Representing Semantified Biological Assays in the Open Research Knowledge Graph[★]

Marco Anteghini[1,2][0000−0003−2794−3853], Jennifer D'Souza[3][0000−0002−6616−9509], Vitor A.P. Martins dos Santos[1,2][0000−0002−2352−9017], and Sören Auer[3][0000−0002−0698−2864]

[1] Lifeglimmer GmbH, Markelstr. 38, 12163 Berlin, Germany
[2] Wageningen University & Research, Laboratory of Systems & Synthetic Biology, Stippeneng 4, 6708 WE, Wageningen, The Netherlands
{anteghini,vds}@lifeglimmer.com
[3] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{jennifer.dsouza,soeren.auer}@tib.eu

Abstract. In the biotechnology and biomedical domains, recent text mining efforts advocate for machine-interpretable, and preferably, semantified, documentation formats of laboratory processes. This includes wet-lab protocols, (in)organic materials synthesis reactions, genetic manipulations and procedures for faster computer-mediated analysis and predictions. Herein, we present our work on *the representation of semantified bioassays in the Open Research Knowledge Graph (ORKG)*. In particular, we describe a semantification system work-in-progress to generate, automatically and quickly, the critical semantified bioassay data mass needed to foster a consistent user audience to adopt the ORKG for recording their bioassays and facilitate the organisation of research, according to FAIR principles.

Keywords: Bioassays · Open Research Knowledge Graph · Open Science Graphs

1 Introduction

More and more scholarly digital library initiatives aim at fostering the digitalization of traditional document-based scholarly articles [1,2,3,6,10,11,18,26]. This means structuring and organizing, in a fine-grained manner, knowledge elements from previously unstructured scholarly articles in a Knowledge Graph. These efforts are analogous to the digital transformation seen in recent years in other information-rich publishing and communication services, e.g., e-commerce product catalogs instead of mailorder catalogs, or online map services instead of printed street maps. For these services, the traditional document-based publication was not just digitized (by making digitized PDFs of the analog artifacts available) but has seen a comprehensively transformative digitalization.

Of available scholarly knowledge digitalization avenues [1,2,3,6,10,11,18], we highlight the Open Research Knowledge Graph (ORKG) [12]. It is a next-generation digital library (DL) that focuses on ingesting information in scholarly articles as machine-actionable knowledge graphs (KG). In it, an article is represented with both (bibliographic) metadata and semantic descriptions (as subject-predicate-object triples) of its *contributions*. ORKG has a number of advantages as: 1) it enables flexible semantic content modeling (i.e., ontologized or not, depending on the user or domain); 2) it semantifies *contributions* at various levels of granularity from shallow to fine-grained; and 3) it publishes persistent KG links per article contribution that it contains. For further technical details about the platform, we refer the reader to the introductory paper [12].

The ORKG DL aims to integrate and interlink contributions' KGs for Science at large, i.e. multidisciplinarily. Thus far, ongoing efforts are in place for integrating scholarly contributions from at least two disciplines, viz. Math [21] (e.g., `https://www.orkg.org/orkg/paper/R12192`) and the Natural Language Processing subdomain in AI [9] (e.g., `https://www.orkg.org/orkg/paper/R44253`). Moreover, the ORKG also has a separate feature to automatically import individual articles' contributions data found tabulated in survey articles [20]. E.g., an ORKG object for Earth Science articles' contributions surveyed: `https://www.orkg.org/orkg/comparison/R38484`. Since surveys are written in most disciplines, this latter feature directly targets the ORKG aim; however, its sole limitation is that it is restricted only to those papers that have been surveyed. On the other hand, with the per-domain semantification models, articles not surveyed can be also modeled in the ORKG.

