





The Digitalization of Bioassays in the Open Research Knowledge Graph

Jennifer D'Souza¹, Anita Monteverdi², Muhammad Haris³,
Marco Anteghini⁴, Kheir Eddine Farfar¹, Markus Stocker¹,
Vitor A.P. Martins dos Santos⁴, and Sören Auer^{1,3}

¹ TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{jennifer.dsouza,kheir.farfar,markus.stocker,auer}@tib.eu

² Brain Connectivity Center, IRCCS Mondino Foundation, 27100 Pavia, Italy
anita.monteverdi01@universitadipavia.it

³ L3S Research Center, Leibniz University Hannover, Hanover, Germany
haris@l3s.de

⁴ Lifeglimmer GmbH, Markelstr. 38, 12163 Berlin, Germany
{anteghini,vds}@lifeglimmer.com

Abstract. Background: Recent years are seeing a growing impetus in the semantification of scholarly knowledge at the fine-grained level of scientific entities in knowledge graphs. The Open Research Knowledge Graph (ORKG, orkg.org) represents an important step in this direction, with thousands of *scholarly contributions* as structured, fine-grained, machine-readable data. There is a need, however, to engender change in traditional community practices of recording contributions as unstructured, non-machine-readable text. For this in turn, there is a strong need for AI tools designed for scientists that permit easy and accurate semantification of their scholarly contributions. We present one such tool, ORKG-ASSAYS. **Implementation:** ORKG-ASSAYS is a freely available AI micro-service in ORKG written in Python designed to assist scientists obtain semantified bioassays as a set of triples. It uses an AI-based clustering algorithm which on gold-standard evaluations over 900 bioassays with 5,514 unique property-value pairs for 103 predicates shows competitive performance. **Results and Discussion:** As a result, semantified assay collections can be surveyed on the ORKG platform via tabulation or chart-based visualizations of key property values of the chemicals and compounds offering smart knowledge access to biochemists and pharmaceutical researchers in the advancement of drug development.

Keywords: Open research knowledge graph · Scholarly digital library · Bioassays · K-means clustering · Artificial intelligence

Supported by TIB Leibniz Information Centre for Science and Technology, the EU H2020 ERC project ScienceGraph (GA ID: 819536) and the ITN PERICO (GA ID: 812968).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
C. Strauss et al. (Eds.): DEXA 2022, LNCS 13426, pp. 63–68, 2022.
https://doi.org/10.1007/978-3-031-12423-5_5

1 Introduction

The Open Research Knowledge Graph (ORKG) [3] digital library addresses scholarly content digitalization as a distributed, decentralized, and collaborative scholarly knowledge creation process that can be powered with automated semantification modules via a continuous, ongoing development cycle of autonomously maintained AI micro-services. To this end, this paper presents ORKG-ASSAYS an AI-based semantification micro-service trained on structured data based on the Bioassay ontology (BAO), and fitted in the ORKG for the rapid assimilation of digitalized biological assays (bioassays). While ORKG-ASSAYS will be the first Life Science domain supported by an automated semantification micro-service in the ORKG, to our knowledge, it fosters the development of the first end-to-end bioassay digitalization workflow in the overall scholarly community as well.

The ORKG-ASSAYS micro-service workflow involves four steps. **1)** Querying a bioassay depositor for their unstructured or semi-structured assays. Commonly, bioassays raw data are obtained via the PubChem depository [12] – a major depositor of bioassays from various research institutes. **2)** Semantifying the assay via the ORKG-ASSAYS AI clustering model. **3)** Linking the depositor-provided assay cross-references to their scientific articles. And, **4)** integrating the bioassay semantic graph in the ORKG. Programmed in Python, ORKG-ASSAYS provides web-based and programmatic tools for semantifying bioassay texts. The semantified bioassay once entered in the ORKG is *editable* via user-friendly front-end interfaces, is *surveyable* via tabulations [11] or 2-D chart visualizations, and is *queryable* for various scientific semantic ORKG relationships. The ORKG-ASSAYS AI clustering method demonstrates high semantification performance F1 scores above 80% and has been chosen after diverse methodological tests including the state-of-the-art, bidirectional transformer-based SciBERT model discussed in prior work [1].

Summing up, ORKG-ASSAYS offers a highly accurate and pragmatic semantification model alleviating unrealistic expectations on scientists to semantify their bioassays from scratch, by instead offering them a mere curatorial role of the automatic annotations. The pace with which novel bioassays are being submitted suggests that we have only begun to explore the scope of possible assay formats and technologies to interrogate complex biological systems. Thus this data domain, specifically, promises long-standing future application discovery many of which remain potentially untapped. Furthermore, inspired by the method we demonstrate, by drastically reducing the time required to semantify data for other scholarly domains as well, digitalization can be realistically advocated to become a standard part of the publication process.

2 Bioassay Digitalization in the ORKG

ORKG-ASSAYS will now be discussed as its implementation w.r.t. the KG Lifecycle requirements consisting of the graph creation, hosting, curation, and deployment modules. The ORKG-ASSAYS micro-service belongs in an early stage of

graph creation, i.e. when generating the graph itself. Thus, while the graph creation module handling the normalization of variously formatted graph data is beyond the scope of ORKG-ASSAYS, it addresses extracting the assay texts from heterogeneous bioassay depositories each with different file formats, generating a BAO-based structured graph. The end-to-end ORKG-ASSAYS semantification pipeline in a micro-service is discussed below.

