A Road Map for Interoperable Language Resource Metadata

Christopher Cieri[1], Khalid Choukri[2], Nicoletta Calzolari[3], D. Terence Langendoen[4], Johannes Leveling[5], Martha Palmer[6], Nancy Ide[7], James Pustejovsky[8]

1. Linguistic Data Consortium (LDC), (ccieri @ ldc.upenn.edu,)
2. European Language resources Association (ELRA), (choukri @ elda.org)
3. Instituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche (glottolo @ ilc.cnr.it)
4. Department of Linguistics, University of Arizona (langendt @ email.arizona.edu)
5. Centre for Next Generation Localisation (CNGL), Dublin City University, (johannes.leveling @computing.dcu.ie)
6. Center for Computational Language and Education Research, Department of Computer Science, University of Colorado, Boulder (Martha.Palmer @ Colorado.edu)
7. Department of Computer Science, Vassar College, USA,(ide @ cs.vassar.edu)
8. Department of Computer Science, Brandeis University, (jamesp @ cs.brandeis.edu)

Abstract

LRs remain expensive to create and thus rare relative to demand across languages and technology types. The accidental re-creation of an LR that already exists is a nearly unforgiveable waste of scarce resources that is unfortunately not so easy to avoid. The number of catalogs the HLT researcher must search, with their different formats, make it possible to overlook an existing resource. This paper sketches the sources of this problem and outlines a proposal to rectify along with a new vision of LR cataloging that will to facilitates the documentation and exploitation of a much wider range of LRs than previously considered.

1. Introduction

Despite years of language resource (LR) creation projects aimed at human language technology (HLT) research and development (R&D), LRs remain expensive to create and thus rare relative to demand across languages and technology types. In this context, the accidental re-creation of an LR that already exists is a nearly unforgiveable waste of scarce resources. Despite the existence of a few large data centers focused on HLT, the European Language Resource Association (ELRA 2010) and the Linguistic Data Consortium (LDC 2010), a prior harmonization project, Networking Data Centers (ELRA 2000) and a union catalog initiative, the Open Language Archives Community (OLAC 2010), HLT researchers must still master multiple metadata sets in order to search multiple locations in order to find needed resources or else risk failing to note the existence of critical LRs and then either recreate them or else do without them.

To address these issues, two cooperating projects: the European FlareNet and the US SILT organized a dedicated workshop with a specific working group that proposed a sequence of short and medium term activities that will result in a transformation of the landscape for HLT LR metadata. Before proceeding to describe this proposal, we describe the current landscape with focus on a series of independent efforts that we propose to coordinate. Currently major data centers (e.g. ELRA, LDC) maintain their own separate catalogs using different metadata languages (categories, terminologies) and export subsets of their metadata categories to the OLAC central data repository. OLAC provides specifications for OAI (Open Archives Initiative) compliant metadata as well as routines for harvesting, interchanging and searching their metadata. ELRA's universal catalog seeks to extend the work of OLAC while focusing on resources intended for HLT R&D. Specifically, the ELRA UC includes a greater percentage of ELRA metadata fields and exploits data mining to discover resources not produced or distributed by ELRA. The LREC Map is an initiative to exploit the biennial LREC (ELRA 2010) abstract submission process as a way to increase the contribution of LR metadata. Each author is asked to complete a simple template answering questions about the LRs described within the proposed paper. The NICT Shachi catalog (NICT 2009) also attempts to serve as a union catalog of resources including resource produced by NICT and elsewhere. Shachi currently differs from OLAC in that catalog records are scraped rather than harvested as part of a bilateral negotiation. Shachi also uses data mining technologies to discover information about LRs that may not be present in their home catalog entries. The LDC LR Wiki (LDC 2010) indentifies LRs for less commonly taught languages organized by language and LR type with individual sections edited by area experts. Some resources, especially plain and parallel text and lexicons,

are identified and even harvested via automated scraping activities. Believing that the community currently lacks the knowledge required to normalize the metadata for LCTLs, the wiki permits free text description and intends to attempt normalization as an activity under the current proposal. Finally the LDC LR Papers Catalog enumerates research papers that introduce, describe, discuss, extend or rely upon another LR. Currently, LDC focuses on papers dealing with LDC data resources and includes full bibliographic information on the paper plus a link to the unique identifier of an LDC data resource. Table 1 summarizes the nature of these organizations and their efforts.

