Linghub: a Linked Data based portal supporting the discovery of language resources

John P. McCrae[1,2]
[1]Insight Centre, National University of Ireland, Galway
Galway, Ireland
john@mccr.ae

Philipp Cimiano[2]
[2]Cognitive Interaction Technology, Cluster of Excellence, Bielefeld University
Bielefeld, Germany
cimiano@cit-ec.uni-bielefeld.de

ABSTRACT

Language resources are an essential component of any natural language processing system and such systems can only be applied to new languages and domains if appropriate resources can be found. Currently the task of finding new language resources for a particular task or application is complicated by the fact that records about such resources are stored in different repositories with different models, different quality and search mechanisms. To remedy this situation, we present Linghub, a new portal that aggregates and indexes data from a range of sources and repositories and applied the Linked Data Principles to expose all the metadata under a common interface. Furthermore, we use faceted browsing and SPARQL queries to show how this can help to answer real user problems extracted from a mailing list for linguists.
Major comments: * Why not integrate the datasets in an already developed catalog? * How to integrate additional datasets * Information about model... give example * Paper more self-contained * Deeper description of use case

1. INTRODUCTION

Language resources are essential for nearly all tasks in natural language processing (NLP) and in particular for the adaptation of resources and methods to new domains and languages. In order to use language resources for new purposes they must first be discovered and this can only be done if there is a comprehensive list of all resources that may be available. To this there have been a number of projects that have attempted to collect such a catalogue using various methods and with differing degrees of data quality. We present a new portal, Linghub, that aims to integrate all these data from different sources by means of linked data and thus to create a website, whereby all information about language resources can be included and queried using a common methodology. The goal of Linghub is thus to enable wider discovery of language resources for researchers in NLP, computational linguistics and linguistics.

Currently, two approaches to metadata collection for language resources can be distinguished. Firstly, we distinguish a *curatorial* approach to metadata collection in which a repository of language resource metadata is maintained by a cross-institution organization such as META-SHARE [7] or CLARIN project's Virtual Language Observatory [17, VLO]. This approach is characterized though high-quality metadata that are entered by experts, at the expense of coverage. A *collaborative* approach, on the other hand, allows anyone to publish language resource metadata. Examples of this are the LREMap [4] or Datahub[1]. A process for controlling the quality of metadata entered is typically lacking for such collaborative repositories, leading to less qualitative metadata and inhomogeneous metadata resulting from free-text fields, user-provided tags and the lack of controlled vocabularies.

Given the nature of this difference we wish to make data available from multiple sources in a homogeneous manner and we saw the development of a new linked data portal as the primary method to achieve this. To this end we adopted a model based on the DCAT data model [10] along with properties from Dublin Core [9]. In addition, we used the RDF version [12] of the META-SHARE model [8] to provide for metadata properties that are specific to language data and linguistic research. As such, in this paper we describe the creation of the largest collection of information about language resources and briefly describe its publication on the Web by means of linked data principles.

The rest of the paper is structured as follows: firstly, we will describe related work in Section 2, then we will describe the collection and processing of data in Section 3. Next, we will describe the portal and how we envision users can access the data in Section 4 and examine how real user queries could be answered with Linghub in Section 5. Finally we conclude in Section 6.

2. RELATED WORK

There have been several attempts to collect metadata about language resources mostly associated with large infrastructure projects. CLARIN has been collecting resources under a project called the Virtual Language Observatory [17], using the Component Metadata Infrastructure [3, CMDI] to collect common metadata values from multiple sources. A similar project is META-SHARE [14] from the META-NET project where language resources are collected and high-quality, manual entries are created for each record. Similarly, the Open Languages Archives Community [2, OLAC] collects data from a number of sources although the metadata collected is not itself open. Another related project called SHACHI has also collected some metadata [16]. There has also been an attempt to track language resources by means of assigning them an International Standard Language Resource Number (ISLRN) similar to an ISBN used to track books [5].

