To be findable, accessible, interoperable and reusable: language data and technology infrastructure for supporting the FAIR data approach

Danguolė Kalinauskaitė
Faculty of Humanities, Centre of Computational Linguistics
Vytautas Magnus University
Kaunas, Lithuania
e-mail: danguole.kalinauskaite@vdu.lt

***Abstract*–In this paper the FAIR principles are summarized, with a focus on (meta)data management infrastructure as the conduit for developing these principles in practice. The paper serves as an overview of the European Open Science Cloud in supporting the FAIR data approach in general, and as an overview of the Lithuanian practice in supporting the sharing, use and sustainability of language resources specifically: the CLARIN infrastructure as a networked federation of repositories and service providers is presented, Lithuanian experience in developing national language resources infrastructure is detailed, and the practical benefits that Lithuania gained with joining the European infrastructure are reviewed.**

Keywords–FAIR; data management; infrastructures; European Open Science Cloud; CLARIN; language resources

I. INTRODUCTION

It is reasonably stated that humans and machines often face distinct barriers when attempting to find and process data on the web. This is because humans have an intuitive sense of "semantics" (the meaning or intent of a digital object): they are capable of identifying and interpreting a wide variety of contextual cues, whether those take the form of structural/visual/iconic cues in the layout of a web page, or the content of narrative notes. The primary limitation of humans, however, is that they are unable to operate at the scope, scale, and speed necessitated by the scale of contemporary scientific data and complexity of e-science [1]. This is due to the fact that humans increasingly rely on computers to undertake different tasks on their behalf.

Machines are necessitated to be capable of autonomously and appropriately acting when faced with the wide range of types, formats, and access-mechanisms/protocols. It also necessitates that the machines keep an exquisite record of provenance such that the data they are collecting can be accurately and adequately cited [1]. Therefore all participants in the data management process (from researchers and data producers to data repository hosts) are of paramount importance in assisting computers and improving this process.

The focus on assisting machines in their discovery and exploration of data through application of more generalized interoperability technologies and standards at the data/repository level, becomes a first-priority for good data management. There are some guidelines, summarized as acronym FAIR that put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. They are discussed in the paper below.

The paper is organized as follows. The FAIR principles are theoretically presented in section II, practical implementation of developing these principles is specified in section III, the Lithuanian practical results in supporting the FAIR approach are detailed in section IV.

The goal of this paper is to present infrastructures as one of the means to overcome data discovery and reuse obstacles in order to be more FAIR, with a focus on Lithuanian experience in developing language resources infrastructure.

II. TO BE "FAIR"

A. Cooperation of humans and machines

Data-intensive sciences meet the challenges to facilitate knowledge discovery. Computational analysis to discover meaningful patterns in massive, interlinked datasets is rapidly becoming a routine research activity. However datasets should be prepared for such analysis, so this kind of activity, in turn, urges an ever closer cooperation of humans and machines in the access to, integration and analysis of, scientific data.

Providing machine-readable data as the main substrate for knowledge discovery and for the e-scientific processes to run smoothly and sustainably is considered to be one of the biggest challenges of e-science. As it is observed, the reason that we often need several weeks (or months) of specialist technical effort to gather the data necessary to answer research questions is not the lack of appropriate technology; the reason is, that we do not pay our valuable digital objects the careful attention when we create and preserve them [1]. So it is worth to note here that cooperation of humans and machines needs to follow certain principles to enable optimal use of data and methods.

B. FAIR principles

A key enabler to achieve international-grade data management is for research data and information to be published in a "FAIR" manner. "FAIR" summarizes several aspects, or in other words, a set of guiding principles, regarding data, both for machines and for people, and these principles assist the interaction between those who want to use data and those who provide them. In the FAIR data approach, data should be [1]:

- *Findable (F)*

F1. (Meta)data are assigned a globally unique and eternally persistent identifier.

F2. Data are described with rich metadata (defined by R1 below).

F3. (Meta)data are registered or indexed in a searchable resource.

F4. Metadata specify the data identifier.

- *Accessible (A)*

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1. The protocol is open, free, and universally implementable.

A1.2. The protocol allows for an authentication and authorization procedure, where necessary.