Data Preparation. This step relies on public access availability to an assay depository’s querying mechanism. PubChem, reported to have over 1 million assays [8], is queryable via its public REST API for its bioassays where some assays have depositor-provided cross-references to scientific articles in PubMed. Depending on the depositor, the data could be returned in JSON, XML, or CSV. We implemented a specific pipeline for “The Scripps Research Molecular Screening Center” which returned JSON query responses. It reported nearly 1,600 bioassays. However, to prepare the data, the bioassay description-specific sections had to be located in its JSON response file and the text then extracted. The text was merged from two separate parts, viz. assay overview and assay protocol summary. We noted that this parsing heuristic can be applied to most depositor responses, although there maybe some exceptions.

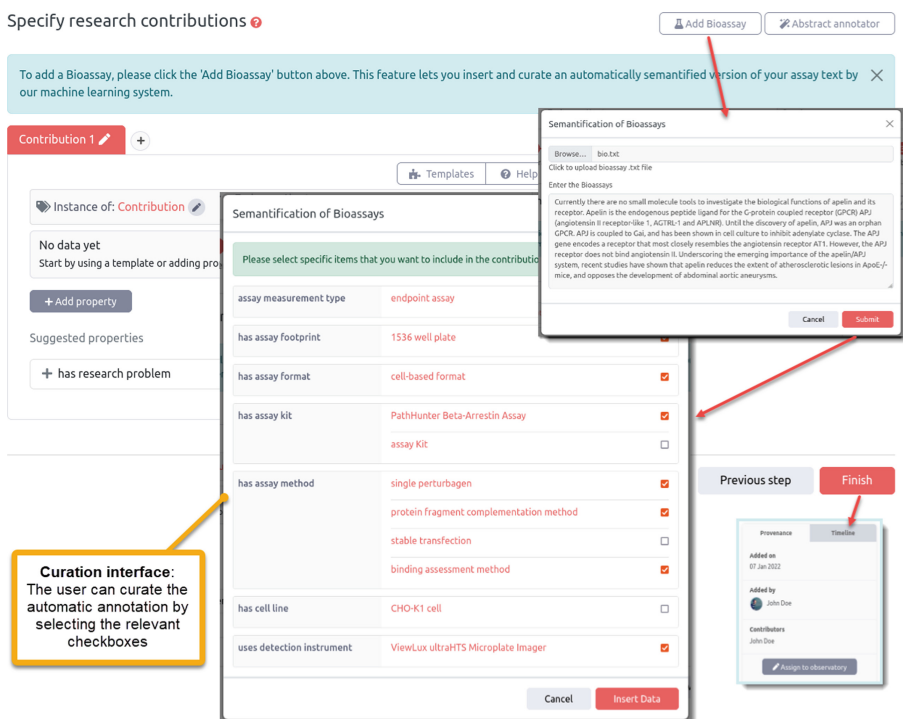
Automated Clustering-Based Semantification. Traditionally, AI-based scholarly KG construction is addressed by the recognition of entities and relations in scientific articles as sequence labeling and classification objectives [5–7, 9, 10]. We instead address the problem of bioassay semantification with a clustering objective. We choose clustering from our corpus observations that bioassays with similar text descriptions are semantified with similar sets of logical statements. Thus, the bioassays could be clustered based on their text descriptions and each cluster could be collectively semantified by the labels of the trained cluster. Indeed while entity and relation classification are sound strategies, they would be unnecessarily more complex and time-consuming methods for the problem at hand. We refer the reader to our prior work [2] which contrasts a classification versus a clustering objective for bioassay semantification.

The final semantification function in ORKG-ASSAYS was arrived at by an experimental process. This entailed testing two different vectorizations, i.e. TF-IDF and SciBERT [4], for the bioassay text to find the optimal representation for clustering by K -means with the elbow optimization strategy to find the best K value. While the TF-IDF vector is fitted on a training collection of assays, the SciBERT embeddings are directly queried for their pretrained 768 dimensional vectors. The results are shown in Table 1. We see that the direct TF-IDF vectorization on bioassay text outperforms the scholarly-articles-based pretrained SciBERT at 0.83 $F1$ vs. 0.77 $F1$ with fewer clusters (450 vs. 550).

Building the Knowledge Graph. We leverage the ORKG to convert our structured annotations to a KG. The assay’s article’s PubMed metadata is first fetched, following which the digitalized bioassay is added in the form of research contributions of the paper via the ORKG KG building functions.

