Language engineering for the Semantic Web: a digital library for endangered languages.

Shiyong Lu, Dapeng Liu, Farshad Fotouhi, Ming Dong , Robert Reynolds
Department of Computer Science

Anthony Aristar, Martha Ratliff, Geoff Nathan Department of English/Linguistics

Joseph Tan, Department of Information Systems Manufacturing

Ronald Powell, Department of Library and Information Science
Wayne State University, Detroit, MI 48202

change font

Abstract

Many languages are in serious danger of being lost and if nothing is done to prevent it, half of the world's approximately 6,500 languages will disappear in the next 100 years. Language data are central to the research of a large social science community, including linguists, anthropologists, archeologists, historians, sociologists, and political scientists interested in the culture of indigenous people. The death of a language entails the loss of a community's traditional culture, for the language is a unique vehicle for its traditions and culture. In this paper, we describe the effort undertaken at Wayne State University to preserve endangered languages using the state-of-the-art information technologies. We discuss the issues involved in such an effort, and present the architecture of a distributed digital library which will contain various data of endangered languages in the forms of text, image, video and audio files and include advanced tools for intelligent cataloguing, indexing, searching and browsing information on languages and language analysis. Various Semantic Web technologies such as XML, OLAC, and ontologies are used so that the digital library is developed as a useful linguistic resource on the Semantic Web.

Introduction

Recently, there is an increasing awareness of the fact that many languages are in serious danger of being lost. As Professor Peter Austin in the School of Oriental and African Studies at the University of London commented in an interview with the BBC World News, half of the world's approximately 6,500 languages will disappear in the next 100 years unless we act immediately. In this interview, Professor Steven Bird of Melbourne University said, "We are sitting between the onset of the digital era and the mass extinction of the world's languages. The window of opportunity is small and shutting fast".

Language data are central to the interests of a large social science research community, including linguists, anthropologists, archeologists, historians, sociologists, and political scientists interested in the culture of indigenous people. When a language disappears there are two major effects. First, there is the loss of the valuable data of the cultural system that produced the language. The death of a language usually entails the loss of a community's traditional poetry, songs, images, stories, proverbs, laments, and religious rites. Secondly, any language loss represents a serious scientific loss: studies of linguistic diversity and cross-linguistic comparisons drive much of linguistic theory. In addition, linguistic material provides valuable information about population movements, contacts, and genetic relationships. When a language becomes highly endangered, efforts to preserve the existing documentation about the language become critically important, not only to that community, but also to academic linguistics and sister sciences such as anthropology, archaeology, history, and ethnobiology.

Digital technologies offer the best promise for the preservation of endangered languages data, for they give permanent storage, wide dissemination, and flexible access. However, to realize that promise, it is important to digitize the material in a way that conforms to the best practice within the language technology community, for otherwise it will not be generally accessible, or machine-interpretable. Because of the urgent need to preserve irreplaceable linguistic material, many linguists and librarians have embarked on the wholesale creation of digital recordings, database records, and Internet displays of texts and lexicons. These reside in disparate and often unrelated sites on the Internet. But the resulting multiplicity of standards and formats has produced a situation in which much of this digitized information is less accessible and less portable than the printed material it was intended to replace. Also, another major problem with searching a distributed set of archives is the incompatibility of markup of data, and the incompatibility of the queries each system requires. One of the most obvious ways that this problem manifests itself is in multilingual searches: for example, a query might require the use of English, or French, or Spanish terms. One simple approach is to limit query terms to English, and then to search translated terms in collections of other languages based on dictionaries which translate from English to other languages. However, this is not a semantic approach, since what we require is not just mapping between words, but between senses of a word. To take an example, the word ***morphology*** has many meanings in English, most of which are irrelevant to linguistic data. Its translation in French, ***morphologie***, has a somewhat different set of senses. How do we know whether a use of ***morphologie*** is relevant to our search? The solution is to use an ontology which precisely defines meaning in an area of knowledge, and then to map the terminology of other languages to that ontology.