A2. Metadata are accessible, even when the data are no longer available.

- *Interoperable (I)*

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles.

I3. (Meta)data include qualified references to other (meta)data.

- *Reusable (R)*

R1. Meta(data) have a plurality of accurate and relevant attributes.

R1.1. (Meta)data are released with a clear and accessible data usage license.

R1.2. (Meta)data are associated with their provenance.

R1.3. (Meta)data meet domain-relevant community standards.

So putting data on the web is not enough. The FAIR principles speak about a "knowledge representation" language for data representation. To be actually interoperable and reusable, data should not only be properly licensed, but the methods to access and/or download them should also be well described and preferably fully automated and using well established protocols [2]. Any combination of the FAIR principles is desirable and of paramount importance (especially regarding data management in small countries), and it is worth to note here that these principles are intended to apply not only to "data" in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. In other words, the principles suggest that contemporary data resources, tools, vocabularies and infrastructures should exhibit specific characteristics, norms, and practices in order to assist discovery and reuse by third-parties and to be considered "FAIR" [1]. In this way, the FAIR principles provide steps toward machine-actionability.

The idea of being machine-actionable applies in two contexts: first, when referring to the contextual metadata surrounding a digital object ("what is it?"), and second, when referring to the content of the digital object itself ("how do I process it/integrate it?"). Here also to be drawn a distinction between machine-actionable data as a result of specific investment in software supporting that data-type, and data that are machine-actionable exclusively through the utilization of general-purpose, open technologies [1]. The paper details the latter.

FAIRness can be achieved with a wide range of technologies and implementations. One of the solutions to supporting FAIRness is data infrastructures. The FAIR principles guide the development of infrastructure and tooling to make data optimally reusable for machines and people alike, which is considered to be a crucial step forward.

III. INFRASTRUCTURES FOR SUPPORTING THE "FAIR" DATA APPROACH

Infrastructure in general is defined as (usually large-scale) basic physical and organizational resources, structures and services needed for the operation of a society or enterprise [3]. Specifically, a research infrastructure refers to an infrastructure intended for carrying out research: facilities, resources and related services used by the scientific community to conduct top-level research [3]. So a research infrastructure covers various means for researchers in different fields. One of those means are digital resources, including both data and software. Infrastructure initiatives enable FAIR principles implementation in practice and are of paramount importance. Some of such initiatives are discussed in the rest of this paper.

A. European Open Science Cloud

The European Commission suggested "European Cloud initiative" [4] (issued in April 2016) which set an ambitious vision for the European Open Science Cloud (EOSC) (see Fig. 1.) to be realised by 2020.

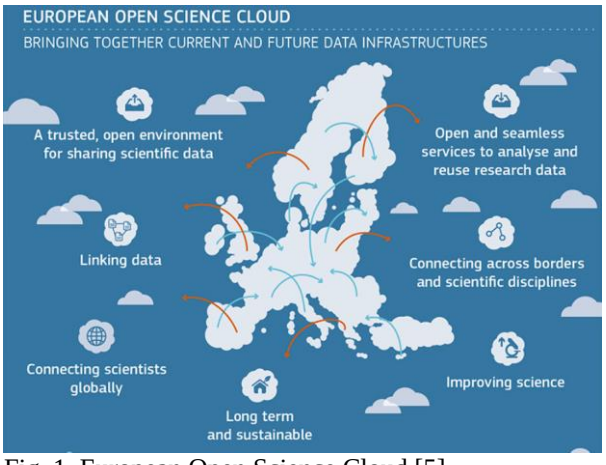

Fig. 1. European Open Science Cloud [5]

The EOSC is a data infrastructure to support Open Research Data and Open Science in Europe: it is intended as an open and trusted environment where research data can be safely stored and made openly available, and is dedicated to enable trusted access to services, systems and the re-use of shared scientific data across disciplinary, social and geographical borders [6]. In other words, the EOSC serves as an ecosystem of infrastructures and although it is indeed a European infrastructure, it is aimed to be globally interoperable and accessible.

