Proceedings of the Linked Data in Linguistics Workshop @ LREC2022, pages 1–6
Marseille, 24 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

The Annohub Web Portal

Frank Abromeit
Applied Computational Linguistics Lab - Goethe University of Frankfurt, Germany
abromeit@em.uni-frankfurt.de

Abstract

We introduce the Annohub web portal, specialized on metadata for annotated language resources like corpora, lexica and linguistic terminologies. The portal will provide easy access to our previously released Annohub Linked Data set, by allowing users to explore the annotation metadata in the web browser. In addition, we added features that will allow users to contribute to Annohub by means of uploading language data, in RDF, CoNLL or XML formats, for annotation scheme and language analysis. The generated metadata is finally available for personal use, or for release in Annohub.

Keywords: Linguistic Metadata, LOD, LLOD, OLiA

1. Introduction

Linguistic metadata has been a research topic for a long time, starting with XML based data formats like TEI[1] and OLAC (Bird and Simons, 2001) and many portals that provide linguistic resource metadata have emerged ever since. For example OLAC[2], the CLARIN infrastructure (Hinrichs and Krauwer, 2014), Meta-Share[3] (Piperidis, 2012), and more recently LingHub[4] (McCrae and Cimiano, 2015). Following the paradigm to distribute data collections as Linked Open Data (LOD) (Bizer et al., 2009)[5], this methodology has been applied to linguistic data[6] (Cimiano et al., 2020), but also to the provenance metadata for linguistic resources. So, for example, LingHub provides linguistic metadata in RDF[7] formats. The RDF framework offers, in contrast to XML based metadata formats, different perspectives, like open data, standardized metadata vocabularies like Meta-Share (McCrae et al., 2015) and DCAT[8], and the ability to process resource metadata along with the actual language data, by means of SPARQL[9] queries. This finally allows tighter integration of NLP-processes that handle corpus, lexicon or terminology language data.

In recent work we have created the Annohub Linked Data set (Abromeit et al., 2020)[10], a metadata collection of annotated language resources, like corpora and lexica. Here, we introduce the Annohub portal (hosted by the Lin|gu|is|tik portal (Chiarcos et al., 2016)) that will provide users with easy access to Annohub's metadata in the web-browser. In addition, NLP-services will enable registered users of the portal to upload annotated language resources in order to perform an analysis on used languages and annotation schemes. The analysis results can then be used to create new entries in the Annohub catalogue. Furthermore, proper editing and commentary functions will help to improve the quality of the gained metadata and to keep the resources listed in Annohub up to date.

2. Annohub web portal

One of the goals of the Annohub portal is to bring metadata of prominent lexical resources and corpus data to a broader audience, but also to advertise new language resources that can not be found on other platforms like LingHub, CLARIN centers[11], Meta-Share or elsewhere. Annohub's metadata combines common resource metadata together with detailed language and annotation information. In addition, the provenance metadata is augmented, by linking annotations that have been used in a language resource, to OLiA[12] ontology classes, as well as to the original annotation scheme providers. All metadata is finally provided in a Linked Data representation, that is well suited for its use with other Linked Data applications, such as querying across multiple LLOD datasets, by means of federated SPARQL queries. Possible use cases of the new portal include:

- Search for publicly available annotated language resources like corpora, lexica or terminologies
- Contribute to Annohub by uploading language resources
- Learn about annotation schemes used in language resources

The portal currently encompasses metadata for over 1000 annotated language resources like corpora, lexica and ontologies. These resources are harvested automatically from different locations like LingHub's RDF data

[1] https://tei-c.org/
[2] http://www.language-archives.org/
[3] https://www.meta-share.org/
[4] https://linghub.org
[5] https://lod-cloud.net
[6] http://www.linguistic-lod.org/
[7] https://www.w3.org/RDF/
[8] https://www.w3.org/TR/vocab-dcat/
[9] https://www.w3.org/TR/sparql11-query/
[10] https://annohub.linguistik.de/en/
[11] https://www.clarin.eu/
[12] https://github.com/acoli-repo/olia