In this paper, we describe the effort undertaken at Wayne State University to preserve endangered languages using the state-of-the-art information technologies. In particular, we discuss the issues involved in such an effort, and present the architecture of a distributed digital library containing various data of endangered languages in the forms of text, image, video and audio, and include advanced tools for intelligent cataloguing, indexing, searching and browsing information on languages and language analysis. We use various Semantic Web technologies such as XML, OLAC, and ontologies so that our digital library is developed as a useful linguistic resource on the Semantic Web. The initial 2-year stage of this multidisciplinary project has been funded by the University Research Enhancement Program at Wayne State University. We expect to secure additional funding from NSF and other agencies to complete the project.

Related work

According to Krauss (Krauss, 1992), around half of the estimated 6,500 languages on earth are spoken only by adults. This is a clear sign of the death of a language: when a minority culture no longer passes on its language to its children, the language will usually survive for only fifty or sixty years at the most.

The decline of linguistic diversity exists on a global scale: in Europe Basque is receding in Spain and France, and Sami in Scandinavia, to name but two examples. The situation is even more dire in the rest of the world: majority languages such as English, Spanish, Chinese, Russian and Arabic are spreading at the expense of a myriad of smaller languages in the Americas, Africa, Australia, and Southeast Asia (Robins & Uhlenbeck, 1991; Brenzinger, 1992; Schmidt, 1990).

The LINGUIST List, whose data derive in large part from the Ethnologue, the most comprehensive database of current languages in existence, lists 7420 languages in its database, of which 543 are already extinct, and a further 432 which have so few speakers left that they will no longer be spoken within 20 years. Dixon (1997) gives us some idea about why languages have come to be endangered and why such a loss is a serious problem. The rate of extinction is clearly accelerating: with the spread of modern communications, the major world languages are becoming more and more essential and pervasive as vehicles for communication.

For a comprehensive list of languages in danger and a concise summary of the worldwide language endangerment situation, the reader is referred to Wurm's "*Atlas of the World's Languages in Danger of Disappearing*" (2001) sponsored by the UNESCO/Japan Trust Fund for the Preservation of the Intangible Cultural Heritage.

The existence of facts like these has recently stimulated a debate on whether endangered languages are worthy to be preserved at all: Sutherland (2003) examined the threat to the world's 6,500 languages and concluded: *"The threats to birds and mammals are well known, but it turns out that languages are far more threatened."* Against this, Prof. David Berreby argued in the science section of the May 27, 2003 edition of the New York Times (Berreby, 2003) that, the extinction of animals is different from that of endangered languages: the former, once gone, are gone forever; the later can be revived later when necessary. He argued that,

> It would be a terrible thing to run out of languages. But there is no danger of that, because the reserve of language, unlike the gas tank, is refueled every day, as ordinary people engage in the creative and ingenious act of talking. Old words, constructions and pronunciations drop away, new ones are taken up, and, relentlessly, the language changes

Prof. Berreby is of course perfectly correct in the assertion that languages recreate themselves. But he forgets two important facts: languages created anew do not also recreate the cultural material that the lost languages conveyed: once lost cultural material is lost forever. And since these new languages will be based upon languages already in existence, they do not show the range of linguistic behaviour which the lost languages displayed. They offer little help, then, in the development of more encompassing linguistic theories, and thus little help in the development of a more comprehensive understanding of the way the human mind can communicate.

Efforts of preserving endangered languages have thus been thus proceeding. To promote the international recognition of the importance of preserving the world's linguistic diversity, a symposium on *Language Loss & Public Policy* was held at the University of New Mexico at Albuquerque in June of 1995, which resulted in the establishment of an organization named Terralingua. Many other organizations, funding foundations, and projects have initiated projects and initiatives world-wide since then to promote, support, and conduct the preservation of endangered languages, including:

- The LINGUIST list,
- E-Meld (Electronic Metastructure for Endangered Languages Data) ,
- Foundation for Endangered Languages,
- The Endangered Language Fund,
- The International Clearing House for Endangered Languages,
- Endangered Language Repository,
- Indigenous Language Institute,
- Endangered Languages of the Pacific Rim,
- Teaching Indigenous Languages,
- UNESCO Red Book on Endangered Languages: Europe, and
- MIT Indigenous Language Intiative.

A complete list is beyond the scope of this paper. The reader is referred to the Yahoo! Directory on Endangered Languages for similar activities.