B. EOSC implementation and opportunities

The European Open Science Cloud will start by federating existing scientific data infrastructures, today scattered across disciplines and geographically. This step will make access to scientific data easier, cheaper and more efficient. It will enable the creation of new market opportunities and new solutions in key areas such as health, environment, or transport [4].

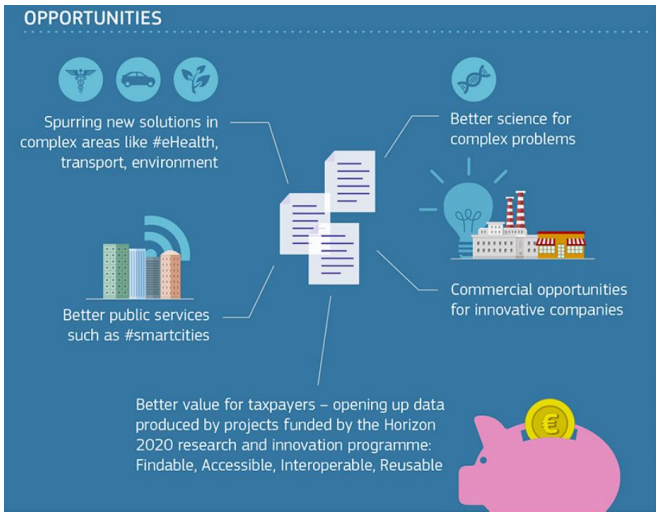

Fig. 2. EOSC opportunities [5]

The European Open Science Cloud will be also open for education and training purposes in higher education and, over time, to government and business users as the technologies developed will be promoted for wider application [4]. The EOSC includes the required human expertise, resources, standards, best practices as well as the underpinning technical infrastructures [6]. In this way, the EOSC is intended to bring tangible benefits to society (see Fig. 3.).

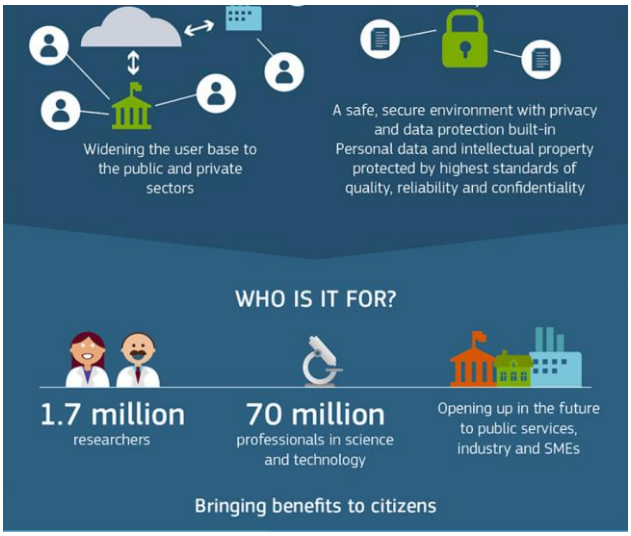

Fig. 3. EOSC benefits [5]

Practically, the European Open Science Cloud will offer 1.7 million European researchers and 70 million professionals in science and technology a virtual environment with free at the point of use, open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines [4]. In other words, this infrastructure is about a federated environment for scientific data sharing and re-use, based on existing and emerging resources in the members of the infrastructure. It will provide only lightweight international guidance and governance and a large degree of freedom regarding practical implementation. So one of the most important aspects of the EOSC is systematic and professional data management and long-term stewardship of scientific data assets and services in Europe and globally [6].

While the European Open Science Cloud is still on the way, the other - European infrastructure dedicated to researches that deal with language resources - already works and is one of the good practices in creating infrastructures.

IV. "FAIR" IN LITHUANIAN PRACTICE. LANGUAGE RESOURCES INFRASTRUCTURE

In January of 2015 Lithuania became a full member of CLARIN ERIC (Common Language Resources and Technology Infrastructure) (see [7]), which is aimed to provide easy and sustainable access to digital language data (in written, spoken, video or multimodal form), and advanced tools to discover, explore, exploit, annotate, analyse or combine such data sets and tools, wherever they are located. Soon national consortium (CLARIN-LT) was founded by three partner universities: Vytautas Magnus University, Kaunas University of Technology and Vilnius University. So this part of the paper is based on reflection of Lithuanian experience in storing and accessing language resources.