	OLAC	ELRA UC	LREC Map	NICT Shachi	LDC LR Wiki	LDC Papers Catalog
external resources	✓	✓	✓	✓	✓	✓
normalized metadata	✓	✓	✓		deferred	✓
raw resources					✓	
scraping				✓	✓	
data mining		✓		✓		
papers as LRs	✓		✓			✓

Table 1: Features of Current Cataloging Efforts

In order to remove the barriers to progress in the catalog described above, we must assure that all are interoperable. For our purposes, interoperability of metadata languages *L1* and *L2* describes the capability of two metadata providers to interchange metadata records *m1* written in *L1* and *m2* written in *L2* for a single LR *r* via a function *f* that maps *L1* to *L2* such that a query that returns *r* in *L2* also returns *r* in *f(L1)*. Less formally a single search should work equally well, retrieving the same LRs, when issued against different but interoperable catalogs assuming the same base metadata in the two catalogs perhaps mapped to a new form in one of the catalogs. The project described below instantiates this definition and applies it to a number of mission critical LR catalogs.

2. Short Term Recommendations

In the short term, we propose a simple and concrete initiative, the harmonization of the LR catalogs of the largest international data centers including initially those of ELRA and LDC as well as the Shachi union catalog produced by NICT. The approach to this harmonization will not be reductionist. That is, we will not attempt to identify the minimal subset of the metadata fields that apply to all LR types. Instead we will begin with focus on the domain of LRs targeted toward HLT R&D and identify the superset of metadata types contained in them. Next we will review the types, one by one, to identify those than can be normalized internally and across data centers and will distinguish those from the smaller subset that encode irreconcilable differences among the business practices of the centers as adapted to their local regulatory constraints. The data center partners will agree to normalize and harmonize practice wherever possible. The outcome of this part of the effort will be new fully functioning LR catalogs to replace those of the data centers partners as well as a definition of the metadata categories, a database structure to hold the metadata and a search engine customized to HLT LR search. These new resources along with a specification of best metadata practices will be made available to other data centers and individual data creators to use in the creation of their own catalogs. To promote the sustainability of LR held outside data centers, we will provide a centralize metadata repository with a harvesting protocol. To address the range of competences among LR searchers, the search engine will permit both use of controlled vocabulary fields and relevance-based search of entire catalog records. A novel contribution of this effort will be searcher assistance based upon the relations among metadata categories (dictionary lexicon) and prior search behavior (those who searched for “Gigaword” also searched for “news text corpora”). Coupled with searcher assistance, we will provide metadata creator assistance based on searcher behavior and behavior of other metadata providers (“93% of

searchers include a language name in their search" but "87% of all providers include ISO 639-3 language codes" and "the metadata you have provided so far also characterize 32 other resources"). In order to effectively manage the harmonization of data center catalogs and the provision of metadata resource, we will construct a governance body specifically for this project. The group will include representatives of the project partners, sponsors, individual and small group resource providers and of LR users.

3. Middle Term Recommendations

In the middle term, we propose to expand the scope of the universal catalog to include two important and frequently overlooked LRs on either end of the processing spectrum, raw unprocessed data and the most carefully processed LRs, research papers. Some of this work has already begun in a number of individual efforts that have not been coordinated across this same span of data centers and LR creators. Specifically, a Less Commonly Taught Language (LCTL) Language Resource wiki was developed by LDC within the REFLEX program. Similar efforts to harvest papers describing LRs are underway at LDC using human effort and within the Rexa project (Rexa 2010) using data mining technologies. Our proposal here is to accomplish this expansion by data type while integrating model workflow methodologies into the workflow including social networking, web sourcing, and data mining. In this project, our intent is to enhance the universal catalog with links to raw resources including web sites rich in monolingual and parallel text and lexicon built for interactive use. The resources are necessary to advance the universal catalog toward the very apt goal of true universality as it affects languages whose representation among formal LRs is insufficient with respect to their global importance. Those who would create HLTs for these languages must resort to primary LR resource creation based upon harvests of these raw resources. As the project moves from short to middle terms objectives it will be necessary to adjust its governance and broaden the scope of its normalization activities. The issues that challenge this expansion to raw resources and papers differed markedly from those that challenge the harmonization of traditional catalog metadata and the collaborators must change in response. The principles centers of these middle term activities will be large data providers that have already implemented sustainable business models.

