FAIR Vocabulary Management at Scale: TERN's Implementation for Ecological Data Integration

Junrong Yu[1], Javier Sanchez Gonzalez[1], Edmond Chuc[2] and Siddeswara Mayura Guru[1]

[1]The University of Queensland, Indooroopilly, QLD, 4068, Australia
[2]KurrawongAI, 72 Yundah St, Shorncliffe, QLD 4017 Australia

1. Introduction

The Terrestrial Ecosystem Research Network (TERN) is Australia's National Collaborative Research Infrastructure for collecting, collating, and publishing key terrestrial ecosystem parameters across space and time. TERN observes the ecosystem across multiple scales using satellite remote sensing, drones, in-situ sensors, and human observations. In addition, TERN also publishes data from partnering institutes. Therefore, data management practices deal with heterogeneous data. Hence, harmonising data from various sources is a challenge.

Most of the datasets published will have parameter names without associated meaning. The same parameters are occasionally measured at different scales; for example, vegetation structure is derived from satellite remote sensing as well as human observations. Therefore, users need to understand the intricacies of the data before using them. One of the approaches to providing better information about the data is to describe them using standard terms with accurate definitions. This will improve the ability to compare data from multiple sources and for analysis [1]. Machine-readable controlled vocabularies are the foundation for semantic interoperability, where data exchange can happen with a shared understanding of the meaning of the terms used. While existing ecological vocabularies like Darwin Core focus mainly on species occurrences, no vocabularies existed for systematic field surveys that include metadata attributes for sites, site visits, sampling activities, and observations. Hence, it needs to be developed for wider community reuse.

The paper will describe the development and management of semantic-enabled controlled vocabularies for broader ecosystem science community. Furthermore, the paper will provide an overview of different vocabularies developed and their use in downstream applications.

2. Background

TERN uses controlled vocabularies to describe and represent all data-related artefacts. These include platforms, sensors/instruments, observable properties, methods, people and organisations. Most vocabularies are developed internally, while subsets representing platforms, instruments, and observed properties are imported from authoritative external sources, including GCMD and CF metadata conventions. Controlled Vocabularies are essential digital assets of the TERN data infrastructure and are used to describe data and related artefacts consistently. Hence, vocabulary development and management are crucial for TERN data management strategies to describe, index, and retrieve data-related artefacts.

TERN develops vocabularies to achieve three key objectives: (1) Support consistent machine-readable descriptions of digital artefacts including parameters, feature types, methods, platforms, instruments, observable properties and measurement units; (2) Improve data discoverability through applications

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan
junrong.yu@uq.edu.au (J. Yu); j.sanchezgonzalez@uq.edu.au (J. S. Gonzalez); edmond@kurrawong.ai (E. Chuc); s.guru@uq.edu.au (S. M. Guru)
0000-0002-7091-6538 (J. Yu); 0009-0007-1472-1596 (J. S. Gonzalez); 0000-0002-6047-9864 (E. Chuc); 0000-0002-3903-254X (S. M. Guru)

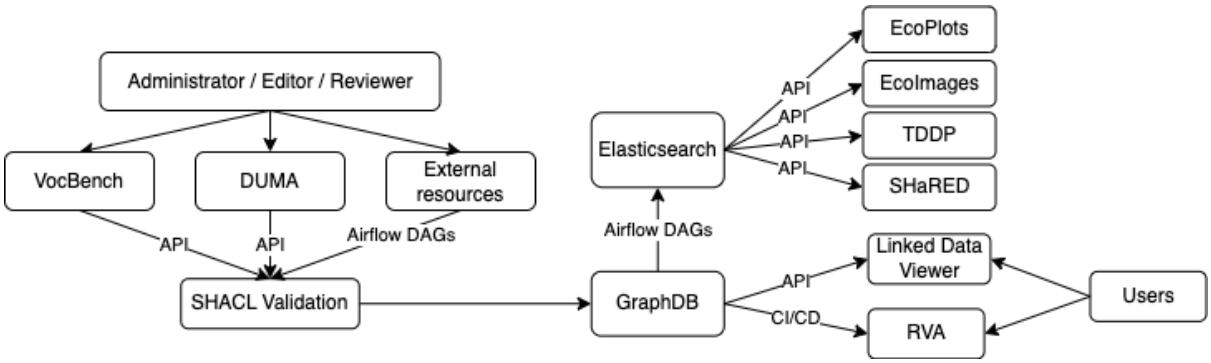

Figure 1: TERN's end-to-end vocabulary management architecture

like the TERN Data Discovery Portal [2], EcoPlots[3] and EcoImages [4]; (3) Facilitate interoperability with other data systems through explicit data representations.

