NAPRALERT, from an historical information silo to a linked resource able to address the new challenges in Natural Products Chemistry and Pharmacognosy.

Jonathan Bisson*, James McAlpine, James Graham, Guido F. Pauli
Center for Natural Product Technologies, University of Illinois at Chicago, United States
* bjo@uic.edu

***Abstract*— NAPRALERT (https://www.napralert.org) is a database on natural products, including data on ethnobotany, chemistry, pharmacology, toxicology, and clinical trials from literature dating back to the 19th century. Established in 1975 by Norman R. Farnsworth, it became a web accessible resource in 2005 but soon became stagnant while literature grew exponentially. After a complete rewrite of the platform, the focus is now on connecting this resource to the rest of the existing databases and expanding its usability. The creation of a Pharmacognosy/Natural Product ontology will foster better understanding of this domain, its linking potential with other resources and the ability to automatize literature annotation and entry efficiently.**

I. Introduction

The late Norman R. Farnsworth established NAPRALERT[1] as a tool to survey Natural Product research in 1975, when relational databases were initially being adopted. Before formal methodologies for developing ontologies existed, Professor Farnsworth developed his own simple but somewhat exhaustive hierarchical classification system and used it to annotate information gathered from the literature by him and his team for the following 30+ years.

NAPRALERT became web-accessible in 2005 with more than 200,000 citations covered as of 2015, but the informal classification system itself remained hidden behind the interface and was only accessible through customized manual queries. NAPRALERT became quiescent due to budget constraints around 2004, at a time of exponential expansion of the literature and development of new resources, new tools and new knowledge.

In 2015, a complete rewrite of the code and database schema (Fig 1.) was undertaken. A new web version (October 1, 2015) provides limited free searches to academics, industry, and governmental agencies. Currently this system is being enhanced with improved search and data entry functionalities. However, the fundamental limitations of the current approach are fully realized as it is now time to connect with existing repositories of bibliographical data, chemical structures, and biological activities including data outside the Natural Product literature previously covered by NAPRALERT.

II. Ontology Development

The recent advances and increasingly prolific field of semantic web technologies and ontology engineering now make it easier to develop formal ontologies. The existing ontologies and their usability for this project, in regard to interoperability and reduced overlap, are currently being explored (Fig 2). Meanwhile, the real entities and concepts of the Natural Product domain and their relationships to compose an ontology for Pharmacognosy are carefully analyzed.

III. Linking resources to the database

Pharmacognosy is a domain at the intersection of Biology, Biochemistry, Botany, Ethnobotany, Pharmacy and Chemistry. The data sources required relative to each discipline are diverse and cover at the same time a wider and a narrower range than what is relevant to this field, and thereby requiring very careful mapping. An example of the apparent dissonance with this domain and existing resources is ChEBI, the ontology of which covers many but not all kinds of Natural Products and at the same time covers non-natural "molecular entities". However, creating such entities *de-novo* in this domain ontology would only reduce both its interoperability and maturity. Moreover, whenever these resources provide a formal and interoperable ontology some inference can be made to enrich the queries of users and potentially the coverage of the domain.

IV. Machine learning based literature annotation

During the first two decades of NAPRALERT, it was still feasible for a relatively small team to work on the annotation and entry of literature data. Nowadays, such an approach is unrealistic both in terms of managing the coherence and validity of the data entry and, particularly, due to the tremendous human resources required. One of the approaches considered for the future of NAPRALERT is similar to what is currently achieved with projects such as GeoDeepDive (https://geodeepdive.org) and PaleoDeepDive [2]. These projects demonstrated the efficiency of mixing Natural Language Processing, Machine Learning, and their already annotated corpus of publications, making it possible to annotate literature more rapidly but still accurately.

V. Use Case

One of the expected use cases of this ontology is to determine whether or not the compounds identified in natural product extracts are truly responsible for, or involved in, the observed bioactivity. This requires the ability to compare and study bioactivity data broadly, from many different data sources and with respect to chemical composition, and to rate the confidence in both. Moreover, as compounds with

promiscuous and unspecific activities are surprisingly often mistaken for active principles, it is important to identify them and take their characteristics into account when searching for bioactives [3]. For this important use case, the new ontology will facilitate ways to annotate and link data from other resources.

ACKNOWLEDGMENT

The authors acknowledge support by grant U41 AT008706 from NCCIH and ODS/NIH. The authors also appreciate the creative advice of Sheila Miguez and Aaron Lav (Pumping Station: ONE, Chicago, IL).

REFERENCES

[1] J. G. Graham and N. R. Farnsworth, "The NAPRALERT database as an aid for discovery of novel bioactive compounds." in Comprehensive Natural Products, vol. II, H.-W. Liu and L. Mander, Eds. Elsevier: Amsterdam, 2010, pp 81– 94.