The CLARIN infrastructure at the European scale is intended for the humanities and the social sciences, and these two domains, in turn, include a wide range of disciplines. It is *distributed*, i.e., implemented in a network of CLARIN

Centres, and *virtual*, i.e., it provides services via the Internet [8]. This European infrastructure covers a wide spectrum of digital data types:

- Data in natural language (texts, lexicons, grammars, etc.).
- Databases about natural language (typological databases, dialect databases, lexical databases, etc.).
- Audio-visual data containing (written, spoken, signed) language (e.g. pictures of manuscripts, audiovisual data for language description, description of sign language, interviews, radio and tv programmes, etc.) [3].

It also includes software to browse and search in digital language data (e.g. software to search in a linguistically annotated text corpus), as well as software to analyze, enrich, process, and visualize digital language data (e.g., a parser, which enriches each sentence in a text corpus with a syntactic structure) [3].

The case of CLARIN ERIC points to the means for overcoming data discovery and reuse obstacles in language technologies. The CLARIN infrastructure supports the sharing, use and sustainability of language data and tools through a networked federation of centres: language data repositories, service centres and knowledge centres, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centres are interoperable, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work. Interoperability is ensured through the standards adopted in the CLARIN framework. A large number of participating centres are offering access services to data, tools and expertise [7].

The access provided to data and content is in principle sustainable and in accordance with the principles of open science, and thus free for scholars. The same criteria for data and services interoperability, access conditions, quality of data and services are adhered to by all members, however countries have a large degree of freedom to decide what they contribute to the CLARIN infrastructure [7]. Principal scheme of the CLARIN services is presented in Fig. 4. to give a general picture of CLARIN functionality.

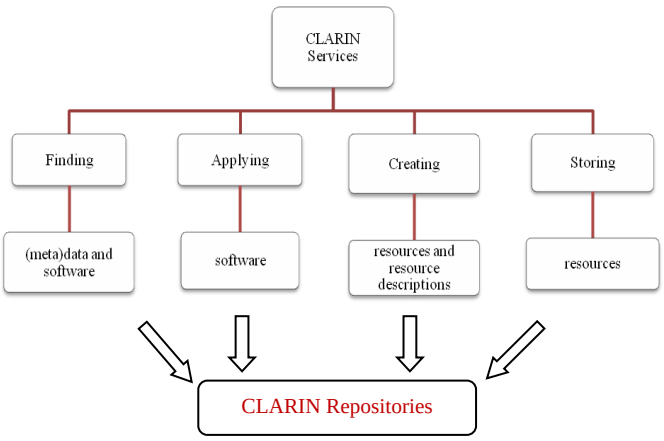

Fig. 4. CLARIN Services

As shown in the figure above, CLARIN offers a number of basic services to researchers, in particular:

- Services to find digital data and get access to digital data.
- Services to find software for processing digital data.
- Services to apply the software to the digital data in a user friendly manner.
- Services to create and describe new data and software.
- Services to store new data and software in CLARIN for long term preservation and for making them accessible to other researchers [8].

A. Infrastructure as the conduit for FAIR data

Before Lithuania became a member of the CLARIN infrastructure, its language-based resources and advanced tools were accessible via special websites. With joining the infrastructure, Lithuanian resources and services are available to the various research communities at large, and Lithuania as a member of the infrastructure gained practical benefits:

- Access to services provided by the infrastructure: existing tools and datasets, e.g. online access to national corpora, lexica, audio and video recordings, annotations, grammars, etc.
- Long-term archiving: a storage guarantee can be given for a long period (up to 50 years in some cases). Resources can be archived and made available to the community in a reliable manner.
- Persistent identifiers for data. Resources can be cited easily with a persistent identifier.
- The resources and their metadata are integrated into the infrastructure, making it possible to search them efficiently.
- Password-protected resources can be made available via an institutional login.
- Once resources are integrated in the CLARIN infrastructure, they can be analyzed and enriched more easily with various linguistic tools (e.g. automated

part-of-speech tagging, phonetic alignment or audio/video analysis).

