Diversifying world agriculture - a common framework for crop data comparison

Agnes **Aboagye**

School of Biosciences, University of Nottingham, Sutton Bonington, Leicestershire, LE12 5R, United Kingdom

Abstract

Underutilised crops (UCs) are plant species that are semi-domesticated and adapted to the local environment, and used as traditional food sources in the past but became neglected by research. One of the obstacles to the wider adoption of (UCs) is the lack of characteristic data that have been acquired or can be readily compared to understand their properties compared with other crops. The development of existing crop-specific trait dictionaries and ontologies including Crop Ontology (CO), Plant Ontology (PO), Trait Ontology (TO), Compositional Dietary Nutritional Ontology (CDNO), and Food Ontology (FoodOn) have improved data sharing in the plant science domain. However, these ontologies do not have appropriate classes, terms, and definitions relating to UCs that can directly map to challenges in the UCs crop trait datasets. Therefore, within the EU-RADIANT project, plant trait datasets generated by project partners for 11 underutilised crops will be collected, collated, and managed in a modified version of the Germinate database https://ics.hutton.ac.uk/germinate-demo/#/home adhering to findable, accessible, interoperable, reusable (FAIR) data management and stewardship. Furthermore, publicly available nutritional datasets have been collected and managed in an appropriate framework and annotated with ontology classes and terms following the Investigation-Study-Assay (ISA) metadata framework to increase their interoperability. Semantic web technologies from World Wide Web Consortium (WC3) standards will be implemented for internet machine readability. Resources Descriptive Framework (RDF) and linked data technologies will be used to provide the foundation for connecting the data. Overall, the benefit of using ontologies is to allow an effective integration analysis of heterogeneous datasets, and comparison of datasets, which will enhance the comparison of UCs with major crops and contribute to their wider adoption and utilisation.

Keywords

Ontologies, FAIR data, underutilised crops, ISA framework, data comparison

1. Introduction

Diversification of staple crops is one of the strategies for addressing the long-term effects of climate change while tackling nutritional deficiencies facing the world's growing population [1]. Moreover, orphan/neglected/underutilised/minor crops are not only resilient to biotic and abiotic stresses but also are often nutritionally dense, and hence can be used to combat nutritional inadequacy and food insecurity [2]. Currently, there are limited data management systems that provide the ability to compare and share underutilised crop trait datasets, with inconsistent and inadequate data management apparent both for major and minor crops. Therefore, there is a need to collect and manage plant trait data with a systematic annotation that adheres to the findable, accessible, interoperable, and reusable (FAIR) guiding principles. In a database management system, ontologies allow for organizing and structuring information so that the dispersed information in the form of documents, manuals, and publications are harmonised and integrated [3] to foster comparison across species [4]. The explicit specification and standardisation provided by ontologies support the interoperability of databases and information processing even where datasets may be found across several databases [5]. Annotating data and metadata with ontology terms in research projects that cover experimental conditions, protocols, and plant trait measurement units will increase the interoperability and reusability of the datasets.

The specification of ontologies required should cover the concepts, properties, and relations that exist within the research field for data sharing, access, and integration. Currently, the development of existing ontologies, including Plant Ontology (PO), Trait Ontology (TO), Compositional Dietary Nutritional Ontology (CDNO), and Food Ontology (FoodOn) has improved data sharing, however, these ontologies are not organised in a way that reflects the acquisition of diverse data relating to underutilised crops [6], [7] [8]. While the Crop Ontology framework has been developed to generate trait dictionaries for ~30 crops, the terms are not consistently organised [9] and are thus unsuitable to represent the diversity of data associated with underutilised crops. At present, there are limited knowledgebase systems that explore ontologies with relevant semantic web technologies for information sharing and exchange for the adoption and utilisation of UCs. Therefore, both upper-level and domain-specific ontologies should be incorporated into the metadata annotation process where ontology terms adequately cover the Investigation-Study-Assay (ISA) framework of the research project [10] [11]. The ISA framework is one of the standardisation methods that have been developed to aid in the management of phenotypic data. [12]. The ISA framework is relevant when harmonising and integrating data from heterogeneous sources to address climate change and food security challenges [10]. In plant phenotyping, standardised and open-sourced 'Minimal Information About Plant Phenotyping Experiments (MIAPPE), provides a generic checklist for describing phenotypic data independent of project or database objectives [13]. Independently, the challenge of database integration for plant breeding data has been addressed through the development of the Breeding Application Programme Interface (BrAPI) [14]. BrAPI specifications describe services, structures, inputs, and protocols required to pass data between one BrAPI-compliant data collection application and databases, and subsequently when data collection is complete datasets may be uploaded into BrAPI-compliant databases. Nevertheless, not all databases are BrAPI compliant thus, breeders are unable to share and receive data from diverse sources [15]. Even though MIAPPE and BrAPI are independent standards and guidelines, both standards leverage ontologies to enhance data access, integration, and semantic understanding within the plant science domain for agricultural productivity [11], [14].

