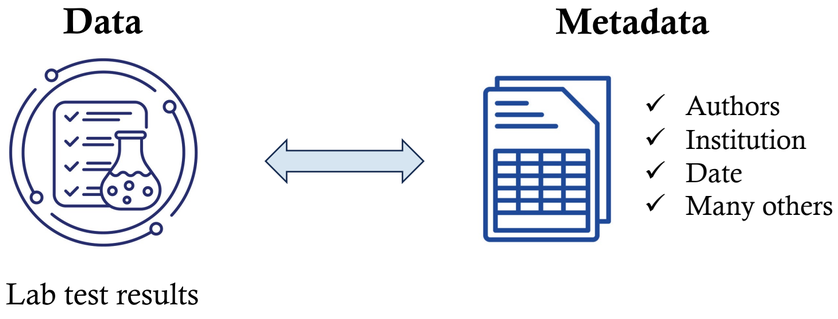
Beyond the Surface: Revealing Researchers' Behaviour in Public Repositories

Maria Juliana Rodriguez-Cubillos[1*], Tomasz Zieliński[1], Jason R. Swedlow[2], T. Ian Simpson[3] and Andrew J. Millar[1]

[1] Centre for Engineering Biology and School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, UK
[2] Divisions of Molecular Cell and Developmental Biology, and Computational Biology, University of Dundee, Dundee, Scotland, UK
[3] School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK
* juliana.rodriguez@ed.ac.uk

Introduction



Promoting data availability and accessibility is a foundational principle of FAIR data guidance[1]. However, better metadata is needed to ensure knowledge dissemination, highlighting the vital role of documenting research studies.

Aim: Develop an AI metadata enrichment tool focusing on named entities within unstructured textual data. Using text mining, Machine Learning, and NLP models like GPT and BERT, my strategic goal is to offer feedback on free text descriptions to improve metadata quality and dataset reusability.

Target repositories: BioDare2[2] (biological time series data repository), DataShare[3] (general repository at the University of Edinburgh) and IDR (Image Data Resource)[4].

Methods

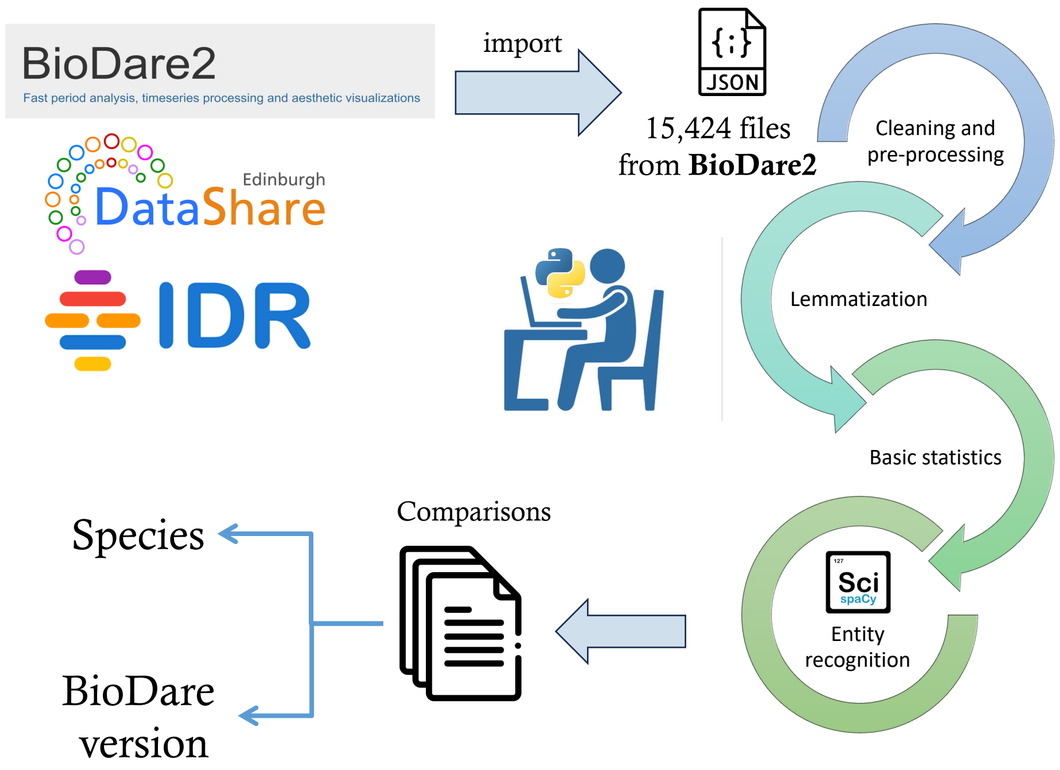


Figure 1. Workflow for the descriptive analysis of the BioDare repository. This process will be repeated with DataShare and IDR datasets in the upcoming months.

