMENTS

This work was funded by the Department for Business, Energy & Industrial Strategy through the UK's National Measurement System. The authors would further like to thank our collaboration partners at the Medicines Manufacturing Innovation Centre consisting of our partners at CPI, University of Strathclyde, UK Research & Innovation, Scottish Enterprise and founding industry partners, AstraZeneca and GSK.

REFERENCES

[1] J-L. Hippolyte, M. Chrubasik, F. Brochu & M. Bevilacqua, "A domain-agnostic ontology for unified metrology data management", Measurement: Sensors, Vol 18, 2021.

[2] Roy Thomas, Fielding. Architectural Styles and the Design of Network-based Software Architectures. Doctoral dissertation, University of California, Irvine, 2000.

[3] Gentile, A.L, Gruhl, D., Ristoski, P. & Welch, S. (2019), Personalized Knowledge Graphs for the Pharmaceutical Domain. In The Semantic Web -- ISWC 2019.

[4] Gansel, X., Mary, M., & van Belkum, A. (2019), Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. European Journal of Clinical Microbiology & Infectious Diseases, 38. pp 1023-1034

[5] Kaelber, D.C., & Bates, D.W. (2007), Health information exchange and patient safety. J. Biomed. Inform, 40.

[6] Lin, M.C., Vreeman, D.J., McDonald, C.J., & Huff, S.M. (2011) A characterisation of Local LOINC Mapping for Laboratory Tests in Three Large Institutions. Methods Inf. Med. 50(02), pp 105-114.

[7] UK CPI (2019), Medicines Manufacturing Innovation Centre project to deliver Just in Time clinical trials. URL: https://www.uk-cpi.com/news/medicines-manufacturing-innovation-centre-project-reaches-next-stage-in-delivering-just-in-time-clinical-trial

[8] SmartBear Software (2021) Swagger Editor Documentation. URL: https://swagger.io/docs/open-source-tools/swagger-editor/

[9] Lamy, J.B. (2019) Welcome to Owlready2's documentation!, URL: https://owlready2.readthedocs.io/en/v0.36/

[10] Reitz, K. (2019) Requests: HTTP for Humans, URL: https://docs.python-requests.org/en/latest/

[11] IBM InfoSphere Information Governance Catalog, (2022) IBM, URL: https://www.ibm.com/us-en/marketplace/information-governance-catalog

[12] Isaac, A., & Summers, E. (2009), SKOS Simple Knowledge Organization System Primer, URL: https://www.w3.org/TR/skos-primer/

[13] Rühle, S., Baker, T., & Johnston, P., (2022), Dublin Core Metadata Initiative, URL: https://www.dublincore.org/resources/userguide/publishing_metadata/

[14] Pallets, (2010) Flask: Web development one drop at a time, URL: https://flask.palletsprojects.com/en/2.0.x/

[15] ISO 11238 (2018). Health informatics — Identification of medicinal products — Data elements and structures for the unique identification and exchange of regulated information on substances. ISO