[1] http://datahub.io

Source	Records	Triples
Datahub	185	10,739
LRE-Map	682	10,650
META-SHARE	2,442	464,572
CLARIN VLO	144,138	3,605,196
All	147,447	4,091,157

Table 1: Size of Linghub datasets by source

On the contrary some resources have instead collected data directly from creators of the resources, for example the LRE-Map [4] collects data from authors of papers submitted to conference, such as LREC. Similarly, Datahub collects resources directly from those submitted to the website, but focusses primarily on linked data resources.

3. DATASET

In order to ensure that all the data from many sources can be queried in a homogenous manner we made sure that the metadata from all the repositories mentioned in Table 1 was available as RDF. In doing this, we aligned the proprietary schemas used in these repositories to well-known semantic Web vocabularies and fixed existing modeling errors, such as using percent-encoded URIs for titles of resources or introducing URL links that would never resolve. Two of our resources, LRE-Map and Datahub, were already available in RDF, so that the conversion mainly involved developing an appropriate URL schema so that datasets were uniquely identified and thus to avoid collisions when uploading data into the Linghub portal. A number of quality issues were also fixed in doing this transformation, such as deciding whether property values should be literals or URIs, reducing the number of blank nodes and reusing existing metadata vocabularies such as VoID [1].

The other resources (CLARIN VLO and META-SHARE) were available in XML. We developed a custom converter for each of these resources building on a transformation language similar to XSLT, which we developed. For META-SHARE, this was a challenging task as there were nearly a thousand unique tags defined and each one was examined to see if it was similar to an existing Semantic Web vocabulary, and in fact we ended up mapping to FOAF [2], SWRC [3] and the Media Ontology [4]. In the case of CLARIN, there was actually a significant difference between the XML schemas used by each contributing instance, with only a small common section giving the resource title and download link. We thus developed distinct mappings for the largest 5 institutes.

Two key issues emerge when collecting data from a heterogenous set of sources such as we are doing. Firstly, the data is likely to be noisy and inconsistent in the properties it uses and more importantly in the values that these properties have. For example, languages may be represented by their English names or alternatively by means of the codes such as the ISO 639 codes[5].

Secondly, it happens relatively frequently that dataset descriptions are duplicate as they are contained in multiple source repositories (currently this affects 5.0% of resources). Furthermore, also intra-repository duplicates exist, resulting from the fact that in some repositories one metadata record is created for each language a resource is available in (this is the case for CLARIN for instance and represents 35.0% of all resources). In order to remove these duplications we used state-of-the-art word sense disambiguation techniques, including Babelfy [13] to identify common controlled vocabularies and duplicate entries. For the case of properties we mapped to several existing resources, including LexVo [6] for languages, and BabelNet for resource types. Duplicate entries were not removed from the dataset but instead were marked with the addition of the Dublin Core property *is replaced by*. In the case that these entries were subsets of resources the target of this link would be a new combined record for the entire resource and in the case of duplicate records collected from distinct sources we referred to the most complete triple, that is the record with the most triples. The harmonization and description is described in more detail in McCrae et al. [11].

Currently, there is no direct method for users to provide metadata to the repository, however it is foreseen that users could submit valid DCAT files to Linghub. We do note that Datahub allows any user to submit a dataset and such datasets will quickly be picked up by Linghub and added to the repository in this manner.