Results & Discussion

The earlier version of BioDare2 demanded more extensive or specific information, resulting in a higher word and character count (Figure 2).

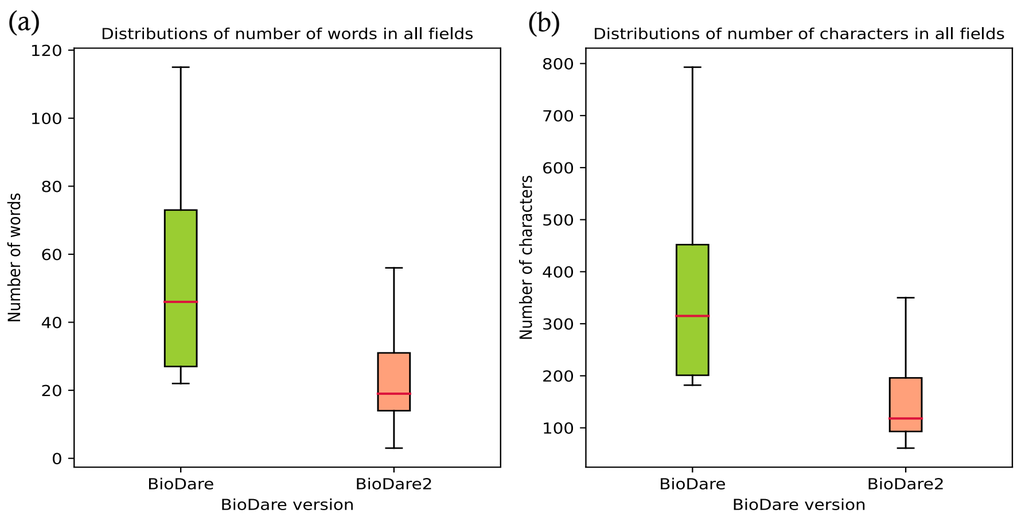
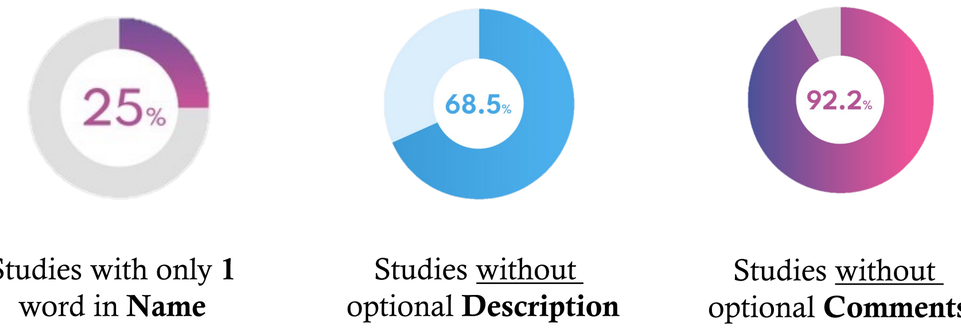


Figure 2. Distribution of the number of (a) words and (b) characters for all fields in the datasets of BioDare (green) and BioDare2 (orange) repositories.