This dire situation has stimulated extensive linguistic research on endangered languages and how to preserve them. Bobaljik (1996), Grenoble (1998), Robins (1991), and Matsumura (1998) are collections of excellent papers presenting case studies, strategies, techniques, tools, and methods of preserving endangered languages. A detailed review of the literature on endangered languages is beyond the scope of this paper. The reader is referred to the Bibliography of Language Revitalization and Endangered Languages.

Architecture of the digital library

In the past few years, there has been a significant increase in language documentation projects, and also in attempts to preserve language documentation using digital technologies. But as mentioned above, this attempt at digital preservation and data sharing is threatened by the multiplicity of incompatible technologies currently in use. The lack of common standards and formats, as well as the lack of supporting software, currently impedes long-term storage, retrieval, display, and even comparative analysis of language data. To address this problem, the E-MELD project was proposed by Anthony Aristar, Helen Aristar-Dry and the LINGUIST list to the NSF in 2000, and funded the following year. The goal of the E-MELD project is to define the infrastructure of a distributed archive, and has focused on the following three main tasks:

- To build a showroom of best practice for digital archives of endangered language data, where the data from *ten* endangered languages is archived in a way such as to demonstrate the best practice, and the best way to design and store material for such an archive,
- To build a linguistic ontology which would serve as an interlingua for the various linguistic markups used, so as to allow searching of diverse material, and
- To build a tool (now called *FIELD*) which facilitates the work of linguists in storing endangered languages material to conform to best practice.

The digital library system we present here generalizes the work of E-MELD, and focuses on a system for the intelligent searching of a distributed endangered language archive across the Internet, with particular emphasis on audio and visual material. The system integrates various data sources and their existing ontologies for the existing endangered languages. E-MELD is only one such data source. Web data integration systems are designed for the purpose of answering user queries by integrating the most accurate and relevant information from a variety of Web data sources. We are currently eveloping a Web data integration system to retrieve data from Web sources and store the necessary data, metadata and language annotations into a virtual data warehouse. We are developing an XML-based global schema and its associated tools for storing, intelligent browsing and querying of endangered language data. The system is not simply another gateway or portal, but an integrated, intelligent, tool-rich research environment, incorporating various multilingual and visual domain-dependent ontologies. In summary, the objectives of this research are:

- To develop and implement a federated multimedia Web-based digital library that integrates various endangered languages data sources including E-MELD (see Figure 1);
- To design and develop an annotation management system to support domain experts to annotate various data related to endangered languages;
- To provide support for intelligent querying and retrieval of language data;
- To provide support for ontology-based multilingual, cross-theoretic querying and retrieval;
- To provide tools for analysis and cross-linguistic comparison;
- To provide methods for accurate display of linguistic documentations of various media formats for various display devices;
- To enhance FIELD to allow for storage of audio and visual materials.

Web data integration systems are designed for the purpose of answering user queries by integrating the most accurate and relevant information from a variety of Web sources. One approach to building a Web data integration system is to retrieve data from Web sources and load it into a data warehouse. This data warehousing approach is implemented by two well-known Web data integration systems: Information Manifold (Levy *etc*, 1996), a project of AT&T Laboratories, and Tsimmis (Garcia-Molina, 1997), a project at Stanford University. In this kind of architecture every data source is wrapped by software that maps the source-specific language and model to a generic data model shared by all sources. Mediators extract information from the shared model and make it available to data requestors. The advantage of this approach is the predictable performance at the time of a query; the drawback is that it requires warehouse data be updated every time the source data are changed.

Another approach to building a Web data integration system is the virtual approach where the data remains in the Web sources and queries are decomposed at run-time into queries on the sources (Florescu, 1998), which has been implemented in the Florid system. The virtual approach has become prevalent in Web data integration systems because it provides a scalable and flexible environment and delivers up-to-date information. At the same time, this approach requires innovative data source selection and query optimization techniques.

The Web data integration system has the following components:

- Web sites serving as data sources.
- Search engines and Web query languages providing interfaces to Web sites.
- Wrappers performing information extraction and mapping to a generic data model shared by all sources.
- Mediators integrating information from disparate data sources.

Figure 1 depicts the architecture of the digital library. In this architecture every data source has an associated 'wrapper' that provides access to the underlying data by using source-specific techniques.

Figure 1: Architecture of the digital library.