4. THE LINGHUB PORTAL

In order to enable users to quickly and easily discover datasets, we set up a portal for browsing the dataset. Naturally we set this up as a site that publishes the individual records as either RDF or HTML, with the actual content delivered to the client decided by means of content negotiation. We developed templates that render the RDF in a readable manner, while still appearing close to the data in such a way that users would get a consistent view of a dataset record even if it came from a different original source and hence had very different properties. In addition, we provide a number of mechanisms by which users and automated agents can discover a dataset. For users, we allowed resources to be discovered by means of faceted browsing by enabling users to select properties and their values. We fixed the list of properties in advance to those that have been harmonized so as to not overload the user with choices for properties that only occur for a few datasets and also to enable the compilation of indexes to speed up page load times. In addition the front page of Linghub contains a free-text search engine allowing the users to query fields by a property. This free-text search engine is powered by a separate index which includes not only the text of data properties but also the labels of URIs which appear as the value of object properties. Machine-based agents may access the endpoint by means of SPARQL querying, although the endpoint limits the agents to a subset of the SPARQL query language. The goal of this is to enable constant query-time without overloading our server. The nature of SPARQL makes it very easy for users to write queries that are of a complexity that would not be easy to answer and other sites have attempted to handle this by enforcing timeouts on SPARQL queries. In general we find this solution to be sub-optimal as it means that queries may fail unpredictably if the server has many concurrent connections. Instead, we limit the complexity of the queries themselves by requiring that the triples have certain properties that can be easily answered. These include:

1. A required limit on the number of results;

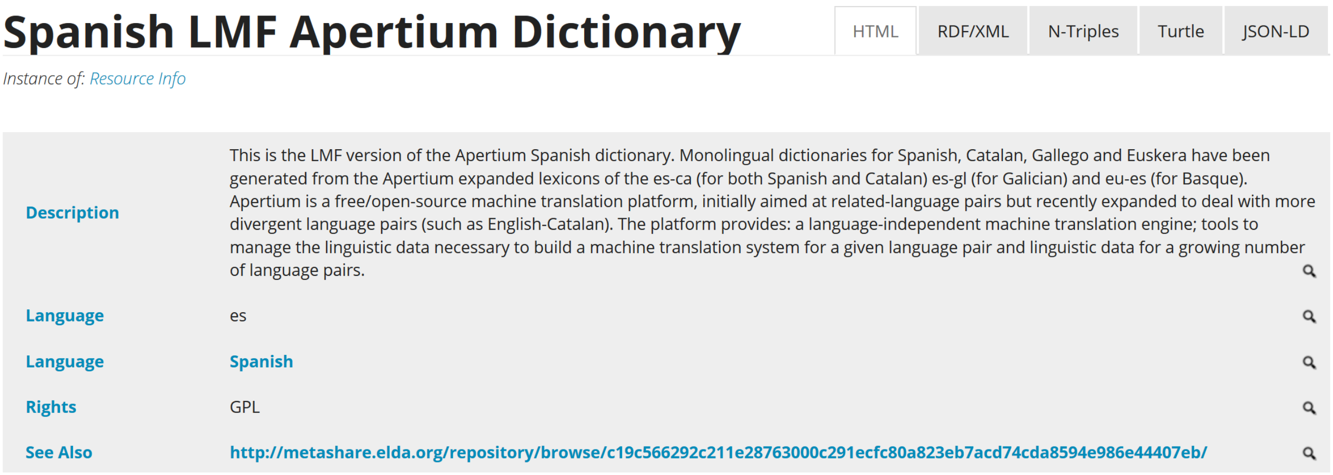
2. The property may not be a variable, thus limiting the number of results;

[2] http://xmlns.com/foaf/spec/

[3] http://ontoware.org/swrc/

[4] http://www.w3.org/TR/mediaont-10/

[5] http://www.iso.org/iso/home/standards/language_codes.htm



Spanish LMF Apertium Dictionary

Instance of: Resource Info

Description	This is the LMF version of the Apertium Spanish dictionary. Monolingual dictionaries for Spanish, Catalan, Gallego and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish and Catalan) es-gl (for Galician) and eu-es (for Basque). Apertium is a free/open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan). The platform provides: a language-independent machine translation engine; tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs.
Language	es
Language	**Spanish**
Rights	GPL
See Also	**http://metashare.elda.org/repository/browse/c19c566292c211e28763000c291ecfc80a823eb7acd74cda8594e986e44407eb/**

Figure 1: A screenshot of the Linghub interface

3. The query must be a 'tree' in that every triples should be connected from a single root node.

Furthermore, the SPARQL endpoint also by default returns SPARQL-JSON results[15], so that the results may be easily applied. This is based on the fact that many clients, notably client-side Javascript in browsers, will not accept XML due to security concerns. Other clients may still obtain SPARQL-XML by supplying the appropriate header or parameter in the query.