In this paper, we describe our ongoing work in extending the ORKG to integrate biological assays from the Biochemistry discipline. For bioassays, a semantification model already exists as the BioAssay Ontology (BAO) [25]. However, we need to design a pragmatic workflow for integrating bioassays semantified by the BAO in the ORKG DL. To this end, we discuss the manual and automatic process of integrating such semantified data in the ORKG DL. Furthermore, we show how these semantified data integrated in the ORKG is amenable to advanced computational processing support for the researcher.

With the volume of research burgeoning [14], adopting a finer-grained semantification as KG for scholarly content representation is compelling. Better semantification means better machine actionability, which in turn means innumerable possibilities of advanced computational functions on scholarly content. One function especially poignant in this era of the publications deluge [13], is computational support to alleviate the manual information ingestion cognitive burden. This is precisely the computational support showcase we depict from the ORKG DL over our integrated bioassay KGs, consequently highlighting the benefits of digitalizing bioassays and of the ORKG DL platform.

2   Our Work-In-Progress Aims and Motivations

Allowing practitioners to easily search for similar bioassays as well as compare these semantically structured bioassays on their key properties.

Why integrate bioassays in a knowledge graph? Until their recent semantification in an expert-annotated dataset of 983 bioassays [7,22,24] based on the BAO [25], bioassays were published in the form of plain text. Integrating their semantified counterpart in a KG facilitates their advanced computational processing. Consider that key assay concepts related to biological screening, including Perturbagen, Participants, Meta Target and Detection Technology, will be machine-actionable. This widens the potential for relational enrichment and interlinking when integrated with machine-interpretable formats of wet lab protocols and inorganic materials synthesis reactions and procedures [15,16,17,19]. Furthermore, in this era of neural-based ML technologies, KG-based word embeddings foster new inferential discovery mechanisms given that they encode high-dimensional semantic spaces [5] with bioassay KGs so far untested for.

Why the ORKG DL [2]? The core of the setup of knowledge-based digitalized information flows is the distributed, decentralized, collaborative creation and evolution of information models. Moreover, vocabularies, ontologies, and knowledge graphs to establish a common understanding of the data between the various stakeholders. And, importantly, the integration of these technologies into the infrastructure and processes of search and knowledge exchange toward a research library of the future. The ORKG DL is such a solution. Implemented within TIB, as a central library and information centre for science and technology, it also promises development longevity: the Leibniz Association institutional networks presents a critical mass of application domains and users to enhance the infrastructure and continuously integrate new knowledge disciplines.

With these considerations in place, the work described in the subsequent sections is being carried forth. Next, we describe our approach in the context of two main research questions.

3   Approach: Digitalization of Biological Assays

RQ1: *What are steps for manually digitalizing a Bioassay in the ORKG?* The digitalization is based on the prior requirement that text-based bioassays are semantified based on the BioAssay Ontology (BAO) [25]. This is the manual aspect of the digitalization process involving domain experts or the assay authors themselves. In Figure 1, we show an example of a manually pre-semantified bioassay integrated in ORKG. This bioassay was semantified on eight properties based on the BAO. It was drawn from an expert-annotated set of 983 bioassays [22,24]. In terms of salient features, the bioassays in this dataset have 53 triple semantic statements on average with a minimum of 5 and a maximum of 92 statements; there are 42 different types of bioassays (e.g., luciferase reporter gene assay, protein-protein interaction assay—see in appendix the full list); and there are 11 assay formats (e.g., cell-based, biochemical). Thus, the manual semantification task complexity can be viewed as 53 modeling decisions.