dump[13], CLARIN centers[14] (by means of the OAI protocol[15]), but also originate from several selected websites like the OPUS portal[16], the Språkbanken [17] website and a collection of corpora and lexica that have been compiled at the ACoLi Lab, Goethe University of Frankfurt (Chiarcos et al., 2020)[18]. The provenance metadata of each dataset is copied from the original metadata provider (RDF/XML/HTML) or has been added manually. Language and annotation information is extracted from the language data by an automated NLP-pipeline (see (Abromeit et al., 2020)). After the analysis, all language and annotation metadata, as well as the provenance metadata can be edited in the web-browser (see (Abromeit and Chiarcos, 2019) [19]) in order to complement missing information or to correct errors from the automatic analysis steps. The portal is built in Java with Apache Jena[20] and the Apache Tinkerpop framework[21] with two Neo4j[22] databases as backend. One of which is used as a backbone for the web-application, whereas the other database is used to map OLiA ontology classes to annotation tags and URLs found in the language data.

3. Ontologies of Linguistic Annotations

The *Ontologies of Linguistic Annotations* (OLiA)[23] provide a formalized, machine-readable view on linguistic annotations for more than 75 different language varieties. They cover morphology, morphosyntax, phrase structure syntax, dependency syntax, aspects of semantics, and recent extensions to discourse, information structure and anaphora, all of these are linked with an overarching reference terminology module. OLiA includes several multi-lingual or cross-linguistically applicable annotation models such as the Universal Dependencies (77 languages), EAGLES (11 European languages) and Multext-East (16 Eastern European and Near Eastern languages). The OLiA core ontology files[24] build the reference terminology module and include over 900 ontology classes. They contain the definitions of fundamental concepts that are commonly used to annotate syntax, morphology and morphosyntax. They are therefore well suited as the basis for an application designed to search linguistic annotations and features in corpora or lexica, independently of used annotation models and languages.

4. Looking up language resources

While browsing and searching Linked Data sets like Annohub with the SPARQL query language is reserved to computer scientists only, the new web-interface will allow non-experts to examine Annohub's metadata in detail. Search parameters include:

- Language (as ISO639-3 code)
- Tagset (e.g. PENN)
- Resource type (corpus, lexicon, ontology)
- Annotation (e.g. part-of-speech tag)
- OLiA class
  (e.g. http://purl.org/olia/olia.owl#Verb)
- Resource URL
- Provenance metadata (e.g. author, title)
- Comments made by users

4.1. Lookup by language / tagset / type / name / provenance / comment

In order to provide exact results the language information in a query has to be provided as ISO639-3 code. The ISO639-3[25] code table encompasses over 7000 languages. Code guessing from a natural language input may be included in upcoming releases. In order to distinguish unilingual, bilingual and multilingual resources the search form has an option to run a query with AND,OR (exclusive AND/OR) operators.
Currently, Annohub supports 41 annotation schemes[26]. These cover annotations commonly used for annotating corpora, as well as RDF vocabularies like OntoLex-Lemon[27], which is actually not an annotation scheme, but rather a RDF vocabulary that is widely used to model lexical data. A model query can include one or multiple annotation schemes with the above-mentioned logical operators. Available resource types include lexica, corpora and ontologies. Another category are wordnets which will be supported in future releases.[28].

[13] https://linghub.org/linghub.nt.gz
[14] https://centres.clarin.eu/restxml/
[15] https://www.openarchives.org/OAI/openarchivesprotocol.html
[16] https://opus.nlpl.eu/
[17] https://spraakbanken.gu.se/
[18] https://github.com/acoli-repo/
[19] https://annohub.linguistik.de/beta/FID-Documentation.pdf
[20] https://jena.apache.org/
[21] https://tinkerpop.apache.org/
[22] https://neo4j.com/
[23] https://acoli-repo.github.io/olia
[24] http://purl.org/olia/olia.owl, http://purl.org/olia/olia-top.owl, http://purl.org/olia/system.owl
[25] https://iso639-3.sil.org/code_tables/download_tables
[26] Alpino, Ancorra, Brown, Connexor, Dzongkha, Eagles, Emille, Genia, Iiit, Ilposts, Lassyshort, Lexinfo, Mamba, Mamba-Syntax, Morphisto, MULTEXT-East, Ontolex, Penn, Penn-Syntax, Ppcme2, Proiel, Qtag,Russ, Russleeds, Sfb632, Stanford, Stts, Suc, Susa, Tcodex, Tibet, Tiger, Tiger-Syntax, Treetagger, Tueba, Urdu, Ycoe, Ubycat, UBY-POS, UD-POS, UD-Dependencies (Universal Dependencies), located at https://github.com/acoli-repo/olia
[27] https://www.w3.org/2019/09/lexicog/
[28] The resource classification process is described in (Abromeit et al., 2020)

In addition, querying resources by URL, provenance data (e.g. author, title, etc.) or comments made by users, is implemented as a full-text query on all provenance attributes / posted comments.