3. Implementation and Technical Architecture

TERN Data infrastructure creates and curates two kinds of vocabularies: SKOS-based vocabularies including concepts and concept schemes [5]; Instances of ontology classes defined in the TERN Ontology [6]. Most of the vocabularies are SKOS-based. However, if we represent an instance of data artefacts, we describe them as an instance of a class. For example, most platforms, sensors, people and organisations are ontology-based vocabularies.

Our semantic approach addresses common vocabulary management challenges: terminology inconsistencies across projects, manual validation workflows, and limited machine readability [7]. It enables automated data validation, federated search across datasets, and sustainable vocabulary preservation through open services.

The vocabulary infrastructure follows FAIR principles (Findable, Accessible, Interoperable, Reusable) using established guidelines for machine-readable vocabulary development [8]. Our collection spans ecological research activities with significant scale: 20 concept schemes covering different ecological data aspects, 138 collections containing 12,603 concepts, 302 research platforms across Australia and New Zealand, and metadata for 72 organisations. The largest concept scheme covers ecological parameters with 6,470 concepts organised in hierarchies of up to 6 levels.

The technical architecture (Figure 1) uses GraphDB as the RDF triple store for vocabulary storage and retrieval. VocBench 3.0 provides the primary editing interface where ecologists create, modify, and deprecate vocabulary terms through a collaborative workflow with automated SHACL validation and editorial review processes. Approved changes are directly committed to GraphDB with named graph versioning for release management.

DUMA, a React-based application, handles people and organisation vocabularies through REST APIs that writes to GraphDB with integrated SHACL validation. Apache Airflow DAGs automate the synchronisation of external vocabulary sources into our system. TERN linked data viewer [9] provides public access to all vocabularies, with its front end and back end supported by Prez [10], which is developed by KurrawongAI [11]. Approved versions are published to Research Vocabularies Australia (RVA) [12] for broader community access through external APIs.

For downstream applications, all vocabularies are indexed in Elasticsearch through scheduled Airflow workflow DAG that maintain daily synchronisation across the infrastructure. When leveraging existing vocabularies from sources like GCMD or CF standard names, we create local versions linked via exactMatch relationships to retain control while preserving interoperability.

The system tracks editorial status through flags (Draft, Published, Under Revision, Deprecated) with automated provenance documentation for all vocabulary modifications.

4. Applications

The Vocabularies are used in multiple TERN applications. The vocabularies drive data discovery in the TERN Data Discovery Portal (TDDP) [2], a gateway to access all TERN published data. The TDDP will enable users to search based on platforms, instruments, parameters, people and organisations. Users can view all controlled vocabularies from each metadata record. EcoPlots [3], a data integration platform for site-based systematic surveys, uses controlled vocabularies to map source data to standard vocabularies to enable harmonisation and integration. EcoImages [4], an image repository for ecology-based image collections, uses vocabularies to map data sources to drive harmonisation and integration. All TERN vocabularies are available through SHaRED (TERN data submission tool) [13] for data librarians and researchers publishing datasets, with over 750 users from more than 70 organisations using existing vocabularies in their data submission and sometimes, contributing vocabulary terms as well. Vocabularies are integrated with the metadata authoring editor so that data librarians can tag data with the pre-defined list. If none of the controlled lists are suitable, the tool enables them to create a new term, which will be reviewed and published.

All proposed vocabularies are available in RDF format and programmatically accessible for maximum reuse. Several external organisations from Australia's state and federal government agencies reuse TERN feature types and observable properties vocabularies. The Biodiversity Data Repository (BDR) [14], developed in partnership with TERN, utilises TERN controlled vocabularies to represent data they collect from industries and state and federal government agencies. Additionally, TERN has developed controlled vocabularies for the Ecological Monitoring System Australia (EMSA) [15] project, covering various aspects of ecological field surveys. These EMSA vocabularies are employed in field survey datasets to provide essential data context, with datasets hosted by the BDR as one of their primary vocabulary systems. The vocabulary system has saved time by avoiding manual harmonisation of different term labels across projects, with nearly 2800 metadata records and 10 million field observations now linked through standardised vocabularies. Government agencies leverage TERN vocabularies in their data publications and representation, while researchers consistently provide feedback to improve vocabulary-based data search and management. TERN vocabularies are continued to be widely adopted across ecosystem communities, contributing to standardised data practices.

