Data search and discovery in language data spaces: challenges and solutions

Stelios Piperidis[1], Penny Labropoulou[1], Georg Rehm[3]

[1,2] Institute for Language and Speech Processing – Athena Research Centre

[3]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

{[1]spip, [2]penny}@athenarc.gr, [3]georg.rehm@dfki.de

Introduction

Language research and technology development has been supported by a fair number of initiatives promoting language data sharing and preservation for the last almost three decades. A wide range of catalogues, repositories, platforms and, in general, infrastructures (hereafter collectively termed as "catalogues") support the publication and dissemination of resources, which can be classified along various parameters. Institutional catalogues hosting all types of resources (publications, datasets, tools, etc.) produced by practitioners affiliated to an institution, disciplinary catalogues restricted to resources produced by specific communities (e.g., OLAC[1] for resources related to language and linguistics, CLARIN[2] and ELRA[3] for language resources, Europeana[4] for cultural works, ELIXIR[5]for bioinformatics, LLOD cloud[6] for linguistic linked data, etc.), catalogues that collect specific content types (e.g., Huggging Face[7] for Machine Learning models and datasets, ELRC-SHARE[8] for Machine Translation related resources, portals for government data), to name a few. In addition, a registry of research data repositories, mostly academic repositories, is maintained by Re3data[9] and registries for metadata schemas and vocabularies (e.g., RDA Metadata directory[10], Linked Open Vocabularies[11]).

At the same time, we witness a strong movement for the sharing of resources from various sources and disciplines through a common endpoint, so that they are easily discoverable, accessible and re-usable by all, fostering interdisciplinary research and cross-community collaborations. Google has implemented Dataset Search[12], a service dedicated to facilitating discovery of datasets stored across the Web based on a simple keyword search. The European Open Science Cloud (EOSC)[13], launched by the European Commission, is conceived as a federated and open multi-disciplinary environment for hosting and processing research data and all other digital objects produced along the research life cycle (e.g., methods, software and publications). Gaia-X[14] seeks to establish an ecosystem in which data is made available, collated and shared in a trustworthy environment, associated with the concept of "data

[1] http://www.language-archives.org/
[2] https://www.clarin.eu/
[3] http://elra.info/en/
[4] https://www.europeana.eu/en
[5] https://elixir-europe.org/
[6] https://linguistic-lod.org/llod-cloud
[7] https://huggingface.co/
[8] https://www.elrc-share.eu/
[9] https://www.re3data.org/browse/
[10] https://rd-alliance.github.io/metadata-directory/standards/
[11] https://lov.linkeddata.es/dataset/lov/
[12] https://datasetsearch.research.google.com/
[13] https://eosc-portal.eu/
[14] https://www.gaia-x.eu/

spaces", a type of data relationship between trusted partners, each of whom apply the same high standards and rules to the storage and sharing of their data.

All these initiatives offer catalogues, or inventories, employing, in many cases, different metadata schemas for the documentation of the resources, due to a) the varying requirements set by the different objects of description (e.g., dataset vs. software or publication or geospatial data), b) the need to cover a wide range of users (for general catalogues) in contrast to the specialized descriptive practices common among scholars of a discipline, c) the different purposes that catalogues may serve (e.g., preservation, dissemination, or processing). Enabling sharing of resources across catalogues presupposes interoperability of the metadata documenting them; initiatives for the adoption of common standards in metadata vocabularies, documentation of the vocabularies themselves, and the creation and publication of crosswalks and mappers between them are among the primary actions in order to achieve such interoperability.

Equally important is the establishment of services for sharing metadata, and/or sharing of the actual resources themselves based on standard protocols. The OAI-PMH protocol[15] is one of the most popular mechanisms used for repository interoperability at the metadata level. The ResourceSync[16] specification is a framework for the synchronization of both metadata and resources. Finally, APIs are frequently offered nowadays as a solution for querying, and retrieving metadata records, while SPARQL services constitute the standard means for accessing and retrieving metadata from catalogues published following the Linked Data paradigm.

In the emerging concept of language data spaces, enabling crosswalks between metadata catalogues of repositories, platforms and infrastructures such as the examples mentioned above, presupposes establishing interoperability bridges to enable cross-catalogue search and discovery. In this paper, we refer to a set of actions taken in order to enable such search and discovery through metadata aggregation in the framework of the European Language Grid Platform (ELG)[17]. The ELG Platform and its Catalogue are based on the ELG-SHARE metadata schema, an evolution and essentially an application profile of the META-SHARE schema. The catalogues of interest and under investigation are disciplinary, targeting the LT/NLP and neighbouring areas (Machine Learning, Artificial Intelligence, Social Sciences and Humanities), but also general repositories and catalogues, like Zenodo[18].