The role of a 'mediator' is to accept a user query, initiate requests to wrappers, accept and integrate returned result sets, and make them available to a user. The functional architecture of a mediator as presented in Florescu (1998) may include a query rewriter to translate a user query into queries against data sources; a logical plan generator to produce alternative query execution plans; a query optimizer to select the most efficient execution plan; an execution engine to initiate data extraction and receive query results. Meta-data and annotations related to endangered languages are stored in the digital library. The user is then querying or browsing the information in the digital library.

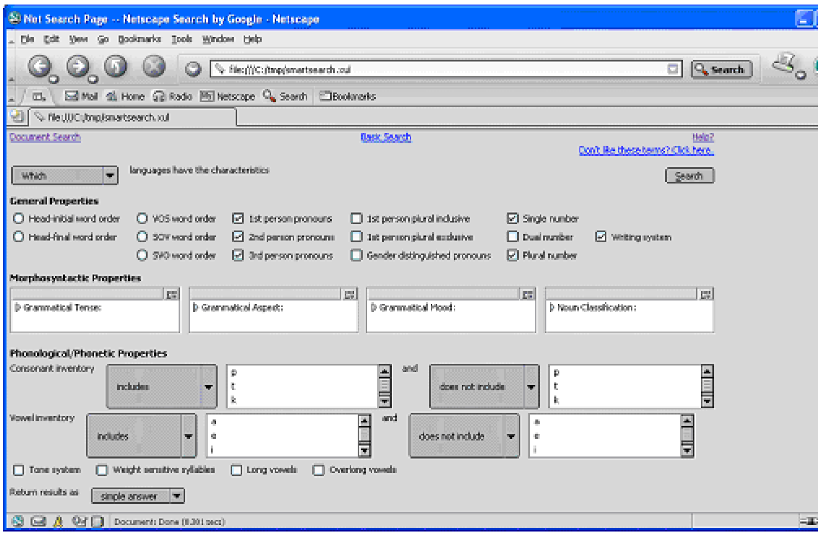
The data repository for metadata and annotations needs to be in formats that will still be readable far into the future. It is generally accepted that the most long-lasting archival format for data is a plain text file with content-oriented markup in an open standard. Currently, the recommended markup standard is XML (Extensible Markup Language), accompanied by a DTD (Document Type Definition, which is a mechanism for specifying document structure, (or, better, a schema) making the markup interpretable by various software programs. We are currently developing an XML-database for the storage of linguistic data of various forms (text, image, video, audio, etc) in multilingual representations (English, French, Chinese, etc). We are designing XML schemas for various linguistic data sources, and use ontological knowledge about their cultural domains in order to integrate them into a Semantic Web for language engineering. Our database approach of mapping semi-structured linguistic data into structured storage will greatly facilitate the intelligent navigation and searching of these linguistic data.

Digital library services

Intelligent querying and retrieval of language data

In a context as vast and rapidly expanding as the Internet, data are only valuable if they are findable, and if their relevance is interpretable through computational means. From a user's point of view, two methods of accessing underlying language data are desirable:

- **Browsing**. Language data are organized into hierarchical taxonomies by topics, language data types, or by linguistic fields, etc., and users can navigate this hierarchy via Web hyperlinks to access various language data organized under different directories.
- **Searching**. A search engine is provided to the user for language data querying and retrieval. (see Figure 2 for a sample linguistic query interface)


Net Search Page -- Netscape Search by Google - Netscape
File Edit View Go Bookmarks Tools Window Help
Search
Mail Home Radio Netscape Search Bookmarks
Document Search
Basic Search
Help?
Don't like these terms? Click here.
Search

file:///C:/tmp/smartsearch.xul

Which languages have the characteristics

General Properties

○ Head-initial word order ○ VOS word order ☑ 1st person pronouns □ 1st person plural inclusive ☑ Single number
○ Head-final word order ○ SOV word order ☑ 2nd person pronouns □ 1st person plural exclusive □ Dual number ☑ Writing system
○ SVO word order ☑ 3rd person pronouns □ Gender distinguished pronouns ☑ Plural number

Morphosyntactic Properties

Grammatical Tense: Grammatical Aspect: Grammatical Mood: Noun Classification:

Phonological/Phonetic Properties

Consonant inventory includes p t k and does not include p t k

Vowel inventory includes a e i and does not include a e i

□ Tone system □ Weight sensitive syllables □ Long vowels □ Overlong vowels

Return results as simple answer

Document: Done (0.301 secs)

