

# **Towards a Web Search Service for Minority Language Communities**

*Baden Hughes*

**Department of Computer Science and Software Engineering  
The University of Melbourne  
VIC 3010, Australia  
badenh@csse.unimelb.edu.au**

## **1. Introduction**

Locating resources of interest on the web in the general case is at best a low precision activity owing to the large number of pages on the web (for example, Google covers more than 8 billion web pages). As language communities (at all points on the spectrum) increasingly self-publish materials on the web, so interested users are beginning to search for them in the same way that they search for general internet resources, using broad coverage search engines with typically simple queries. Given that language resources are in a minority case on the web in general, finding relevant materials for low density or lesser used languages on the web is in general an increasingly inefficient exercise even for experienced searchers. Furthermore, the inconsistent coverage of web content between search engines serves to complicate matters even more.

A number of previous research efforts have focused on using web data to create language corpora, mine linguistic data, building language ontologies, create thesaurii etc. The work reported in this paper contrasts with previous research in that it is not specifically oriented towards creation of language resources from web data directly, but rather, *increasing the likelihood that end users searching for resources in minority languages will actually find useful results* from web searches. Similarly, it differs from earlier work by virtue of its focus on search optimization directly, rather than as a component of a larger process (other researchers use the seed URIs discovered via the mechanism described in this paper in their own varied work). The work here can be seen to contribute to a user-centric agenda for locating language resources for lesser-used languages on the web.

Building on earlier work in the development of effective and robust strategies for finding web resources for lesser used languages (Hughes, 2005) we describe the architecture and main functions of new web search service for minority language communities, allowing end users to efficiently discover web based resources in their language of choice. The web search service consists of three components: the LangGator web crawler (Hughes 2005) which exclusively targets language resources; the Open Language Archives Community (OLAC) metadata set (Bird and Simons, 2003) for the description of these found resources; and the Open Language Archives Community Search Engine (Hughes

and Kamat, 2005), by which users can execute searches. This web search service is grounded to the Australian context through support for more than 3000 languages world wide, including language communities which are present in Australia, both as officially recognized languages and languages used by community groups of various origins. Furthermore, a gateway service gateway service allows generic web search engines such as Google and Yahoo to find these specialized collections of language specific material.

## **2. Australia's Linguistic Diversity**

It is well recognized that Australia is a nation with considerable linguistic diversity, derived from its increasingly rich multicultural heritage. In terms of global statistics, according to the Ethnologue (14<sup>th</sup> Edition, 2000) there were 311 distinct languages spoken in Australia (compared to 7299 world wide). Similarly according to the Ethnologue (15<sup>th</sup> Edition, 2005) there were 318 distinct languages spoken in Australia (of 7299 world wide). A simple statistical comparison to the linguistic situation in all other countries shows that Australia is within the top 10 countries for linguistic diversity (determined by number of languages spoken in a country as a percentage of all languages globally). Locally, the Australian Bureau of Statistics reports that there are 364 languages spoken in Australia (Australian Standard Classification of Languages 2005a.)

Assuming that the multilingual communities represented in Australia are typical of those elsewhere, there is a strong motivation for enabling these communities to discover web based resources of interest to their linguistic and cultural heritage, and indeed to contribute these for other users. Hence we are motivated strongly by a local, as well as international use case for facilitating discovery of minority language resources on the web.

## **3. The Open Language Archives Community**

The Open Language Archives Community (OLAC) is a consortium of linguistic data archives, at the time of writing consisting of 34 archives and a corresponding catalogue of 28,000 objects described by metadata. (For a more detailed description of OLAC, we refer interested readers to Bird and Simons (2003) and Simons and Bird (2003a)). OLAC metadata is based on Dublin Core, with a number of extensions (Simons and Bird, 2002) to the Dublin Core Metadata Set (Dublin Core, n.d.) for relevant conceptual domains such as language (Simons and Bird, 2003c), subject language (Simons and Bird, 2003b), linguistic type, linguistic subject and linguistic role.