- Practical workshops at which a diverse group of people, including those with research questions and software developers, get together to work on producing some real solutions over a short but intense period. The outputs of those workshops - tools, datasets, linked data, etc. that are to be maintained. Such activity stimulates the adoption or better use of digital methods for the community of the infrastructure.
- Repository. Lithuanian language data and services are stored in CLARIN-LT Repository [9]. It contains 5 resources: 1. *Lemmatised Wordlist of 1 m. Corpus of Contemporary Lithuanian*; 2. *LITIS v.1*; 3. *Lithuanian Treebank ALKSNIS*; 4. *Lithuanian morphologically annotated corpus MATAS*; 5. *Wordlist of the Contemporary Corpus of Lithuanian language*.

B. *CLARIN-LT Repository*

Data in CLARIN-LT Repository (Fig. 5.) are made available under licences attached to resources. In case there is no licence, data is made freely available for access, printing and download for the purposes of non-commercial research or private study [9]. The Repository is like a library for linguistic data and tools: to search for data and tools and easily download them; deposit the data and be sure it is safely stored, everyone can find it, use it, and correctly cite it (*Findable*).

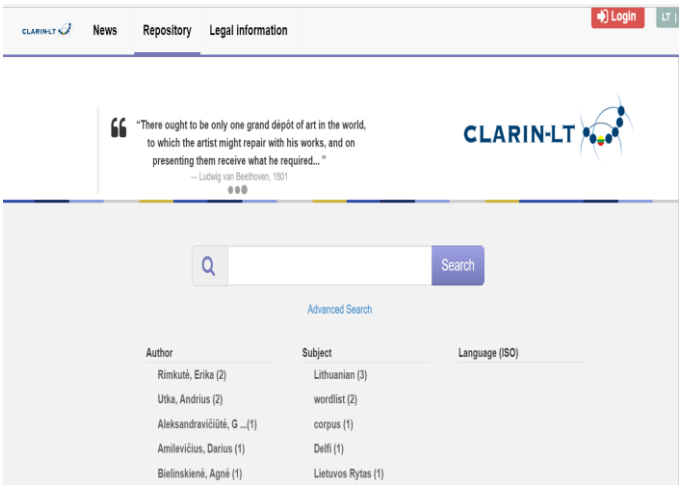

Fig. 5. CLARIN-LT Repository

So Lithuanian language resources and tools can now be reached not only via special websites of those resources and tools, but also via a specific platform provided by the CLARIN infrastructure (*Accessible*).

C. *Metadata*

An important step in the FAIR data approach is to publish existing and new datasets in a semantically interoperable format that can be understood by computer systems. By semantically annotating data items and metadata, we can use computer systems to (semi)automatically combine different data sources, resulting in richer knowledge discovery activities. Rich metadata facilitate such discovery, including clear rules regarding the process for accessing the data.

Metadata for language resources and tools exists in a multitude of formats. Often these descriptions contain specialized information for a specific research community (e.g. TEI headers for text, IMDI for multimedia collections [10]). To overcome this dispersion CLARIN has initiated the Component MetaData Infrastructure (CMDI). It provides a framework to describe and reuse metadata blueprints. Description building blocks ("components", which include field definitions) can be grouped into a ready-made description format (a "profile") (*Interoperable*). Both are stored and shared with other users in the Component Registry [11] to promote reuse (*Reusable*).

For metadata purposes, CLARIN ERIC provides a special metadata-based portal for language resources - Virtual Language Observatory [12]. It is completely based on the Component Metadata (CMDI) and ISOcat standards. This approach allows for the use of heterogeneous metadata schemas while maintaining the semantic compatibility [13].

D. *Benefits to individual groups of people and to institutions*

There are numerous and diverse stakeholders who benefit from being a part of the infrastructure:

- researchers share and reuse each other's data;
- professional data publishers offer their services;
- software and tool-builders provide data analysis and process services;
- funding agencies (private and public) increasingly concerned with long-term data stewardship;
- a data science communities mine, integrate and analyse new and existing data.