Figure 2: A sample of querying languages by various properties

While the LINGUIST List focuses on the development of the browsing access method, and only provides very limited searching functionalities (queries are based on only a couple of attributes and are submitted to a single, centralized database), one of the main goals is to develop an advanced technique to fully support intelligent querying and retrieval of various language data. To develop such a technique, we focus on the following issues:

- *Standard metadata format*. To facilitate resource identification and retrieval, and the interoperability between the system and other related systems, we use the metadata standards proposed by the Open Language Archives community (OLAC). The OLAC metadata standard is based on the widely-used Dublin Core standard, adapted for use with language resources. This allows for material in the digital library to be directly accessible to OLAC-compliant search engines.
- *Multimedia document indexing and retrieval*. Since the digital library will include large amounts of video, image, and audio data, we need to develop efficient and effective multimedia indexing and retrieval techniques. Initially, content analysis and retrieval techniques need to be tailored to the media from the language engineering domain. Then a scheme of combining attributes from various media will be developed. We are developing ontology-based semantic retrieval techniques for various multimedia data. In real-world applications, content based retrieval alone usually can not provide semantically meaningful information (Vailaya, 2001). For example, a search for a red flower by the colour red on a very heterogeneous database cannot be expected to generate meaningful results. In addition, the semantics of the data in multimedia databases is imprecise, and depends on the user's interpretation (Santini, 2001). Even for the same user, the semantic content of an image will change depending on his goal of the search. In order to support effective retrieval, both user's interests and the shift of the user's interests over time have to be reflected in the hierarchical structure of the database. Our approach is to organize the multimedia data into semantic classes and dynamically evolve the class hierarchy based on user's relevance feedback. By including the users in the loop, we could not only accurately capture the meaning of multimedia data, but also build a class hierarchy that reflects the interests of most users.

Display

Accurate display of linguistic documentation is currently inhibited by a number of practices. As is well known by anyone who has tried to bring up an Excel file in Word Perfect, the use of proprietary software often restricts display of the material to those who own a specific program. However, even the use of open formats (e.g., HTML), which are readable by multiple software programs, does not necessarily end display problems, since many such problems are linked to character representation. As is well known by anyone who has ever confronted a browser screen full of tiny squares instead of character glyphs, accurate display of non-Latin characters cannot yet be taken for granted. Also one language might have several encoding mechanisms: for example, popular encoding mechanisms for Chinese characters include GB2312, BIG5, HZ, etc. While the digital library probably would only store linguistic data in one encoding, Web services are needed to convert linguistic data from one encoding to another on the fly. As more and more devices including mobile devices such as PDAs are connected to the Internet, a content adaptation technique is needed to transform the linguistic data from the server to client devices with different display capabilities.

Analysis and cross-linguistic comparison

Efficient data analysis and comparison also require a level of standardization in content markup schemes. Currently, data from different sources may use different structural tags for the same concept (e.g., possessive vs. genitive) or a single tag to reference different concepts (e.g., "absolutive" in a Uto-Aztecan language description does not mean the same thing as "absolutive" in the description of an ergative[1] language). And without compatible markup, no two bodies of data are comparable. Linguistic similarities and differences will not be discoverable by computational means, since no search-engine can be expected to know that differently named entities are equivalent. The solution to this problem is again ontology-based searching. An ontology represents what is essentially a standardized form of the interrelated semantics of an area of knowledge. Thus an ontology can function essentially as an interlingua, to mediate between the terms that a particular linguist or archive might use, and the terms found in other archives, allowing different markups to be related intelligently. The ontology allows the search to know that the two "absolutives" are different, but that "possessive" in one data-set is the same as "genitive" in another. Ontologies can thus be used as integration tools to support unified querying and retrieval from many linguistic data sources simultaneously. Developing an ontology for a discipline is in itself a huge research problem. Fortunately, we are in a position to use the linguistic GOLD ontology being developed by the Wayne State-based E-MELD project at the University of Arizona for this purpose.

Annotation management system

We are currently developing an annotation management system to support domain experts across the world to annotate various data related to endangered languages. This will facilitate advanced queries and retrieval. By means of annotation the semantics of non-text data such as images, audios, videos can be enriched by the knowledge of the domain experts when they describe the contents of these data in metadata. In addition, the resulting metadata can be structured and organized using database technologies to facilitate querying and integration.