Derived from the model adopted within the OAI, the OLAC model has a two tiered approach to implementation. Data providers are the institutional language archives which publish their XML-based metadata according to the OAI Static Repository standard (Hochstenbach et al 2003). Individual archives use a variety of software to manage their catalogues internally. Service providers leverage the OAI Protocol for Metadata Harvesting (Lagoze et al 2002) to harvest the XML expressions of metadata catalogues. Within the OLAC community, typical practice is to aggregate these into an SQL database

using the OLAC Harvester and Aggregator. Service providers can then build services which utilize the union catalogue of OLAC metadata.

Of particular interest in this context is the fact that OLAC metadata distinguishes the language a resource is in ('language') as opposed to the language a resource is about ('subject language'). In the particular context of web search for minority language communities, we believe that resources in a given language (hence, the 'language' referent) are likely to be more relevant to users than resources which are about a given language ('subject language').

## **4. The LangGator Service Architecture**

Building on earlier work in the development of effective and robust strategies for finding web resources for lesser used languages (Hughes, 2005) the main focus of this paper is to describe a new web search service for minority language communities, allowing end users to efficiently discover web based resources in their language of choice. The web search service consists of three components: the LangGator web crawler (Hughes 2005) which exclusively targets language resources; the Open Language Archives Community (OLAC) metadata set (Bird and Simons, 2003) for the description of these found resources; and the Open Language Archives Community Search Engine (Hughes and Kamat, 2005), by which users can execute searches. We proceed by discussing each of these components in turn.

### **4.1 Crawler**

The crawler component of the LangGator is responsible for the content identification and acquisition tasks.

In the first instance, LangGator proceeds by gathering a list of seed URIs for retrieval based on a principled set of queries derived from the Ethnologue list of language names and variants, the Getty Thesaurus list of country names and variants, and lexical items from the Rosetta Project. These queries are then programmatically dispatched to several search engines including Google, Yahoo, A9 and DogPile.

The query results from the search engines are then combined using rank aggregation techniques relevant to the number and diversity of links (the interested reader is referred to Meng, Yu and Liu, 2002 for the techniques used). The aggregated results then form a priority list for retrieval of the web documents, and their subsequent scope crawling.

Once a specific document is retrieved, we conduct focused crawling around the document instance, searching and retrieving both inbound and outbound links for a candidate resource to find potentially additionally relevant content. We generate and compare a term frequency / inverse document frequency (TF/IDF) for low frequency items in candidate document against the additional located documents from link analysis, and if a close TF/IDF is found, then the additional document is included in the collection (it should be noted that it can be excluded based on language identification performance in the next phase).

At the time of writing LangGator has identified and catalogued (ie created metadata for) more than 1.6 million ‘language-centric’ resources on the web in over 3000 languages. Not all of these are directly accessible through the OLAC infrastructure, although the number of metadata records published this way is incrementally increasing as quality control measures are scaled up to ensure high quality resources are published first and that copyright requirements are observed. The majority of these catalogued resources are in fact exposed to general purpose search engines directly via a DP9 gateway service (see discussion in the ‘Search Facilities’ section below).

For a more detailed description of the resource discovery strategy implemented in the LangGator software, the interested reader is referred to Hughes (2005).

## **4.2 Metadata Descriptions**

A critical part of any resource discovery framework is the ability to describe a resource separately from its realization. OLAC metadata provides a mechanism for resource description with a particular focus on linguistic properties, of which the language property as discussed earlier is of most interest here. The more general problem of automatic creation of metadata for electronic resources has been addressed elsewhere within the digital libraries community (most recently, see Paynter (2005)).

The most specialized instance of metadata creation in the context of web search for minority language resources is the language identification task itself, since the language of a resource is of primary importance to users searching for minority language resources on the web. To determine the language a specific resource is written in, we use a combination of several machine learning techniques which compare and classify the language from a candidate resource against reference samples drawn from hand-curated collections of language data sourced from the Rosetta Project (covering approximately 2400 languages). In particular, our data points are encoding, word n-grams, and character n-grams. The machine learning framework is implemented using a Java based machine learning toolkit, Weka.