Table 1. Semantification results by K-means clustering of vectorized bioassays

# Clusters (K)	TF-IDF			SciBERT		
	P	R	$F1$	P	R	$F1$
400	0.80	0.85	0.82	0.72	0.79	0.75
450	0.81	0.85	0.83	0.74	0.79	0.76
500	0.82	0.85	0.83	0.75	0.78	0.76
550	0.82	0.84	0.83	0.75	0.78	0.77
600	0.83	0.84	0.83	0.77	0.78	0.77

**Fig. 1.** ORKG frontend screens for user curation of an automatically semantified bioassay.

Data Workflows. 1. Add Paper Wizard. In the ORKG Frontend, as shown in Fig. 1, the user can add an assay by clicking the 'Add Bioassay' button. The assay gets automatically semantified with the result on a screen with checkboxes enabling accept or reject user interactions. On clicking 'Insert Data,' all selected statements and the user provenance form the ORKG. **2. Bulk Import via REST API.** To ingest the data in bulk, iterative calls to the ORKG REST API with article metadata and structured bioassay as contributions encapsulated in a JSON object can be made. This process is depicted in Figs. 2 and 3.

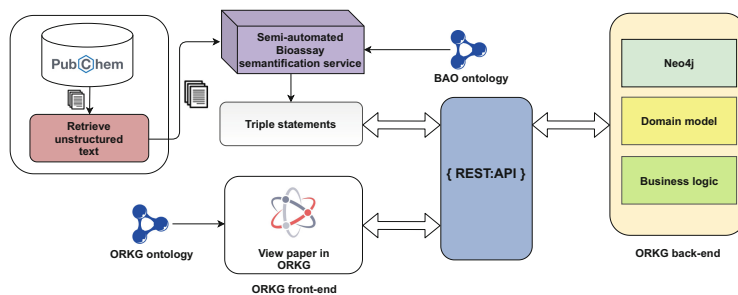


Fig. 2. End-to-end ORKG-ASSAYS semantification pipeline which practically realizes the digitalization of digitized data involving data sources, data retrieval, an annotation service, and resulting triple statements.



Fig. 3. Conversion of an unstructured Bioassay to its equivalent digitalized representation and finally presented in the ORKG frontend (<https://tinyurl.com/orkg-assay>).

3 Conclusion

We presented ORKG-ASSAYS—an end-to-end digitalization workflow of unstructured descriptions of bioassays within a next-generation digital library, the ORKG. Its supplementary information is released online <https://github.com/jd-coderepos/bioassays-ie>. The hybrid design of ORKG-ASSAYS complementarily integrates automated and manual semantification methods since pure machine learning on its own tends to be insufficiently accurate and expecting scientists to find the time to semantify their assays from scratch is unrealistic.

Bioassays being highly diverse are clearly a non-trivial semantification domain posing challenges to standardizing and integrating the data with the goal to maximize their scientific and ultimately their public health impact as the assay screening results are carried forward into drug development programs with intelligent machine assistance. The current coronavirus pandemic

situation sheds critical light on advancing the drug development research life-cycle for which bioassays are crucial, offering credence to our domain choice for semantification research. In this respect, the ORKG will not serve as a mere mirror of other Bioassay depositories, but will itself be a unique application of a highly-structured science-wide knowledge graph of scholarly contributions which incorporates semantified bioassays as well.

References

1. Anteghini, M., D'Souza, J., Dos Santos, V.A.M., Auer, S.: Scibert-based semantification of bioassays in the open research knowledge graph. In: EKAW-PD 2020, pp. 22–30 (2020)
2. Anteghini, M., D'Souza, J., Santos, V.A., Auer, S.: Easy semantification of bioassays (2021). arXiv preprint [arXiv:2111.15182](https://arxiv.org/abs/2111.15182)
3. Auer, S., et al.: Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis* **44**(3), 516–529 (2020)
4. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3606–3611 (2019)
5. Brack, A., D'Souza, J., Hoppe, A., Auer, S., Ewerth, R.: Domain-independent extraction of scientific concepts from research articles. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12035, pp. 251–266. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_17
6. Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E., Sack, H.: AI-KG: an automatically generated knowledge graph of artificial intelligence. In: Pan, J.Z., et al. (eds.) ISWC 2020. LNCS, vol. 12507, pp. 127–143. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_9
7. D'Souza, J., Auer, S., Pedersen, T.: SemEval-2021 Task 11: NLPContribution-Graph - structuring scholarly nlp contributions for a research knowledge graph. In: Proceedings of the 15th SemEval-2021, pp. 364–376. ACL, August 2021
8. Kim, S., et al.: Literature information in pubchem: associations between pubchem records and scientific articles. *J. Cheminformatics* **8**(1), 1–15 (2016)
9. Liu, H., Sarol, M.J., Kilicoglu, H.: UIUC_BioNLP at SemEval-2021 task 11: A cascade of neural models for structuring scholarly NLP contributions. In: Proceedings of the 15th SemEval-2021, pp. 377–386. ACL, August 2021
10. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: Proceedings of the 2018 EMNLP, pp. 3219–3232. ACL, October–November 2018
11. Oelen, A., Jaradeh, M.Y., Farfar, K.E., Stocker, M., Auer, S.: Comparing research contributions in a scholarly knowledge graph. In: CEUR Workshop Proceedings, vol. 2526, pp. 21–26. RWTH, Aachen (2019)
12. Wang, Y., et al.: Pubchem's bioassay database. *Nucleic Acids Res.* **40**(D1), D400–D412 (2012)