4. Requirements

The collaborators needed to assure the success of this project include some of the members of the initial FlareNet/SILT planning group. New collaborators will include:

- Representatives of the relevant data and metadata centers; ELRA/ELDA, LDC, NICT, and OLAC.
- Representatives of interested professional organizations: ACL, LSA, LinguistList, SIL, ISCA
- Editors of journals who will agree to implement a version of the LREC map questionnaire including: LRE, LILT.
- Organizers of conferences who similarly agree to require the LREC Map questionnaire to be answered: LREC, AFLR
- Representative from related catalogin projects, for example the Rexa Project
- Representative of leading industrial partners possibly Microsoft, Google and partners in the CNGL
- Representatives of the LanguageGrid project, the Centre for Development of Advanced Computing (CDAC) in India and the Technology Development for Indian Languages project.

The resources needed to assure success include:

- funding support for Europe (for example FlareNet, T4ME)
- funding support in US (beginning with SILT but seeking additional support from NSF or other federal agencies with an interest in LR)
- funding support for NICT (current funding finished)
- database schema (existing or one we create)
- search engine
- technology for data mining
- taxonomy/controlled vocabulary of applications, data types, etc.

In addition to the activities described above we believe the following activities need to be highlighted

- outreach
- sample output early in the project
- endorsement of institutions
- evaluation of metadata (user feedback)
- evaluation of performance of the Catalog in terms of LRs required.

5. Cross Cutting Recommendations for Harmonization Projects

Given the central position of language resources in HLT research and development and the critical role of metadata in the distribution and exploitation of language resources, it is

important that the initiatives described herein are integrated within the broader range of SILT/FlareNet activities. We propose to define a use case that will stress not only the work of this group focused on metadata but also the other working groups focused on data categories, semantics of annotation and requirements for LR publication. The case we envision is an infrastructure that supports the automatic discovery and processing of resources needed to build an HLT. Consider the training of a parser from Treebanked data. Although metadata harmonized across data centers that fully describe these Treebanks is a necessary condition of this work it is not a sufficient condition. In order to automatically discover and process Treebanks it will also be necessary that the corpora themselves are described in a structured way that can be read by machines in order to identify the location and format of files and relevant annotations within them. To assure that two non-identical annotations are in fact compatible, it will also be necessary link the structured, machine-readable description of the corpus to a data category registry to assure that different annotators have adopted the same definitions of variable terms. Finally, the corpora will need to have been created and distributed according to best practices and standards including a thorough human readable description of the methodology. These four areas of need match the SILT/FlareNet working groups created during the November workshop at Brandeis. Adopting such a use case across SILT/FlareNet working groups will have the positive effect of focusing the groups' efforts toward a common goal.

6. Conclusion

We have described impediments to progress in the current landscape of language resource metadata and sketched a proposal for addressing these. The proposed work would begin by removing the largest impediments that affect users of the two largest international data centers. Outcomes of this work would include infrastructure for mapping the LDC and ELRA catalog entries to one another, a unified catalog and infrastructure for building new compatible, instantiations of LR catalogs for example, among smaller data centers. Middle term work would expand the notion of LR catalog in two directions to include both raw resources needed for under-resourced languages and papers describing or relying upon LRs. The implementation of this new concept of LR catalog would include raw resource scouting and harvesting tools and data mining tools for filling in the missing parts of existing metadata records. We have identified the resources needed to accomplish the short and middle term goals. Finally we have proposed a use case that would unite the activities of the SILT/FlareNet working groups toward a common goal. Coordinated efforts on this scale have the potential to significantly alter the landscape of language resources, reducing impediments to progress for current and new researchers, and for well and poorly resourced languages.

7. Acknowledgements

This work was made possible through the U.S. National Science Foundation's INTEROP Program's support of the "Sustainable Interoperability for Language Technology" (SILT) project and the European Commission's eContentPlus Framework's support of the Fostering Language Resources Network (FLaReNet,www.flarenet.eu).

8. References

ELDA (2000), NetDC Arabic Broadcast News Speech Corpus, http://www.elda.org/article45.html

ELRA (2010) European Language Resource Association Home Page, www.elra.info.

ELRA (2010) Language Resource and Evaluation Conference 2010 Home Page, www.lrec-conf.org

LDC (2010) Linguistic Data Consortium Home Page, www.ldc.upenn.edu.

LDC (2010) Language Resource Wiki lrwiki.ldc.upenn.edu.

NICT (2009) Shachi Catalog www2.shachi.org

OLAC (2010) Open Language archives Community Home Page www.language-archives.org

Rexa (2010) Rexa Project Home Page, rexa.info