It has been shown elsewhere that there is a strong positive correlation between minority language documents and the locations mentioned within the documents; the interested reader is referred to MacKinlay (2005) for further discussion on this topic. Therefore, we also extract geographic named entities from the candidate text using a generalized named entity tagger, and resolve these to a gazetteer (the Getty Thesaurus of Geographic Names).

Having conducted the extraction and classification phase, we then ascribe a probability to the language classification: probabilities of greater than 0.8 result directly in a language element entry based on the Ethnologue / ISO 639-3 coding system which is entered as OLAC metadata for the language resource.

For a more detailed description of the automatic metadata creation within OLAC, the interested reader is referred to Hughes (2006 to appear).

Having created metadata for the resources in question, the metadata records are selectively aggregated into a single OLAC data provider, and then harvested by the central OLAC infrastructure, from which searching is supported.

### **4.3 Search Facilities**

The specific search component is delivered within the standard search infrastructure of OLAC. The search engine resembles a typical form based web search engine, and acts similarly, leveraging many users familiarity with this paradigm.

The user interface at the search entry point is deliberately simple: a single text input field which supports full UTF-8 encoding in the string. Similarly to general purpose web search there is no presupposition as to what the input keywords may be, although as will be seen below, certain heuristics are brought to bear to determine linguistic properties of interest. Input is treated as case insensitive, and supports standard search operators such as AND, OR, etc. A variety of additional inline syntax is supported.

The search engine logic first uses exact word matching against a full text index of the OLAC metadata element content, soundex values for approximate string matching, and Ethnologue data providing information on alternate language names. The default behavior of the search is to match exact words in the search string and to return records that contain all words, using a full text index on the content of metadata elements. If exact matching is not successful, a range of heuristics is applied to determine the closest relevant matches and weight these for result ranking.

When a search by default yields no results, our aim is to minimize the number of mouse clicks a user requires to find information on either correct spellings of the search term, or information on topics related to the original search. In the event that there are no matching records found in the database, the query string is first checked to determine whether it is a language name. If it is the case that there are no records about an input language name, it is assumed to be more helpful to present alternate names for the language, rather than similarly spelled words. (This approach is grounded in the context of the complexity of language name standardization, a well-known problem in the domain).

Where this is not the case, similarly spelled words (including languages) are suggested. These words are retrieved from a table of words and their soundex values, the vocabulary consisting of all distinct words found in the content of all OLAC metadata elements. The table is indexed on the soundex value, to provide faster lookup. The words with soundex equal to that of the search term are retrieved and sorted according to their Levenshtein edit distance from the search term. Similarly, language names from the Ethnologue tables that have identical soundex values are ranked by Levenshtein distance and displayed as possible corrections to the original query string. Again, this separation of similarly

spelled words and similarly spelled language names makes an assumption that language names are most frequently searched.

The search engine also features an integrated domain thesaurus and ontology, derived from the authoritative source of language name classification, the Ethnologue. If the search string is identified as derivative of a language or country name, this is detected (by use of the Ethnologue data on alternate language and dialect names) and the user is presented with links to launch queries related to that language or its country of origin. Where the original search is for a language name, a link is available to view records that use an alternate name for that language or records that provide information on a dialect of the originally searched language. Similarly, if the original search string is a country name, a list of languages that are spoken in that country can be viewed. Recognizing language and country names allows a link to the Ethnologue entry on that search term to be presented where relevant. Furthermore, the use of the domain thesaurus and ontology also allows users with linguistic domain knowledge to search by language code, which is a commonly used approach in other language resource repositories.

Search results are grouped by archive, with the archives sorted based upon the aggregate of metadata quality scores for the retrieved records. Each retrieved record displays the element containing the matching word with key-word-in-context (KWIC) highlighting. Elements that have been deemed useful in providing information of record content (title, description, subject, date, identifier) are used to provide additional summary information.

A number of additional utilities are provided to the user in the context of search results. These include links to instantiate pre-composed queries for alternate names; language-code based searching; links to the Ethnologue; and pre-composed combinatorial web queries for language names and linguistic domain-specific terms.