[2] S. E. Peters, C. Zhang, M. Livny and C. Ré, "A machine reading system for assembling synthetic paleontological databases." PLoS ONE, vol 9(12), pp: e113523

[3] J. Bisson, J. B. McAlpine, J.B. Friesen, SN Chen, J. G. Graham, G. F. Pauli, "Can invalid bioactives undermine Natural Product-based drug discovery" Journal of Medicinal Chemistry, vol 59(5), pp 1671-1690

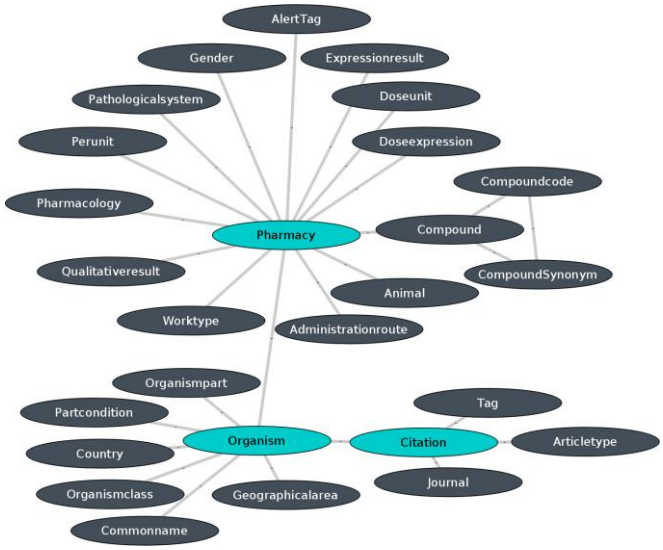

Figure 1- Structure of a portion of the actual NAPRALERT database scheme.

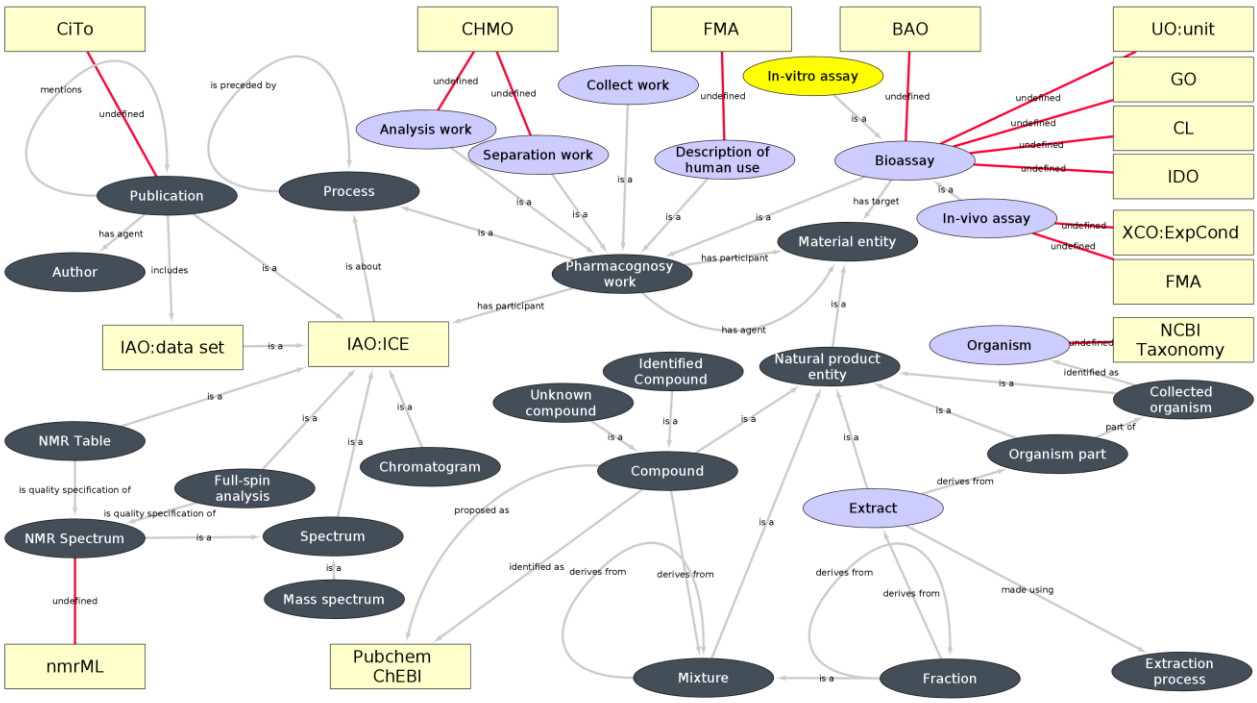

Figure 2 –Portion of a draft of the PHarmacognosy Ontology(PHO) showing in yellow boxes the ontologies that could be linked. Red lines depict the yet undetermined relationships.