In gist, the manual digitalizaton of a bioassay in the ORKG includes: 1) *a BAO-based semantification step*: forming subject-predicate-object triples of
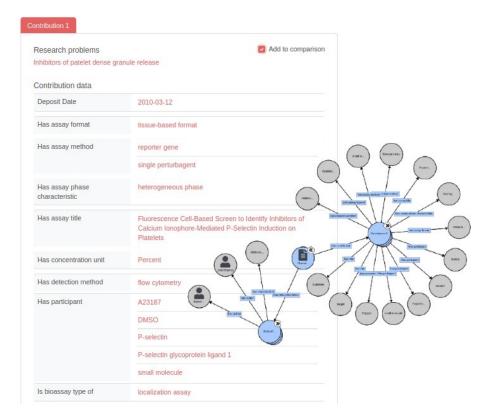
Fig. 1: An ORKG representation of a semantified Bioassay with an overlayed graph view of the assay. Accessible at: https://www.orkg.org/orkg/paper/R48178

the bioassay text content based on the BAO. E.g., for the assay in Fig. 1, a few of its semantic triples are: (Contribution, Has assay format, tissue-based format), (Contribution, Has assay method, reporter gene), among others. And as a recommended step, 2) associating each ontologized resource (i.e., a subject, a predicate, an object) with a URI as its defining class in the original ontology, which for bioassays is the BAO.

Having just described the manual digitalization workflow, we next present our hybrid workflow that is currently in development. In this, we decide to incorporate automated semantification which levies pragmatic considerations in the digitalization of bioassays in the ORKG. Relatedly, there is an existing hybrid system [7] for semantifying bioassays involving machine learning and expert interaction which inspires our work. Nonetheless, we differ. While their learning-based component relies heavily on explicitly encoded syntactic features of the text, ours relies on neural networks based on the current state-of-the-art transformer models [23] trained on millions of scientific articles [4]. Such systems by

encoding high-dimensional semantic spaces of the underlying text, obviate the need to make explicit considerations for features of the text. Moreover, they significantly outperform systems designed based on explicit features [8]—with due credit to the system by Clark et al. [7] designed prior to the onset of this revolutionary technology. Next, our hybrid workflow is designed toward a practical end—to be integrated in the ORKG DL which has a predominant focus on the digitalization of scholarly knowledge content multidisciplinarily, thus setting it apart from any existing DL.

RQ2: *What are the modules needed in the hybrid digitalization of Bioassays in the ORKG?* Essentially, given a new bioassay text input, we are implementing two modules in a two-step workflow as follows: 1) an automated semantifier; and 2) a human-in-the-loop curation of the predicted labels either by the assay author or a dedicated curator. Unlike the manual workflow, this presents a much easier and less time-intensive task for the human. They would be merely selecting the correctly predicted triples, deleting the incorrect ones, or defining new ones as needed. Assuming a well-trained machine learning module, the latter two steps may be entirely omitted. Toward this hybrid workflow, as work in progress, the automated semantifier is in development, and we are also implementing extensions in the ORKG infrastructure to include additional front-end views as assay curation interfaces.

4 Solving the Cognitive Information Ingestion Hurdle: Comparison Surveys across KG-based Bioassays

Premise: We need an information processing tool that can be used by biomedical practitioners to quickly comprehend bioassays' key properties.

The ORKG DL has a computational feature to generate and publish surveys in the form of a tabulated comparisons of the KG nodes [20]. To demonstrate this feature, we manually entered the data of three semantified bioassays in the ORKG DL. Applying then the ORKG survey feature on the three assays aggregates their semantified graph nodes in tabulated comparisons across the assays. This is depicted in Figure 2. With such structured computations enabled, we have a novel approach to uncovering and presenting information relying on aggregated scholarly knowledge. The computation shown in Fig. 2 aligns closely with the notion of the traditional survey articles, except it is fully automated and operates on machine-actionable knowledge elements. The BAO-semantified assays are compared side-by-side on their graph nodes. Thus, tracking the progress on bioassays, can be eased from a task of several days to a few minutes.

5 Conclusion

Thus in this paper, we outlined a vision in two separate workflows for integrating bioassay knowledge in the ORKG DL and our ongoing work to this end. The implications of bioassay structured and machine-actionable knowledge are broad.