The entire search engine interface being designed for localization. The mechanism used is essentially language (and encoding) preference selection, which in turn drives the selection of an appropriate XSL stylesheet for displaying the search results. Beyond this, a user can select an alternative language (and encoding) display after results are returned. At the time of writing we are completing a comprehensive code audit which will result in a set of strings for translation for efficient implementation of the localized interface: these will be made available directly to interested parties. Already we have interest for localization into French, Spanish, Bahasa Indonesia, Vietnamese and Thai.

For a more detailed description of the architecture of the search service, the interested reader is referred to Hughes and Kamat (2005). A user-centric review is being undertaken to determine the types of resources and searches that are specifically of interest to users. The interested reader is referred to Hughes (2006 to appear) for more details of user search behavior and preferences.

In addition to directly facilitated search, the LangGator collections are exposed via the Open Archives Initiative Protocol for Metadata Harvesting to any OAI compliant search engine, and via a DP9 gateway as static resource collections to traditional search engines

such as Google and Yahoo. This flexibility allows for additional modes of interaction with the LangGator collections outside their default discovery framework of the Open Language Archives Community.

## 5. Future Work

While a significant amount of functionality is currently deployed, there are a number of areas for improvement; here we list a number to provoke further interaction.

LangGator currently harvests a complete collection data set once every 3 months, largely owing to computational constraints. As such, resources which appear in the OLAC Search context may be up to 3 months old. It is clearly desirable to have a smaller temporal extent between harvests and this is currently being engineered by parallelizing the crawler phase for execution in on high performance computing infrastructure. The completion of this task will mean a considerably lower resource age, the target is that a resource found via the search service will have been verified minimally within the last 7 days.

The need for more, and more accessible end user documentation for specific search construction is quite obvious. The API for the OLAC search engine allows for very flexible interaction by third party services, and yet is not easily deciphered. For the end user the ability to create queries based on example (or queries derived from earlier searches) and the ability to save searches etc would certainly add to the user experience in working with the LangGator/OLAC search interfaces.

A natural extension of the search interface, particularly for searching by geographical information is the provision of a graphical user interface for locational-context aware searching. We have recently acquired the relevant mapping data on a global scale, and the corresponding language to location relational data to allow for this to be implemented, allowing users to interact with data in a different mode. Another extension in the area of location-based services which is under development is the ability to locate resources by geographical location or proximity. The existing service is quite limited in the sense that the user can find resources ‘by country’ rather than being able to find resources by finer grained locations eg regions, cities, or natural features. Additionally, the simple ‘country of origin’ classification for languages does not adequately address the situation where speaker dispersion has led to vibrant immigrant communities in other locations (‘country of use’); resources from which may be more appropriate to a users information needs.

It is possible that with extremely low density languages that LangGator will not find resources of interest to users. Given this possibility, we are considering how best to implement a ‘related language’ discovery service which may address this need. While we recognize that related language resources are not necessarily ideal, many minority language speakers are multilingual, and using linguistic relation information we may be able to provide a service to allow discovery of these resource types.

A related area of activity is the derivation of language resources for minority language communities based on the collections of data that are curated by LangGator. Many minority language communities lack basic electronic resources such as dictionaries, spell checkers, keyboarding support (what many distribution agencies call the “Basic Language Resource Kit” or BLARK): we are currently exploring the exploitation of free online resources to create some of these objects (see Baldwin and Hughes, 2006 for examples).

Finally, it is hoped that a synergistic relationship with the newly established MyLanguage portal will be established specifically to allow Australian-resident minority language speakers to more efficiently discover web resources of interest. The MyLanguage effort effectively engages with minority language communities in Australia in a way that is difficult to replicate. The LangGator service and the Open Language Archives Community provide a scalable resource discovery framework on an international scale which is likewise inefficiently duplicated. As an initial step toward collaboration, we are implementing a simple harvester for the content of the MyLanguage portal, which will allow language resources curated within that context to be found from the generic OLAC search.

## 6. Conclusion

While to date the Open Language Archives Community has largely focused on the needs of the documentary linguistic research, the LangGator service, and the future development activity identified will make this rich resource more accessible to those interested in minority language content on the web, both via specialized and general purpose search engines. While English remains the dominant language of content and users on the web, multilingual content and communities are flourishing online, and services which address their specific needs are required.