Semantic web technologies are World Wide Web Consortium (WC3) standards that aim to ensure data on the internet are machine readable. The standard allows the encoding of data with the Resources Descriptive Framework (RDF) and Web Ontology Language (OWL) [16]. RDF is a linked data technology that provides the foundation for connecting data from heterogeneous sources designed to represent data in triples (subject-predicate-objects). RDF triples represent information in the form of statements about a subject that has property represented by a predicate and has a value as an object. Based on the semantic web technology including RDF and Web Ontology Language, a SPARQL query protocol can be executed to retrieve information from the web of data [17]. Thus, the use of semantic inferences and ontologies is an innovative approach to the management of information and can significantly enhance the representation of data for downstream analysis and use [17].

1.1 Motivation

The world's sensitivity to climate change and its effects combined with population growth resulting in urbanisation has further decreased the availability of arable lands for agricultural production contributing to hunger and poverty in different parts of the world [18], [19]. To address these global challenges, the adoption and utilisation of UCs can be used to tackle crop diversification. However, due to neglect in research and investment, UCs lack sufficient systematic characterisation datasets to make direct and relevant comparisons between UCs and major crops [9]. Therefore, there is a need to develop interoperable database frameworks detailing their adaptive response to agro-ecologies and end-use properties such as nutritional composition. Information about UC breeding and production remains poorly documented in knowledge systems compared to major crops. Moreover, trait data that are collected and managed within the plant science domain must have sufficient practice to abide by the (FAIR) guiding principles to increase research efficiencies. Therefore, there is a need to develop knowledgebase systems that incorporate ontologies and semantic web technologies to advance information sharing on UCs and facilitate the comparison of data within the plant science domain [10].

Access and utilisation of such databases could help improve and speed up biological knowledge discovery that is relevant for addressing the consequences of climate change and solving food security issues around the globe.

1.2 Problem statement

While various ontologies in the plant science domain exist, they are not organised in a way that represents the appropriate classes, terms, and definitions relating to underutilised crops. Furthermore, UCs lack interoperable database management systems that collect relevant data, making the datasets FAIR. Additionally, this ability is no longer sufficient for relational databases that store most information useful in the agriculture research domain. To address this gap in database interoperability, semantic web technologies can help make information more meaningful and understandable for people and computational systems by providing well-structured searchable databases that support humans and computers to understand the information hosted on the databases [20].

1.3 Hypothesis and Objectives

The overarching hypothesis is that datasets relating to UCs can be made FAIR, and reusable with a well-constructed data curation pipeline annotated with a consistent set of ontologies. Ontology standards along with phenotyping and breeding API standards could foster interoperability of data collection applications and databases for downstream data exchange, dissemination, usage, and comparison with major crops. The objectives set to test this hypothesis include:

- Database schema optimization and data curation of EU-RADIANT project trait data into modified versions of the Germinate database.
- Development of curation tools and a database based on semantic web technology to enable easy access and retrieval of information from the database with minimal interventions.
- Ensure the database is compatible with phenotyping standard (MIAPPE) and Breeding API (BrAPI) standards to guarantee data findability, accessibility, interoperability, and reusability.
- Assess, annotate, and contribute to the development of appropriate ontology classes, terms, definitions, and relationships to annotated crop trait data from EU-RADIANT project.

2. Materials and Methods

1. Datasets collection and curation from field trials sites across the EU for 11 UCs, to develop a curation tool and compatible plant phenotypic data management standards for data exchange, reuse, and analysis.
2. Knowledgebase development relying on semantic web technologies and crop domain-specific ontologies for plant phenotypic data integration using the Investigation, Study, Assay (ISA) framework to assign the ontology annotation to the data.

3. Acknowledgments

- UKRI BBSRC- Doctoral Training Partnership
- University of Nottingham, UK
- Crops For the Future, UK
- James Hutton Institute, UK

4. References

[1] L. Fleskens *et al.*, 'Overcoming barriers to crop diversification uptake in Europe: A mini review'. [Online]. Available: https://www.cropdiversification.eu/about.html

[2] S. N. Azam-Ali, P. J. Gregory, and E. Jahanshiri, 'Diversifying the UK Agrifood System: A Role for Neglected and Underutilised Crops', *Agronomy*, vol. 14, no. 4. Multidisciplinary Digital Publishing Institute (MDPI), Apr. 01, 2024. doi: 10.3390/agronomy14040853.