5. USE CASES

As a proof-of-concept for Linghub, we discuss a number of realistic use cases that demonstrate the type of queries that can be answered using Linghub. In order to get realistic use cases, we collected queries for language resources from the *Corpora List*, a mailing list used by researchers in corpus linguistics to discuss corpora. From questions posed in February 2015, 3 queries are considered below as they are clear and well-stated questions that would have feasible answers. We chose these queries to provide an illustrative example of queries that can be directly answered using the Linghub portal, while discarding many other questions that were vague, unclearly stated or misused linguistic terminology. We discuss these queries and show how they can be formalized as SPARQL queries against Linghub and discuss whether reasonable answers were retrieved.

1. "[...] desparately needs an Igbo corpus." (Thapelo J. Otlogetswe, Feb. 5th 2015[6])

   Igbo is a language of Nigeria and Equatorial Guinea and is identified with the language code `ibo`. Simply typing "Igbo" into the search interface of Linghub finds a number of resources that could be used. For many of these resources Igbo is the value of the Dublin Core property *subject*. Although, there is a language property some sources decided not to use this Dublin Core category. In addition these resources are marked with a *type* that is mapped to the META-SHARE `corpus` individual even though the resources do not originate from META-SHARE due to our harmonization. We can search for both language and subject with the following query[7]:

   ```
   SELECT ?resource WHERE {
    ?resource
      dct:language iso639:ibo |
      dc:subject "Igbo" ;
      dct:type metashare:corpus .
   }
   ```

2. "I am looking for a Lithuanian gigaword corpus for a research project." (Márton Makrai, Feb. 24th 2015[8])

   Finding a corpus for a European language such as Lithuanian is generally not a challenge, however this user also has the requirement that the resource has over one billion words. We can easily use the META-SHARE properties to return the user a list of corpora with their associated sizes, as follows:

   ```
   SELECT ?resource ?size WHERE {
     ?resource
       ms:corpusInfo [
         ms:languageInfo [
           dct:language iso639:lit ;
           ms:sizePerLanguage [
             ms:size ?size ;
             ms:sizeUnit ms:words
           ]
         ]
       ] .
   }
   ```

   Unfortunately, the results of this query show that no resource in Linghub is over one billion words in size for Lithuanian.

3. "I am looking for freely available geotagged tweets collection for research purpose." (Md. Hasanuzzaman, Feb. 16th 2015[9])

[6] `http://mailman.uib.no/public/corpora/2015-February/021993.html`

[7] Note: We have implemented some syntactic extensions to SPARQL. The `|` operator is a `UNION` with the same subject.

[8] `http://mailman.uib.no/public/corpora/2015-February/022103.html`

[9] `http://mailman.uib.no/public/corpora/2015-February/022044.html`

> Several of the search terms here are unfortunately not found anywhere in our data, namely 'geotagged' and 'tweets'. It would still be possible for this query to be answered by looking at related keywords such as 'Twitter', and other aspects of this query can be handled (e.g., 'for research purpose'), can be handled by means of the META-SHARE vocabulary.

In summary, we saw that in two of the three cases the users' need could be clearly expressed as a SPARQL query and that in one of those cases, the query would return an answer as required, in the second case no suitable dataset is recorded. In the final case, the user's query does not match the structured data found in Linghub, but related resources can be found by using free text search. As such, we see that Linghub enables users to better find their resources than with previous approaches, although it is still not satisfactory for all user queries. In particular, the crucial defect in the final query is that there is no specific metadata that would indicate if a resource is from a social media site or not, and this would require deeper understanding of the textual components of resource descriptions to better handle.