4.2. Lookup by annotation / OLiA class

Words in corpus or lexicon data have tags (strings) or classes (URLs) attached to, that are used to classify them. For example, the tag *Pp3fpi* is used to mark instrumental-case in the Multext-East annotation scheme. Examples[29] for the usage of OLiA annotation classes (URLs) can be found in corpus data that is annotated with the NLP Interchange Format (NIF)[30]. The OLiA ontologies cover over 30.000 annotation tags. By means of the search forms (see Fig.1, 2) resources can be located that explicitly contain an occurrence of a tag or an OLiA annotation class.

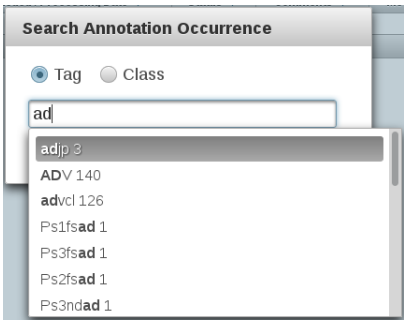

Figure 1: Annotation tag search form

By selecting a tag / OLiA class the number of resources is shown that contain a reference to it.

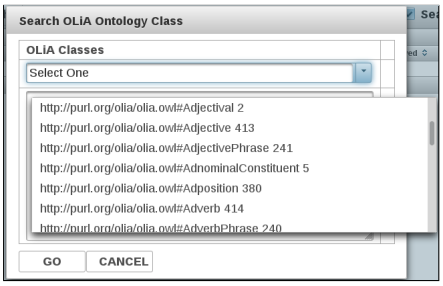

Figure 2: OLiA class search form

5. Contributing to Annohub

In order to benefit from the input of the language community, the portal offers an upload-service that allows registered users to analyze language data. Supported data types include Linked Data formats like rdf, nt, n3, etc., CoNLL[31] style data and to some degree XML encoded data[32], also as part of zip, tar and gzip archives. Limits on the size and amount of data files a user can upload are granted individually. Uploading works by providing the download URL of a language resource.[33] Before an upload is started it is checked if a resource is already contained in the catalogue or has been previously unsuccessfully processed. For this purpose the download URL, HTTP header information (e.g. *etag* information [34]) as well as MD5 and SHA256 hashes of already processed resources are kept in a database. Nevertheless, further manual duplicate checking has to be applied since a resource can have different versions and is possibly hosted at multiple locations. Finally, new resources will be queued for processing and progress information as well as the analysis results can be examined in the web-browser. In addition, registered users can comment on individual datasets listed on Annohub. Based on this feedback corrections can be made and it is decided by the reviewers at the linguistic portal[35] which user uploaded datasets will be included in the official Annohub RDF release[36]. General requirements for language resources to be included in the Annohub release are:

- A resource is publicly available via an URL as a downloadable file
- A resource is in RDF, CoNLL or XML format
- A resource includes word annotations from the syntactical or lexical domain. Otherwise only language information will be extracted
- Provenance metadata like a description for a dataset and author, licence, etc. information is provided

Because Annohub does not host the uploaded language resources, but merely the extracted metadata from it, anybody can upload data, despite of any license restrictions. Since the ability to upload content to a website poses a severe risk to fraud, by creating manipulated data packages with the intention to hack services, possible threats have to be carefully investigated.

6. Performance analysis

A qualitative analysis of the automatic tagset and language detection for CoNLL data is presented in (Abromeit and Chiarcos, 2019). Here, we focus on the analysis speed for three different data formats used for

[29] https://lider-project.eu/sites/default/files/referencecards/NIF-Corpus-reference-card.pdf

[30] https://persistence.uni-leipzig.org/nlp2rdf/

[31] https://www.signll.org/conll

[32] For a description of the supported XML data formats please see(Abromeit et al., 2020), chapter 6.1

[33] The processing of URL lists is supported as well

[34] https://docs.w3cub.com/http/headers/etag.html

[35] https://linguistik.de

[36] https://annohub.linguistik.de

language data, namely RDF, XML and CoNLL. Runtime is crucial, especially when large numbers of files with unknown content have to be processed in an unsupervised fashion, which is the case for any uploaded content to Annohub, but also applies when processing harvested file lists from CLARIN centers or other language resource metadata providers. A problem that occurs with language data encoded in RDF and XML formats is, that these formats are also widely used for non-linguistic purposes. Therefore, sampling techniques have been implemented in order to rule out unusable data quickly, but also to minimize computation times when processing large files or large collections of files (e.g. in tar archives) by testing a small fraction of a file first and by limiting the total number of data files to be processed.