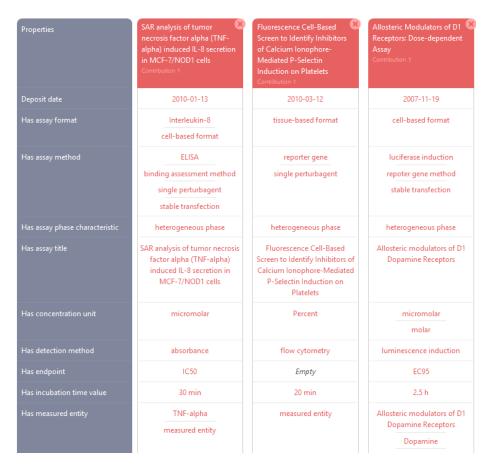| Properties	SAR analysis of tumor necrosis factor alpha (TNF-alpha) induced IL-8 secretion in MCF-7/NOD1 cells Contribution 1	Fluorescence Cell-Based Screen to Identify Inhibitors of Calcium Ionophore-Mediated P-Selectin Induction on Platelets Contribution 1	Allosteric Modulators of D1 Receptors: Dose-dependent Assay Contribution 1
Deposit date	2010-01-13	2010-03-12	2007-11-19
Has assay format	Interleukin-8 cell-based format	tissue-based format	cell-based format
Has assay method	ELISA binding assessment method single perturbagent stable transfection	reporter gene single perturbagent	luciferase induction repoter gene method stable transfection
Has assay phase characteristic	heterogeneous phase	heterogeneous phase	heterogeneous phase
Has assay title	SAR analysis of tumor necrosis factor alpha (TNF-alpha) induced IL-8 secretion in MCF-7/NOD1 cells	Fluorescence Cell-Based Screen to Identify Inhibitors of Calcium Ionophore-Mediated P-Selectin Induction on Platelets	Allosteric modulators of D1 Dopamine Receptors
Has concentration unit	micromolar	Percent	micromolar molar
Has detection method	absorbance	flow cytometry	luminescence induction
Has endpoint	IC50	*Empty*	EC95
Has incubation time value	30 min	20 min	2.5 h
Has measured entity	TNF-alpha measured entity	measured entity	Allosteric modulators of D1 Dopamine Receptors Dopamine

Fig. 2: Comparisons of semantified bioassays in the ORKG digital library. Online
`https://www.orkg.org/orkg/comparison?contributions=R48195,R48179,R48147`

To mention just one in the particular context of the current Covid-19 pandemic: The discovery of cures for diseases can be greatly expedited if scientists are given intelligent information access tools, and our work toward automatically semantifying bioassays are a step in this direction.

To this end, the workflows prescribed in this work offer the possibilities to chose between a manual or a semi-automatic strategy for bioassays' semantification within a real-world digital library.

We would like to invite interested researchers to collaborate with us on the following topics: 1) generating a large dataset of semantically structured bioassays; 2) user evaluation of our semi-automated system for semantically structuring bioassay data.

We deem this as a starting point for a discussion in the community ultimately leading to more clearly defined technical requirements, and a roadmap

for fulfilling the potential of the ORKG as a next-generation digital library for fine-grained semantified access to scholarly content.