## References

- Australian Bureau of Statistics, 1997. *Australian Standard Classification of Languages*. (First Edition; 1267.0). Australian Bureau of Statistics, Canberra.
- Australian Bureau of Statistics, 2005a. *Australian Standard Classification of Languages*. (Second Edition; 1267.0-2005-06). Australian Bureau of Statistics, Canberra.
- Dublin Core, n.d. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. <http://dublincore.org/documents/dces>
- Timothy Baldwin and Baden Hughes, 2006 to appear. *Deep Lexical Exploitation of Language Resources on the Web*. To appear in Proceedings of the ESSL 2006 Workshop on Resource Scarce Language Engineering.
- Steven Bird and Gary Simons, 2003. *Extending Dublin Core Metadata to support the description and discovery of language resources*. *Computing and the Humanities* 37, pp.375-388.
- Baden Hughes, 2005. *Towards Effective and Robust Strategies for Finding Web Resources for Lesser Used Languages*. Proceedings of Lesser Used Languages and Computational Linguistics. EURAC, Bolzano.

Baden Hughes, Towards a Web Search Service for Minority Language Communities, Open Road 2006

Baden Hughes, 2006a to appear. *Searching for Language Resources on the Web: User Behavior in the Open Language Archives Community*. To appear in Proceedings of the 5<sup>th</sup> International Conference on Language Resources and Evaluation. European Language Resources Association, Paris.

Baden Hughes, 2006b to appear. *Automatic Creation of OLAC Metadata for Web-based Language Resources*. To appear in Proceedings of the 5<sup>th</sup> International Conference on Language Resources and Evaluation. European Language Resources Association, Paris.

Baden Hughes and Amol Kamat, 2005. *A Metadata Search Service for Digital Language Archives*. D-Lib Magazine 11(2).

Raymond G. Gordon Jr. (ed.), 2005. *Ethnologue: Languages of the World* (15<sup>th</sup> Edition). SIL International, Dallas.

Barbara F. Grimes (ed.), 2000. *Ethnologue: Languages of the World* (14th Edition). SIL International, Dallas.

Patrick Hochstenbach, Henry Jerez, Herbert Van de Sompel, 2003. *The OAI-PMH Static Repository and Static Repository Gateway*. Proceedings of the IEEE/ACM Joint Conference on Digital Libraries 2003 (JDCL'03). pp. 210-220.

Carl Lagoze, Herbert Van de Sompel, Michael Nelson and Simeon Warner, 2002. *The Open Archives Initiative Protocol for Metadata Harvesting*. <http://www.openarchives.org/OAI/openarchivesprotocol.html>

Andrew MacKinlay, 2005. *Using Diverse Information Sources to Retrieve Samples of Low Density Languages*. Proceedings of the Australasian Language Technology Workshop 2005. Australasian Language Technology Association, Sydney. pp. 64-70.

Weiyi Meng, Clement Yu and King-Lup Liu, 2002. *Building Efficient and Effective Metasearch Engines*. ACM Computing Surveys 34(1), March 2002. pp.48-69.

Gordon W. Paynter, 2005. *Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources*. Proceedings of the Fifth ACM/IEEE Joint Conference on Digital Libraries, Denver, Colorado, June 7-11, 2005, ACM Press, pp. 291-300.

Gary Simons and Steven Bird, 2002. *Recommended Metadata Extensions*. <http://www.language-archives.org/REC/olac-extensions.html>

Gary Simons and Steven Bird, 2003a. *The Open Language Archives Community: An infrastructure for distributed archiving of language resources*. Literary and Linguistic Computing 18. pp.117-128.

Gary Simons and Steven Bird, 2003b. *OLAC Subject Language Vocabulary*. <http://www.language-archives.org/REC/language.html>

Gary Simons and Steven Bird, 2003c. *OLAC Language Vocabulary*. <http://www.languagearchives.org/REC/language.html>

## **Acknowledgements**

Portions of the research in this paper has been supported by the Australian Research Council under the funding program for Special Research Initiatives in E-Research, grant SR0567353 “An Intelligent Search Infrastructure for Language Resources on the Web”.