5. Conclusions

TERN's semantic-enabled vocabulary system has significantly improved data representation, discoverability and interoperability in multiple TERN applications. Vocabularies are used by other system run by Australian federal and state government agencies. The machine-readable format enables consistent descriptions of all digital artefacts and supports automated data validation plus federated search. The system and processes developed have significantly lowered the barrier in the community to represent any terms with associated meaning.

What made this work? Three things: getting domain scientists involved early to describe terms and definitions, developing semantic-enabled systems and processes to manage vocabularies, and automating quality checks with SHACL validation. The robust publication processes, while keeping vocabularies available through open services and from multiple endpoints.

Based on user feedback requesting better vocabulary discovery, including cross-domain examples like agriculture, we plan to integrate Large Language Models (LLMs) into our vocabulary management workflow to enhance semantic inference and search capabilities. This integration will enable more intuitive vocabulary discovery, where users searching for broad concepts like "cover" can automatically retrieve related terms such as "vegetation cover," "ground cover," and "canopy cover," while searches for specific terms like "vegetation cover" will surface semantically similar vocabularies including "plant cover" and "forest cover." The LLM-enhanced system will leverage natural language processing to understand user intent and provide contextually relevant vocabulary recommendations.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] F. Z. Amara, M. Hemam, M. Djezzar, M. Maimour, Semantic web technologies for internet of things semantic interoperability, in: Y. Maleh, M. Alazab, N. Gherabi, L. Tawalbeh, A. A. Abd El-Latif (Eds.), Advances in Information, Communication and Cybersecurity, Springer International Publishing, Cham, 2022, pp. 133–143.

[2] Terrestrial Ecosystem Research Network (TERN), TERN Data Discovery Portal, 2025. URL: https://portal.tern.org.au/.

[3] Terrestrial Ecosystem Research Network (TERN), EcoPlots, 2025. URL: https://ecoplots.tern.org.au/.

[4] Terrestrial Ecosystem Research Network (TERN), EcoImages, 2025. URL: https://ecoimages.tern.org.au/.

[5] A. Miles, S. Bechhofer, SKOS simple knowledge organization system reference, W3C Recommendation, 2009. URL: https://www.w3.org/TR/skos-reference/.

[6] Terrestrial Ecosystem Research Network (TERN), TERN Ontology, 2025. URL: https://github.com/ternaustralia/ontology_tern.

[7] C. Di Muri, M. Pulieri, D. Raho, et al., Assessing semantic interoperability in environmental sciences: variety of approaches and semantic artefacts, Scientific Data 11 (2024) 1055. URL: https://doi.org/10.1038/s41597-024-03669-3. doi:10.1038/s41597-024-03669-3.

[8] S. J. D. Cox, A. N. Gonzalez-Beltran, B. Magagna, M.-C. Marinescu, Ten simple rules for making a vocabulary fair, PLOS Computational Biology 17 (2021) 1–15. URL: https://doi.org/10.1371/journal.pcbi.1009041. doi:10.1371/journal.pcbi.1009041.

[9] Terrestrial Ecosystem Research Network (TERN), TERN Linked Data, 2025. URL: https://linkeddata.tern.org.au/.

[10] KurrawongAI, Prez, 2025. URL: https://github.com/RDFLib/prez.

[11] KurrawongAI, KurrawongAI, 2025. URL: https://kurrawong.ai/.

[12] Australian Research Data Commons, Research Vocabularies Australia, 2025. URL: https://vocabs.ardc.edu.au/.

[13] Terrestrial Ecosystem Research Network (TERN), TERN SHaRED Data Submission Tool, 2024. URL: https://shared.tern.org.au/.

[14] Australian Government Department of Climate Change, Energy, the Environment and Water, Biodiversity Data Repository, 2025. URL: https://bdr.gov.au/.

[15] Terrestrial Ecosystem Research Network (TERN), Ecological Monitoring System Australia, 2025. URL: https://emsa.tern.org.au/.