6. CONCLUSION

Linghub is a new site that collects data from a large number of sources and makes it queriable through a common mechanisms. Furthermore, the data has not only been converted to RDF it has also been homogenized and linked to other bubbles in the Linguistic Linked Open Data Cloud. As such, this resource is likely to pay a pivotal role in enabling not only humans but also software agents to find new resources and use them for applications in natural language processing and artificial intelligence.

Acknowledgments

This work has been funded by the LIDER project funded under the European Commission Seventh Framework Program (FP7-610782), the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG), and the Insight Centre which is funded by the Science Foundation Ireland under Grant Number SFI/12/RC/2289.

7. REFERENCES

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets with the VoID vocabulary. Technical report, The World Wide Web Consortium, 2011. Interest Group Note.

[2] S. Bird and G. Simons. Extending dublin core metadata to support the description and discovery of language resources. *Computers and the Humanities*, 37(4):375–388, 2003.

[3] D. Broeder, M. Windhouwer, D. Van Uytvanck, T. Goosen, and T. Trippel. CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, page 1, 2012.

[4] N. Calzolari, R. Del Gratta, G. Francopoulo, J. Mariani, F. Rubino, I. Russo, and C. Soria. The LRE Map. Harmonising community descriptions of resources. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1084–1089, 2012.

[5] K. Choukri, V. Arranz, O. Hamon, and J. Park. Using the international standard language resource number: Practical and technical aspects. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 50–54, 2012.

[6] G. de Melo. Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web*, page 7, 2013.

[7] C. Federmann, I. Giannopoulou, C. Girardi, O. Hamon, D. Mavroeidis, S. Minutoli, and M. Schröder. META-SHARE v2: An open network of repositories for language resources including data and tools. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3300–3303, 2012.

[8] M. Gavrilidou, P. Labropoulou, E. Desipri, S. Piperidis, H. Papageorgiou, M. Monachini, F. Frontini, T. Declerck, G. Francopoulo, V. Arranz, et al. The META-SHARE metadata schema for the description of language resources. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1090–1097, 2012.

[9] J. Kunze and T. Baker. The Dublin Core metadata element set. RFC 5013, Internet Engineering Task Force, 1997.

[10] F. Maali, J. Erickson, and P. Archer. Data catalog vocabulary (DCAT). W3C recommendation, The World Wide Web Consortium, 2014.

[11] J. P. McCrae, P. Cimiano, V. Rodriǵuez Doncel, D. Vila-Suero, J. Gracia, L. Matteis, R. Navigli, A. Abele, G. Vulcu, and P. Buitelaar. Reconciling heterogeneous descriptions of language resources. In *Proceedings of the 4th Workshop on Linked Data in Linguisitcs*, 2015.

[12] J. P. McCrae, P. Labrapoulou, J. Gracia, M. Villegas, V. R. Doncel, and P. Cimiano. One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web. In *Proceedings of the 4th Workshop on the Multilingual Semantic Web*, 2015.

[13] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244, 2014.

[14] S. Piperidis. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 36–42, 2012.

[15] A. Seaborne, K. G. Clark, L. Feigenbaum, and E. Torres. SPARQL 1.1 query results JSON format. W3C recommendation, The World Wide Web Consortium, 2013.

[16] H. Tohyama, S. Kozawa, K. Uchimoto, S. Matsubara, and H. Isahara. Shachi: A large scale metadata database of language resources. In *Proceedings of the 1st International Conference on Global Interoperability for Language resources*, pages 205–212, 2008.

[17] D. Van Uytvanck, H. Stehouwer, and L. Lampen. Semantic metadata mapping in practice: the virtual language observatory. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1029–1034, 2012.