Minimalism in metadata



Similar quality in all species

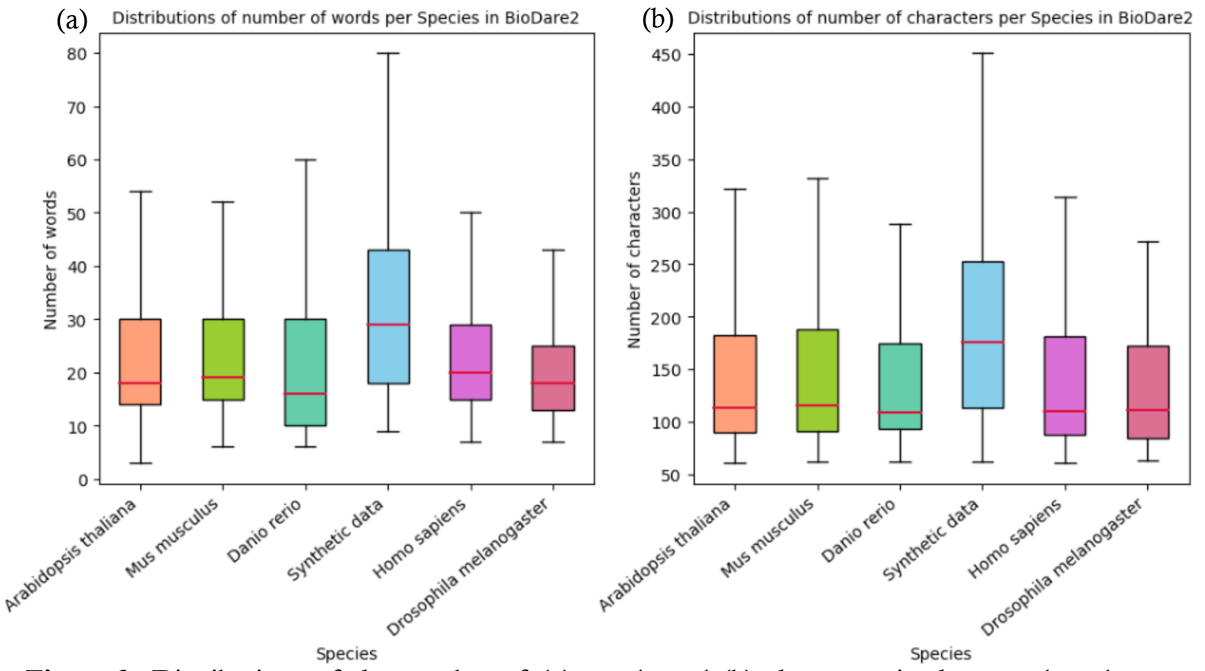


Figure 3. Distributions of the number of (a) words and (b) characters in the complete dataset of BioDare2 for *Arabidopsis thaliana* (orange), *Mus musculus* (green), *Danio rerio* (turquoise), Synthetic data (blue), *Homo sapiens* (magenta) and *Drosophila melanogaster* (dark pink). The species or datasets represented counts with 500 or more entries deposited in the repository. They also were selected to represent different taxa in the comparison.

All species have consistent counts. However, synthetic data registered a notable difference in the values, showing the tendency to describe those entries in more detail. That information comes from different organisms, which explains this variation.