Depending on the contents, metadata schemas and vocabularies used, as well as export functionalities of the source catalogues, we have experimented with different approaches, briefly sketched in the following use cases.

Harvesting metadata using OAI-PMH and closely related schemas.

To harvest catalogues that already share their metadata with the OAI-PMH protocol, ELG has implemented a client that accepts metadata records compliant with the minimal version of the ELG schema. This has already been used in two cases with catalogues that expose records compliant with a schema version based on the META-SHARE model, hence facilitating the adaptation of the mappers.

[15] https://www.openarchives.org/pmh/
[16] http://www.openarchives.org/rs/1.1/resourcesync
[17] https://live.european-language-grid.eu/
[18] https://zenodo.org/

The CLARIN (Common Language Resources and Technology Infrastructure) Research Infrastructure consists in a federated network of centres, targeting Social Sciences and Humanities. As part of the technical interoperability specifications, CLARIN data repositories are required to expose their metadata records to the Virtual Language Observatory[19] using OAI-PMH. LINDAT/CLARIAH-CZ, one of these centres, exposes metadata in a META-SHARE compliant form on which the ELG schema is based. The repository solution of the LINDAT/CLARIAH-CZ is deployed by other CLARIN centres, too, thus making the process replicable in very few steps.

OAI-PMH harvesting is also deployed for bridging to the ELRC-SHARE repository, which is used for storage and access to language resources collected through the European Language Resource Coordination[20] initiative, and uses an application profile of META-SHARE tuned to text resources for Machine Translation purposes.

Querying using custom APIs and proprietary schemas

A different procedure has been tried for catalogues that expose metadata records through custom APIs and proprietary metadata schemas, as is the case of Hugging Face (HF)[21], which includes a large collection of Machine Learning (ML) models and datasets, used for training models with a focus on transformers. HF encourages users to add descriptions of their resources in the form of a "card" (different for datasets and models), with a combination of free text fields and a set of tags (e.g., language, licence) with values from recommended controlled vocabularies, which are, however, not strictly validated. HF exposes two distinct APIs with JSON files for datasets and models respectively, with only a subset of the metadata elements and no guarantee that these are filled in for all records. To fulfil the aggregation goals set, records are imported from HF only when the description, language and licence elements are filled in. A conversion process has been set up based on the mapping of the elements and, in the case of controlled vocabularies, their values. Further enrichment of the resulting records has been performed with semi-automatic means for specific elements, most prominent being the licencing information, where ELG requires, besides the name of the licence, a URL with the text of the licence. Finally, where required, default values have been used for mandatory elements whose values could not be inferred from the original metadata records.

*Harvesting general catalogues using standard schema*s

Language and language technology related datasets are also included in general catalogues like Zenodo, a repository established and run by CERN, created in response to the European Commission's (EC) assignment to the OpenAIRE project for storing and sharing EC funded research outcomes in support of Open Science. The uptake of Zenodo by researchers for the upload of datasets, and, most recently, software, makes it interesting for language technology purposes. Zenodo exposes the metadata records in two channels: a) through a REST API, which outputs records as JSON files, and b) an OAI-PMH API in a set of standard metadata formats, namely DC[22], DataCite[23], MARC21[24] and DCAT[25].

[19] https://vlo.clarin.eu
[20] https://lr-coordination.eu/
[21] https://huggingface.co/
[22] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[23] https://schema.datacite.org/meta/kernel-4.4/
[24] https://www.loc.gov/marc/bibliographic/
[25] https://www.w3.org/TR/vocab-dcat-3/

Due to restrictive properties of Zenodo's OAI-PMH endpoint, a combined solution was used whereby a full dump (from the REST API) was filtered according to resource-type, and then OAI-PMH is used for incremental harvesting. Metadata conversion is based on the DCAT metadata schema, currently one of the most popular schemas across repositories, with the richest information among the ones exposed from Zenodo. Mappers built for Zenodo can be reused for metadata aggregation from other DCAT based catalogues; however, the DCAT vocabulary comes with various application profiles which may use non-DCAT metadata elements and/or impose specific constraints on cardinality and/or value vocabularies for specific elements. Thus, any new catalogue may require further calibration of the mapper.

Challenges in metadata aggregation and cross-catalogue search and discovery

Interoperability of repositories and their catalogues proves to be of utmost importance in a federated environment of data and services, as envisaged in initiatives like the EOSC and the emerging Data Spaces, and propounded by the FAIR principles[26]. We briefly discuss here below some of the challenges we faced in our aggregation processes and the issues that need to be addressed.