Lithuania as a member of the CLARIN infrastructure makes its resources (Fig. 6.) available and accessible in the infrastructure for other researchers. Other members, in turn, provide access to their tools and resources incorporated in the infrastructure.

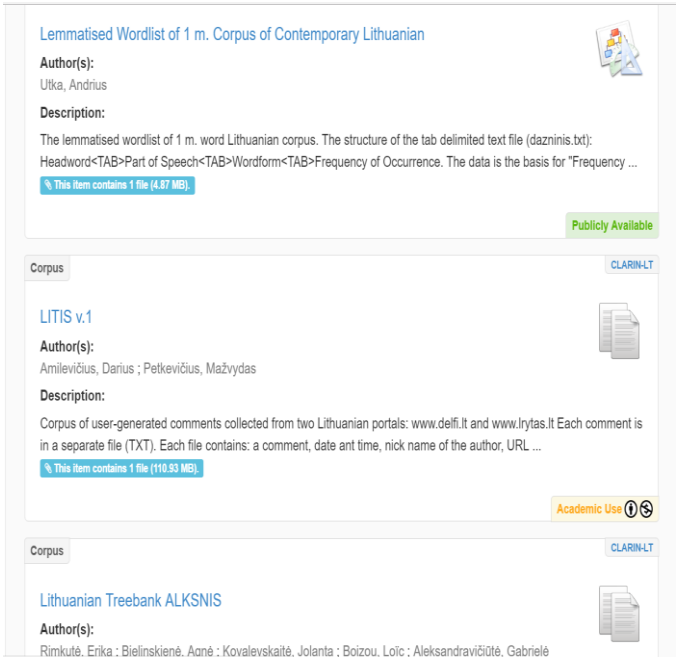

Fig. 6. Lithuanian resources in CLARIN-LT Repository

It should be stated that CLARIN has a lot to offer: it has adapted existing data and software, and it has created new easy and user-friendly software for searching, analysing and visualising data. However, there is still work to do: some improvements and extensions are needed, especially in terms of (meta)data functionality (the same applies also to CLARIN-LT).

V. CONCLUDING REMARKS

In the paper the FAIR principles were summarized, with a focus on infrastructure initiatives as the conduit for developing these principles in practice: the European Open Science Cloud for supporting the FAIR data approach was presented, and Lithuanian experience in developing language resources infrastructure was overviewed.

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process.

Infrastructures are one of the means to overcome data discovery and reuse obstacles for those wishing to become more FAIR. Lithuanian case of membership in the infrastructure shows that to be a part of infrastructures means to reach more and wider.

REFERENCES

[1] Wilkinson, M. D. *et al.* "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data 3, 15 March 2016, doi:10.1038/sdata.2016.18. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/.

[2] "Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0," FORCE11. [Online]. Available: https://www.force11.org/fairprinciples.

[3] J. Odijk, "The CLARIN infrastructure in the Netherlands: What is it and how can you use it?" 2014. [Online]. Available: http://www.clarin.nl/sites/default/files/CLARIN%20General%20Introduction.pdf.

[4] "Communication: European Cloud Initiative – Building a competitive data and knowledge economy in Europe." 2016. [Online]. Available: http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52016DC0178.

[5] "European Open Science Cloud," European Research & Innovation. [Online]. Available: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

[6] "Realising the European Open Science Cloud." 2016. [Online]. Available: https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf.

[7] CLARIN ERIC: https://www.clarin.eu/.

[8] J. Odijk, "Linguistic research using CLARIN," Lingua, 178, pp. 1-4, 2016.

[9] CLARIN-LT Repository: https://clarin.vdu.lt/xmlui/?locale-attribute=en.

[10] CLARIN ERIC Glossary: https://www.clarin.eu/glossary.

[11] CMDI Component Registry: https://catalog.clarin.eu/ds/ComponentRegistry/#/?_k=arj093.

[12] CLARIN Virtual Language Observatory: https://vlo.clarin.eu/?2.

[13] D. Van Uytvanck, H. Stehouwer, L. Lampen, "Semantic metadata mapping in practice: the Virtual Language Observatory," in Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation, N. Calzolari (Ed.), pp. 1029-1034, 2012.