References

1. Aryani, A., Poblet, M., Unsworth, K., Wang, J., Evans, B., Devaraju, A., Hausstein, B., Klas, C.P., Zapilko, B., Kaplun, S.: A research graph dataset for connecting research data repositories using rd-switchboard. Scientific data **5**, 180099 (2018)
2. Auer, S.: Towards an open research knowledge graph (Jan 2018). https://doi.org/10.5281/zenodo.1157185
3. Baas, J., Schotten, M., Plume, A., Côté, G., Karimi, R.: Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies **1**(1), 377–386 (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. Bianchi, F., Rossiello, G., Costabello, L., Palmonari, M., Minervini, P.: Knowledge graph embeddings and explainable ai. arXiv preprint arXiv:2004.14843 (2020)
6. Birkle, C., Pendlebury, D.A., Schnell, J., Adams, J.: Web of science as a data source for research on scientific and scholarly activity. Quantitative Science Studies **1**(1), 363–376 (2020)
7. Clark, A.M., Bunin, B.A., Litterman, N.K., Schürer, S.C., Visser, U.: Fast and accurate semantic annotation of bioassays exploiting a hybrid of machine learning and user confirmation. PeerJ **2**, e524 (2014)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
9. D'Souza, J., Auer, S.: Nlpcontributions: An annotation scheme for machine reading of scholarly contributions in natural language processing literature (2020)
10. Fricke, S.: Semantic scholar. Journal of the Medical Library Association: JMLA **106**(1), 145 (2018)
11. Hendricks, G., Tkaczyk, D., Lin, J., Feeney, P.: Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies **1**(1), 414–427 (2020)
12. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S.: Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. pp. 243–246 (2019)
13. Jinha, A.E.: Article 50 million: an estimate of the number of scholarly articles in existence. Learned Publishing **23**(3), 258–263 (2010)
14. Johnson, R., Watkinson, A., Mabe, M.: The stm report. An overview of scientific and scholarly publishing. 5th edition October (2018)
15. Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., Ceder, G.: Text-mined dataset of inorganic materials synthesis recipes. Scientific data **6**(1), 1–11 (2019)

16. Kulkarni, C., Xu, W., Ritter, A., Machiraju, R.: An annotated corpus for machine reading of instructions in wet lab protocols. In: NAACL: HLT, Volume 2 (Short Papers). pp. 97–106. New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2016

17. Kuniyoshi, F., Makino, K., Ozawa, J., Miwa, M.: Annotating and extracting synthesis process of all-solid-state batteries from scientific literature. In: LREC. pp. 1941–1950 (2020)

18. Manghi, P., Atzori, C., Bardi, A., Shirrwagen, J., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Lhden, A., Bcker, A., Mannocci, A., Horst, M., Baglioni, M., Czerniak, A., Kiatropoulou, K., Kokogiannaki, A., De Bonis, M., Artini, M., Ottonello, E., Lempesis, A., Nielsen, L.H., Ioannidis, A., Bigarella, C., Summan, F.: Openaire research graph dump (Dec 2019). https://doi.org/10.5281/zenodo.3516918, `https://doi.org/10.5281/zenodo.3516918`

19. Mysore, S., Jensen, Z., Kim, E., Huang, K., Chang, H.S., Strubell, E., Flanigan, J., McCallum, A., Olivetti, E.: The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In: Proceedings of the 13th Linguistic Annotation Workshop. pp. 56–64 (2019)

20. Oelen, A., Jaradeh, M.Y., Stocker, M., Auer, S.: Generate fair literature surveys with scholarly knowledge graphs. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. p. 97106. JCDL 20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3383583.3398520

21. Runnwerth, M., Stocker, M., Auer, S.: Operational research literature as a use case for the open research knowledge graph. In: Bigatti, A.M., Carette, J., Davenport, J.H., Joswig, M., de Wolff, T. (eds.) Mathematical Software - ICMS 2020 - 7th International Conference, Braunschweig, Germany, July 13-16, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12097, pp. 327–334. Springer (2020). https://doi.org/10.1007/978-3-030-52200-1_32

22. Schürer, S.C., Vempati, U., Smith, R., Southern, M., Lemmon, V.: Bioassay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. Journal of biomolecular screening **16**(4), 415–426 (2011)

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)

24. Vempati, U.D., Przydzial, M.J., Chung, C., Abeyruwan, S., Mir, A., Sakurai, K., Visser, U., Lemmon, V.P., Schürer, S.C.: Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the bioassay ontology (bao). PloS one **7**(11), e49198 (2012)

25. Visser, U., Abeyruwan, S., Vempati, U., Smith, R.P., Lemmon, V., Schürer, S.C.: Bioassay ontology (bao): a semantic description of bioassays and high-throughput screening results. BMC bioinformatics **12**(1), 257 (2011)

26. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft academic graph: When experts are not enough. Quantitative Science Studies **1**(1), 396–413 (2020)

A   Bioassay types

Bioassay types	
protein-protein interaction	hydrolase activity
kinase activity	protein-small molecule interaction
viability	beta lactamase reporter gene
cytochrome P450 enzyme activity	luciferase enzyme activity
luciferase reporter gene	oxidoreductase activity
protein unfolding	chaperone activity
lyase activity	transporter
plasma membrane potential	dye redistribution
calcium redistribution	apoptosis
beta lactamase reporter gene	beta galactosidase reporter gene
phosphatase activity	cAMP redistribution
IP1 redistribution	cell morphology
phosphorylation	transferase activity
isomerase activity	protein redistribution
radioligand binding	signal transduction
ion channel	platelet activation
fluorescent protein reporter gene	protein-DNA interaction
protease activity	cell permeability
protein stability	protein-turnover
localization	organism behavior
cytotoxicity	cell growth

Table 1: List of the different bioassay types present in our dataset

B   Preliminary Results of Automated Semantification: SciBERT-based Bioassay Semantifier

The semantic statements depicted in Figure 3 were automatically generated from SciBERT-based [4] neural semantification system. These predictions were made for the same bioassay text depicted in Figure 1. Comparing the automatically generated one against the reference, we see that almost all the manually curated labels are correctly predicted. Among 16 manually curated labels, excluding those we omit in our training procedure (e.g., has title, PubChem AID, Deposit Date, has incubation time value, has concentration unit), the model accurately predicts 12 statements, while the remaining were deemed by a domain-specialist as valid additional candidates to incorporate in the reference set (e.g., has significant direction, has concentration throughput).

```
Labels:
has percent response -> efficacy
has role -> culture medium
has signal direction -> signal increase corresponding to inhibition
has manufacturer -> IntelliCyt Corporation
has assay method -> single perturbagen
has organism -> Homo sapiens
involves biological process -> platelet activation
has concentration throughput -> multiple concentration
has participant # has role -> membrane protein # target
has assay method -> binding assessment method
has role -> instrumentation manufacturer
has assay control -> negative control
has participant # has role -> small molecule # perturbagen
has assay medium -> FACS buffer, BD Biosciences
has manufacturer -> Perkin Elmer
has repetition throughput -> single repetition
has role -> biologics and screening manufacturer
has assay readout content -> single readout
has assay format -> tissue-based format
has target -> blood plasma
has percent response -> percent inhibition
has signal direction -> signal decrease corresponding to inhibition
is bioassay type of -> platelet activation assay
has assay format -> cell membrane format
has primary assay -> primary assay
has concentration throughput -> single concentration
has assay readout content parametricity -> single parameter
has repetition throughput -> multiple repetition
antibody -> Red-fluorescent labeled anti-P-selectin antibody CD62P, BD Biosciences
has participant # has role -> G protein coupled receptor # target
has bioassay type -> functional
has preparation method -> recombinant expression
is bioassay type of -> localization assay
has assay medium -> assay medium
has role -> inducer
has role -> assay provider
has participant -> P-selectin glycoprotein ligand 1
assay measurement type -> endpoint assay
has assay kit -> assay Kit
has participant -> DMSO
has assay control -> positive control
has participant # has role -> A23187 # substrate
has confirmatory assay -> confirmatory assay
compound library -> MLSMR library
has assay method -> stable transfection
has assay method -> reporter gene method
has measured entity -> measured entity
has function -> binding
has assay phase characteristic -> heterogeneous phase
has role -> culture serum
has assay format -> cell-based format
has alternate target assay -> alternate target assay
has percent response -> percent activation
has participant -> P-selectin
has detection method -> flow cytometry
has role -> potentiator
has manufacturer -> BD Biosciences
```

Fig. 3: Automatically semantified bioassay (human-annotated reference in Fig. 1)