Service for Digital Annotation of Scientific Data

Tomáš Kulhánek[1,*] and Yann Le Franc[1]

***Abstract*— A note in the margins of a book or a scientific paper, a comment on a manuscript or a reference to other related content: we are all using annotations to add information to existing physical documents. To offer a similar experience with digital content, we developed B2NOTE, a service that allows to associate additional information to a file in a computer-readable format based on the W3C Web Annotation standard. These digital annotations can then be searched to retrieve the related files and datasets using the user-defined content without changing the data records itself. This service has been developed within the EUDAT Common Data Infrastructure and is now part of the EOSC-Hub service portfolio.**

I. INTRODUCTION

Scientific datasets are published with their associated metadata in dedicated data repositories [1], [2], [3], [4]. To support the reuse of the published datasets, data consumers should be able to easily extend the dataset description by adding extra metadata and semantics or by tagging some part of a dataset without changing the underlying record or data element. Written annotations are often used for this purpose on existing physical documents such as books, scientific papers, images,... A similar approach for digitally annotating published datasets with structured user-defined information would be extremely valuable for data consumers, however is not currently possible in the existing repositories.

In this paper, we present B2NOTE, an integrable data annotation service based on the Web Annotation W3C standard. This web service, developed in the context of the EUDAT Common Data Infrastructure (CDI) is part of the European Open Science Cloud (EOSC) hub service catalogue [9].

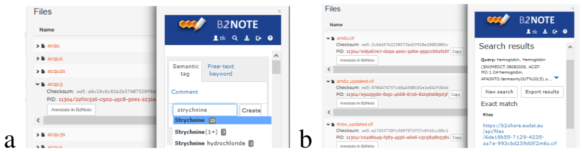
II. METHODS

The core database model follows W3C standard Web Annotation Data Model [5] which proposes a well-defined data model for annotation, an ontology and a protocol to publish and share annotations. Annotations are serialized using JSON-LD format [6]. Current implementation stores the annotations in a central MongoDB database and provide a User Interface (UI) and REST API using the Python Django, Python Eve frameworks. B2NOTE is integrated as a web widget directly within repository web UI and initialized by an API call providing the unique identifier of the file (UUID, DOI, PID,...) and the URL of the file to be annotated.

III. RESULTS

The service can be accessed at `https://b2note.eudat.eu/`. As of now, B2NOTE has been integrated with B2SHARE [1] and the semantic data publication workflow developed by the University of Porto [4].

B2NOTE allows users to manually create three types of annotation (Fig. 1). (1) **Semantic annotation** using domain specific ontologies. (2) **Free text keywords**, allowing users to annotate with domain specific keywords not defined in any ontology. (3) **Comments** providing a placeholder for long comments. Semantic annotations are supported by an index of semantic concepts/classes based on NCBO bioportal content [7]. The user can retrieve these semantics resources during the annotation process as an autocomplete list (see fig.1). An extension of this index to cover other domains has been proposed and is being further developed [8].



Checksum: md5:d4c19c6c91b2b57407529f94

PID: 11304/22f0c516-c902-45c8-90e1-2231e

strychnine

Strychnine 23

Strychnine(1+) 3

Strychnine hydrochloride 3

Checksum: md5:2c64d57b2230573b42f910e28083002c

PID: 11304/ed5a67e7-d29a-4a0c-926e-959cc662f26f

Checksum: md5:6704474757c46a4580192e1e842fd8dd

PID: 11304/e51295d0-815c-4b68-8716-81b9fa68gd3f

Checksum: md5:e17453738fc5697972f37c6fd2cc88c1

PID: 11304/c04d8a49-f583-4556-a6e6-c5c5848b5381

Query: hemoglobin, Hemoglobin (SNOMEDCT:38082009, ACGT-MO:1.0#Hemoglobin, APAONTO:termsonlyOUT%20(5).o...

https://b2share.eudat.eu/api/files/6de18b55-7129-4235-aa7e-993cbd239d0f/2m6z.cif

Fig. 1. Annotating (a) and searching (b) in B2NOTE selected dataset file from B2SHARE with 'strychnine' term and autocomplete dialog.

IV. DISCUSSION

Digital annotations provide a powerful tool to support various data management processes (data curation, lightweight semantic integration, structuring and enriching data descriptions without modifying it). B2NOTE is currently being integrated with various data services. Such integration will allow data consumers to leverage annotations to retrieve, aggregate and analyse datasets from heterogeneous and distributed data sources, fostering the reuse of scientific datasets.

ACKNOWLEDGMENT

This work is supported by EUDAT and EOSC-Hub, EU Horizon 2020 research and innovation program respectively under grant agreement No 654065 and No 777536.

REFERENCES

[1] B2SHARE https://b2share.eudat.eu/

[2] Zenodo https://zenodo.org/

[3] DataCite https://datacite.org/

[4] Karimova, Y., Castro, et al., Description + annotation: semantic data publication workflow with Dendro and B2NOTE, International Journal of Metadata, Semantics and Ontologies,12(4), pp 182-194.2017

[5] Web Annotation Data Model, W3C Recommendation 23 February 2017, https://www.w3.org/TR/annotation-model/

[6] A JSON-based Serialization for Linked Data, W3C Recommendation 16 January 2014, https://www.w3.org/TR/json-ld/

[7] Bioportal https://bioportal.bioontology.org/

[8] Goldfarb, Doron, and Yann Le Franc. "Enhancing the Discoverability and Interoperability of Multi-Disciplinary Semantic Repositories." S4BioDiv@ ISWC. 2017

[9] EOSC-Hub service catalogue https://www.eosc-hub.eu/catalogue

[*] `t.kulhanek at esciencefactory.com`

[1] T. Kulhánek and Y. Le Franc are with e-Science Data Factory S.A.S.U., Paris, France