Species have different entities

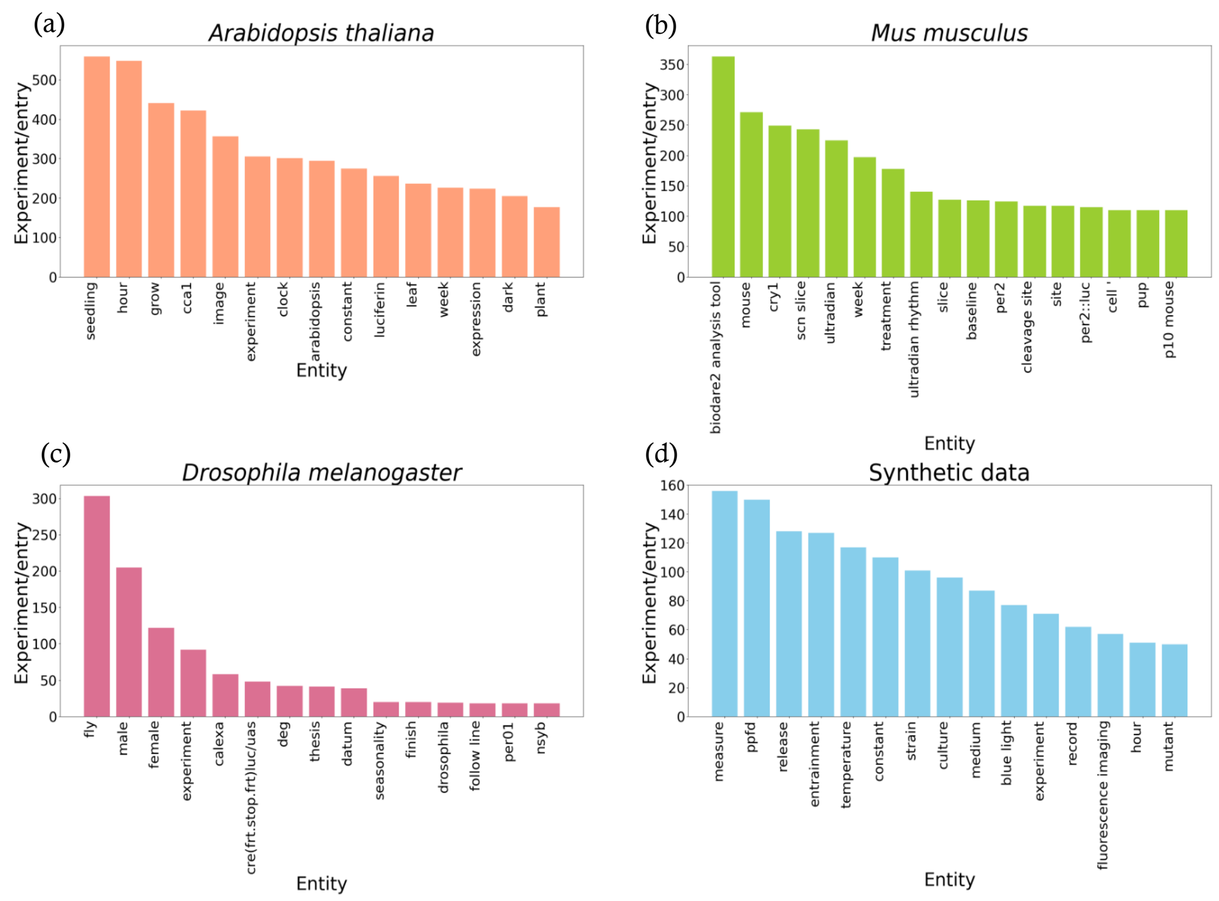


Figure 4. The number of entities in the complete BioDare2 dataset. After removing special characters and the 10 most common entities, the BioDare2 dataset shows entity counts for the 15 most common entities. (a) *Arabidopsis thaliana* (orange), (b) *Mus musculus* (green), (c) *Drosophila melanogaster* (dark pink) and (d) Synthetic data (blue).

The counts and entities identified vary depending on the organism. Nevertheless, the entities are coherent to the species. The entity recognition analysis (Figure 4) shows the presence of the gene CCA1 in *Arabidopsis thaliana*. It is part of a feedback loop closely associated with the circadian clock in this genus[5], and CRY1 in *Mus musculus* encodes an essential component protein of the circadian core oscillator complex[6].

Conclusions and upcoming research

- ✓ The results elucidated the current state of the art for the BioDare repository. For instance, they highlight the user's behaviour and provide information that could be associated with specific species. This analysis will be extended using time series and chronological information by testing data characteristic correlation with metadata.
- ✓ Entity recognition is a promising resource with immense potential for identifying similarities to describe and define datasets. When reinforced and extended to further comparisons, its application could open up new avenues for data analysis and interpretation.
- ✓ The code and protocol implemented have proven reliable in the exploration of the BioDare2 repository. Therefore, they will be used to analyse DataShare and IDR subsequently.

Reference

[1] Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 2016 3:1*, *3*(1), 1–9. doi: 10.1038/sdata.2016.18

[2] Zieliński, T., Hay, J., & Millar, A. J. (2022). Period Estimation and Rhythm Detection in Timeseries Data Using BioDare2, the Free, Online Community Resource. doi: 10.1007/978-1-0716-1912-4_2/FIGURES/4

[3] University of Edinburgh. (2008). *DSpace Principal*. Retrieved from https://datashare.ed.ac.uk/

[4] Williams, E., et al., (2017). Image Data Resource: a bioimage data integration and publication platform. doi: 10.1038/nmeth.4326

[5] Wang ZY, & Tobin EM. Constitutive expression of the CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression. doi: 10.1016/s0092-8674(00)81464-6. PMID: 9657153.

[6] van der Spek PJ, et al. Cloning, tissue expression, and mapping of a human photolyase homolog with similarity to plant blue-light receptors. DOI: 10.1006/geno.1996.0539. PMID: 8921389.

eastBIO East of Scotland Bioscience Doctoral Training Partnership
UKRI Biotechnology and Biological Sciences Research Council