6.1. Processing RDF files

RDF data is processed in a streamline fashion by utilizing the Apache Jena streaming interface[37]. This has the advantage that RDF files do not have to be loaded into a dedicated RDF triple store, which can take long for large datasets. In a first step the RDF data is validated[38] for correct URI specification of the included triples (checking forbidden characters), because this may lead to processing errors later. In case a non-conform URI is found, the RDF data is then converted to an RDF-XML representation by means of the rapper[39] RDF-utility. This has proven to fix any issues reliably. After these prepossessing steps the actual parsing of the RDF data starts. More details about the parsing process can be found in (Abromeit et al., 2020).

6.2. Processing CoNLL and XML files

The CoNLL file format is a tabular data format (TSV), where each line contains a word together with lemma, annotation and dependency information (see `https://universaldependencies.org/guidelines.html`). Parsing a CoNLL file works by identifying first the type of data included in the individual columns, because the CoNLL data format is not standardized to a certain order or number of columns (e.g. extra columns can be used to include language specific annotations). Subsequently, the language used in the word and lemma column as well as the annotation schemes used in 'annotation' columns are determined. XML files are treated in the same way as CoNLL files after they have been converted from the XML format to a CoNLL representation.

6.3. Evaluation

Table 1 shows the computation times for some well known datasets. Tests were performed on a Xeon server CPU (quad-Core) with 20GB RAM. The processing time in the last column of the table is composed of three parts (a) download time (b) validation time (only RDF) and (c) the time for NLP analysis. For better comparison, (a) and (b) are omitted for the RDF files. Download times for the CoNLL and XML examples could be neglected.

- All triples in a RDF file are examined. Since the runtime scales linear with the number of triples this alone can explain the different runtimes. A second performance factor is the number of database writes which scales linear with the amount of identified tags[40]. Since lexica generally do not contain word annotations, but rather word definitions in different languages (Wiktionary: eng, Wordnet: eng, DBnary (de): 515 languages), this factor is rather small[41]. A substantial part of the computation time is spent for validating a dataset before parsing (Wordnet: 20s, Wiktionary: 110s, DBnary: 120s). However, disabling the validation step could lead to errors while parsing, with finally no results.

- Similarly to RDF files, the runtime for CoNLL files scales linear with the number of words in a dataset. However, the extraction process for annotation data is much simpler than for RDF and XML files, since tags only have to be read from a column of a tsv file. In fact, the runtimes for the two example CoNLL files are nearly identical, although one of them is 3 times larger and also has more database writes.

- For each XML file a sample of 5000 sentences was used. The different runtimes can be explained with the number of database write operations.

Dataset	Type	Tripples/Lines	Writes	t[s]
Wordnet[42]	RDF lexicon	2637168	6	53[43]
Wiktionary[44]	RDF lexicon	3501697	41	123[45]
DBnary[46]	RDF lexicon	11267006	79	190[47]
UD_Hindi-HDTB[48]	CoNLL corpus	320968	385	42
UD_Arabic-NYUAD[49]	CoNLL corpus	90286	221	44
kubhist-stockholms-posten[50]	XML corpus	2812692	525	48
Pride and Prejudice[51]	XML corpus	339639	1683	120

Table 1: Annohub processing times

[37] `https://jena.apache.org/documentation/javadoc/arq/org/apache/jena/riot/system/StreamRDF.html`

[38] Jena command-line-tool *riot –validate*

[39] `http://librdf.org/raptor/`

[40] The persisted annotation data includes matched, but also unmatched annotations (for CoNLL and XML data only). Storing unmatched annotations ensures that these can later be automatically matched if the database if updated with an appropriate OLiA annotation model description that includes the definition of a formerly unknown tag