Technical interoperability across repositories: sharing of metadata and the construction of a "common" catalogue, a sort of metadata space, presuppose the availability of exchange services. OAI-PMH, despite its confinement to metadata exchange, constitutes the most widespread and usually preferred option. REST and SPARQL services are becoming more popular, with the underlying metadata schemas non standardized, customized solutions are called for.

Semantic interoperability across repositories: The use of "shared" vocabularies for the documentation of resources is a necessary step towards interoperability. Standardization and documentation of metadata schemas is a requirement articulated by many initiatives. Certain metadata vocabularies (e.g., DC, DCAT, schema.org, DataCite) have become more or less de facto standards. Still, these are general schemas and can be used to generically express core metadata elements of any type of resource, competing with the more fine-grained documentation needs of communities and more detailed requirements set to achieve machine actionability. For example, "resource type" is an element that poses problems for all catalogues: in contrast to the general vocabularies (e.g., COAR resource type vocabulary[27], DCMI Type vocabulary[28], Zenodo) specific communities prefer finer distinctions, thus creating a burden when moving from more general to more specialized catalogues (e.g., from Zenodo to ELG).

Issues that affect metadata interoperability can be attributed to the use of different metadata elements for similar concepts, different data types (e.g., free text vs. controlled vocabulary) and value spaces (different vocabularies) of the elements, as well as to the different granularity level of metadata schemas and of the metadata elements themselves. This is due to established practices of different communities and/or serving different documentation needs. For instance, for the "language" property, a value taken from the ISO 639 standard may suffice for general catalogues, but for language-related catalogues, a more detailed value space is required, one that takes into account regional and other variants, language varieties and dialects (which are not included in the ISO 639). In ELG we have decided to use the BCP

[26] https://www.nature.com/articles/sdata201618
[27] https://vocabularies.coar-repositories.org/resource_types/
[28] https://www.dublincore.org/specifications/dublin-core/resource-typelist/

47 recommendation alongside values taken from the Glottolog[29] vocabulary; the fact that glottolog includes a mapping to ISO 639-3 values facilitates metadata exchange with catalogues that prefer the ISO 639 vocabulary or other vocabularies that include a mapping to the ISO 639 values (e.g., lexvo[30], EU language authority vocabulary[31]).

Crosswalks and mappers between the various vocabularies are built especially between the popular schemas. Yet this is not a scalable approach, as for each new vocabulary a new crosswalk has to be built. Instead, an "open shared semantic space" where these metadata concepts can be mapped is needed. Selecting a single ontology which could be used in this space to cover all metadata concepts is an impossible task. Instead, this space conceived as a common catalogue of metadata concepts (elements and values from various vocabularies) linked to each other (following the Linked Data paradigm) can pave the way to semantic interoperability. In this space, links between concepts of standard vocabularies with a general scope (e.g., DC, DCAT, schema.org, prov-o) can act as the seed. Metadata concepts from community-specific vocabularies are maintained and curated by the relevant communities; links (through similarity and broader relations) can be established among concepts of community-specific vocabularies as well as with the more general concepts.

Minimal metadata requirements: The different targets of the various catalogues have an impact on metadata exchange. Zenodo, for example, is used for the publication of research outcomes by many individuals, it requires a few mandatory elements and providers do not have strong incentives to make their resources findable, thus metadata quality is rather lighter. Training and incentivizing resource owners on the importance of metadata together with continuous curation is a possible solution. Semi-automatic methods for metadata enrichment by extracting information from the datasets themselves, as well as other sources, will also play an important role in ensuring that minimal documentation requirements are met.

Cross-platform sharing of metadata records: supporting cross-platform search by offering search and discovery APIs used by a platform to third parties would allow their integration in third parties' own search spaces. This way, a query would return matches from all platforms whose search APIs are integrated in the platform queried by the user. In this case, search results would possibly show only a minimal set of metadata redirecting the user to the platform that offers the respective resource for richer descriptions. A shared common semantic space is required but only for a limited set of metadata (similarly to the general catalogues case above). Scalability will probably be an issue as soon as the collaborating platforms and search APIs grow in numbers. In this respect, decentralized federated infrastructures such as Gaia-X, where individual trusted platforms following a common standard (the Gaia-X standard) become a networked system freely sharing and exchanging data and services across multiple actors, offers a viable solution addressing this challenge.

[29] https://glottolog.org/
[30] http://www.lexvo.org/
[31] https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/language