[3] D. Stevenson and C. Zumajo-Cardona, 'From Plant Ontology to Gene Ontology and back', *Curr Plant Biol*, vol. 14, pp. 66–69, Sep. 2018, doi: 10.1016/j.cpb.2018.09.009.

[4] R. L. Walls *et al.*, 'The plant ontology facilitates comparisons of plant development stages across species', *Front Plant Sci*, vol. 10, May 2019, doi: 10.3389/fpls.2019.00631.

[5] N. Morales *et al.*, 'Breedbase: A digital ecosystem for modern plant breeding', *G3: Genes, Genomes, Genetics*, vol. 12, no. 7, Jul. 2022, doi: 10.1093/g3journal/jkac078.

[6] L. Cooper, J. Elser, M. A. Laporte, E. Arnaud, and P. Jaiswal, 'Planteome 2024 Update: Reference Ontologiesãnd Kno wledg ebase f or Plant Biology', *Nucleic Acids Res*, vol. 52, no. D1, pp. D1548–D1555, Jan. 2024, doi: 10.1093/nar/gkad1028.

[7] D. M. Dooley *et al.*, 'Food on: A harmonized food ontology to increase global food traceability, quality control and data integration', *NPJ Sci Food*, vol. 2, no. 1, pp. 1–10, Jan. 2018, doi: 10.1038/s41538-018-0032-6.

[8] L. Andrés-Hernández *et al.*, 'Establishing a Common Nutritional Vocabulary - From Food Production to Diet', *Front Nutr*, vol. 9, Jun. 2022, doi: 10.3389/fnut.2022.928837.

[9] L. Andrés-Hernández, R. A. Halimi, R. Mauleon, S. Mayes, A. Baten, and G. J. King, 'Challenges for FAIR-compliant description and comparison of crop phenotype data with standardized controlled vocabularies', *Database*, vol. 2021, May 2021, doi: 10.1093/database/baab028.

[10] K. Dumschott *et al.*, 'Ontologies for increasing the FAIRness of plant research data', *Frontiers in Plant Science*, vol. 14. Frontiers Media SA, 2023. doi: 10.3389/fpls.2023.1279694.

[11] E. A. Papoutsoglou *et al.*, 'Enabling reusability of plant phenomic datasets with MIAPPE 1.1', *New Phytologist*, vol. 227, no. 1, pp. 260–273, Jul. 2020, doi: 10.1111/nph.16544.

[12] B. Dipta *et al.*, 'Digitalization of potato breeding program: Improving data collection and management', *Heliyon*, vol. 9, no. 1. Elsevier Ltd, Jan. 01, 2023. doi: 10.1016/j.heliyon.2023.e12974.

[13] H. Ćwiek-Kupczyńska *et al.*, 'Measures for interoperability of phenotypic data: Minimum information requirements and formatting', *Plant Methods*, vol. 12, no. 1, pp. 1–18, 2016, doi: 10.1186/s13007-016-0144-4.

[14] P. Selby *et al.*, 'BrAPI - An application programming interface for plant breeding applications', *Bioinformatics*, vol. 35, no. 20, pp. 4147–4155, 2019, doi: 10.1093/bioinformatics/btz190.

[15] S. Jung *et al.*, 'The Breeding Information Management System (BIMS): An online resource for crop breeding', *Database*, vol. 2021, 2021, doi: 10.1093/database/baab054.

[16] A. Lawan, A. Rakib, N. Alechina, and A. Karunaratne, ‘The Onto-CropBase – A semantic web application for querying crops linked-data’, *Communications in Computer and Information Science*, vol. 613, pp. 384–399, 2016, doi: 10.1007/978-3-319-34099-9_30.

[17] A. Venkatesan *et al.*, ‘Agronomic Linked data (AGROLD): A knowledge-based system to enable integrative biology in agronomy’, *PLoS One*, vol. 13, no. 11, Nov. 2018, doi: 10.1371/journal.pone.0198270.

[18] A. Ahmadalipour, H. Moradkhani, A. Castelletti, and N. Magliocca, ‘Future drought risk in Africa: Integrating vulnerability, climate change, and population growth’, *Science of the Total Environment*, vol. 662, pp. 672–686, Apr. 2019, doi: 10.1016/j.scitotenv.2019.01.278.

[19] J. Hufnagel, M. Reckling, and F. Ewert, ‘Diverse approaches to crop diversification in agricultural research. A review’, *Agronomy for Sustainable Development*, vol. 40, no. 2. Springer, Apr. 01, 2020. doi: 10.1007/s13593-020-00617-4.

[20] B. Pitollat *et al.*, ‘AgroLD: A Knowledge Graph Database for plant functional genomics’, p. 10, 2021, doi: 10.1101/325423ï.