[41] Nevertheless, there exist RDF corpora as well

7. Summary & outlook

We introduced a new web portal that hosts metadata of publicly available annotated language resources. In addition to automated harvesting processes for such resources, and following the crowd-sourcing idea, registered users of the portal can contribute to Annohub by uploading datasets in order to extend the metadata in the Annohub catalogue which is released as a Linked Data set. The portal (**`https://annohub.linguistik.de/beta`**[52] is currently in the beta testing phase. Guest users (*login=acoli* and *password=guest*) can search all released resources in the Annohub dataset, but can not upload data or post comments. For registration as a beta-tester, please contact us with some information about your research interests. The source code of the project will be available at `https://github.com/ubffm/Annohub` under MPL 2.0 license.

Additional services can be provided in future releases, for example to convert language data listed in Annohub into a different format and make it available for download. For example from XML to CoNLL or CoNLL-RDF[53] format. Furthermore, providing a SPARQL endpoint, in order to query datasets listed in Annohub directly, could ease access to language data for researchers even more. This effort however, would require a considerable powerful technical infrastructure, which is not available right now. Finally, existing OLiA annotation models are steadily refined, but also new OLiA models will be added over time to cover yet unsupported annotation schemes.

[42] `http://wordnet-rdf.princeton.edu/wn31.nt.gz`

[43] 68s, including download and RDF-validation

[44] `https://lemon-model.net/lexica/wiktionary_en/en/en.nt.gz`

[45] 273s, including download and RDF-validation

[46] `https://kaiko.getalp.org/static/ontolex/latest/de_dbnary_ontolex.ttl.bz2`

[47] 275s, including download and RDF-validation

[48] `https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-3424/ud-treebanks-v2.7.tgz` UD_Hindi-HDTB/hi_hdtb-ud-train.conllu

[49] `https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-3424/ud-treebanks-v2.7.tgz` UD_Arabic-NYUAD/ar_nyuad-ud-test.conllu

[50] `https://spraakbanken.gu.se/lb/resurser/meningsmangder/kubhist-stockholmsposten-1830.xml.bz2`

[51] `https://opus.nlpl.eu/download.php?f=Books/v1/parsed/en.zip`, Austen_Jane-Pride_and_Prejudice.xml

[52] Not `https://annohub.linguistik.de/de/beta` !

[53] `https://github.com/acoli-repo/conll-rdf`

8. Acknowledgements

The research described in this paper was conducted in the context of the Specialized Information Service Linguistics (FID), funded by German Research Foundation (DFG/LIS, 2017-2022). The author would like to thank Christian Chiarcos for providing expert advice throughout the project. We would also like to thank Thorsten Fritze and Yunus Söyleyici for technical support and Vanya Dimitrova for helpful comments.

9. Bibliographical References

Abromeit, F. and Chiarcos, C. (2019). Automatic Detection of Language and Annotation Model Information in CoNLL Corpora. In Maria Eskevich, et al., editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 23:1–23:9, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Abromeit, F., Fäth, C., and Glaser, L. (2020). Annohub – annotation metadata for linked data applications. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 36–44, Marseille, France, May. European Language Resources Association.

Bird, S. and Simons, G. (2001). The OLAC metadata set and controlled vocabularies. In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources*.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data: The story so far. *International Journal on Semantic Web and Information Systems*, 5:1–22, 07.

Chiarcos, C., Fäth, C., Renner-Westermann, H., Abromeit, F., and Dimitrova, V. (2016). Lin|gu|is|tik: Building the Linguist's Pathway to Bibliographies, Libraries, Language Resources and Linked Open Data. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4463–4471, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Chiarcos, C., Fäth, C., and Ionov, M. (2020). The ACoLi dictionary graph. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France, May. European Language Resources Association.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data - Representation, Generation and Applications*. Springer.

Hinrichs, E. and Krauwer, S. (2014). The CLARIN research infrastructure: Resources and tools for eHumanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1525–1531, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

McCrae, J. P. and Cimiano, P. (2015). Linghub: a linked data based portal supporting the discovery of language resources. In *SEMANTiCS*.

McCrae, J. P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., and Cimiano, P. (2015). One ontology to bind them all: The meta-share owl ontology for the interoperability of linguistic datasets on the web. In *MSW@ESWC*.

Piperidis, S. (2012). The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 36–42, Istanbul, Turkey, May. European Language Resources Association (ELRA).