As the data related to endangered languages are multimedia and multilingual in nature, and annotation will be performed from different geographical locations, our annotation management system should satisfy the following requirements:

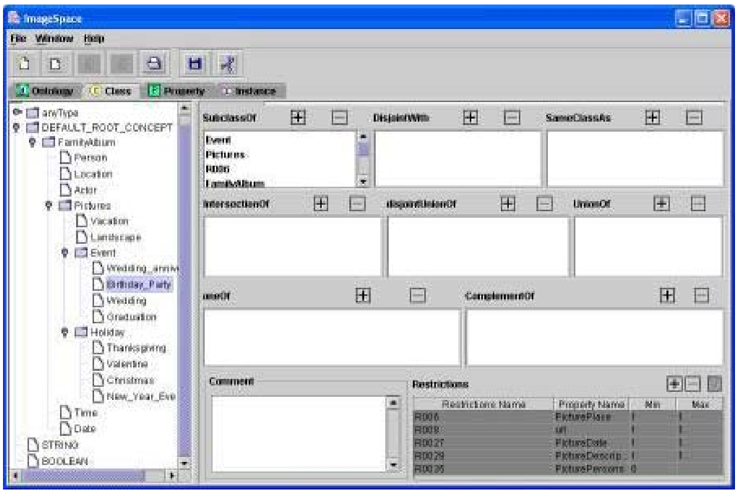
- *Standard metadata format.* To support interoperability within different components of the system and with other related systems, and to store and preserve the metadata over the long term, metadata should be saved in a standard format. The standard Web Ontology Language OWL (Bechhofer, 2003) can be used to specify ontologies for metadata.
- *Support for different data forms and languages.* Our annotation management system will not only support the annotation of Web pages and other standard text formats, but also the annotation of data in various forms. As the data related to endangered languages are multilingual, so will be the metadata. In addition, data in different forms and different languages are usually related from one to another and the annotation of such relationships should be supported by the annotation system.

To meet these requirements and address the relevant challenges, the annotation management system should address the following issues:

- *Creating domain ontologies with standard languages*. Based on OWL (Bechhofer, 2003), and using an off-the-shelf ontology creation tool such as Protege and Ontoedit, we are working together closely with domain experts to create various domain ontologies for different collections of data of the endangered languages. These ontologies will be compatible with the online metadata standard Dublin Core to promote metadata interoperability.

- *Deriving ontologies from existing data sources*. Rather than creating each ontology from scratch, we are investigating the methodologies for automatic or semi-automatic generation of ontologies from existing data sources. In the past, we have investigated a methodology of deriving ontologies automatically for image databases based on the notion of multi-user relevance feedback, we expect this methodology can be extended and generalized to one for this system.
- *Storage and indexing of metadata in XML*. As standard ontology languages for the Semantic Web (Lu, 2002) support XML serialization, our metadata will be in XML format, which has become the standard for representing and exchanging data over the Web (Bray, 1998). Currently, we are developing an XML database based on the approach of mapping the XML model to the relational model (Lu, 2003); we are investigating how the metadata can be stored in an XML database and indexed in a way to facilitate advanced queries and retrieval of metadata using the information specified in the ontologies.

We have developed an ontology creation and annotation tool, called *ImageSpace* to create ontologies and perform ontology-based annotation of text and images. Figure 3 shows a snapshot of creating an ontology using *ImageSpace*. The four tabs, labeled by *Ontology*, *Class*, *Property*, and *Instance*, facilitate the specification of these components and their relationships in a graphical fashion. For example, in Figure 3, the *Class* tab is enabled, the left frame displays the class hierarchy, and the right frame shows the relationships of this class with other classes (including restriction classes). With this interface, one can easily insert, delete, and update a class. In addition, using the right frame, one can specify the relationships of this class with other classes. At the right-bottom corner of the right frame, is a panel that corresponds to property restrictions, where a user can specify both value constraints and cardinality constraints. Note that those shaded property restrictions are automatically inherited from their parent classes unless they are overridden. Also note that, since a class might have multiple parents, other parent classes are shown in the *SubClassOf* field.

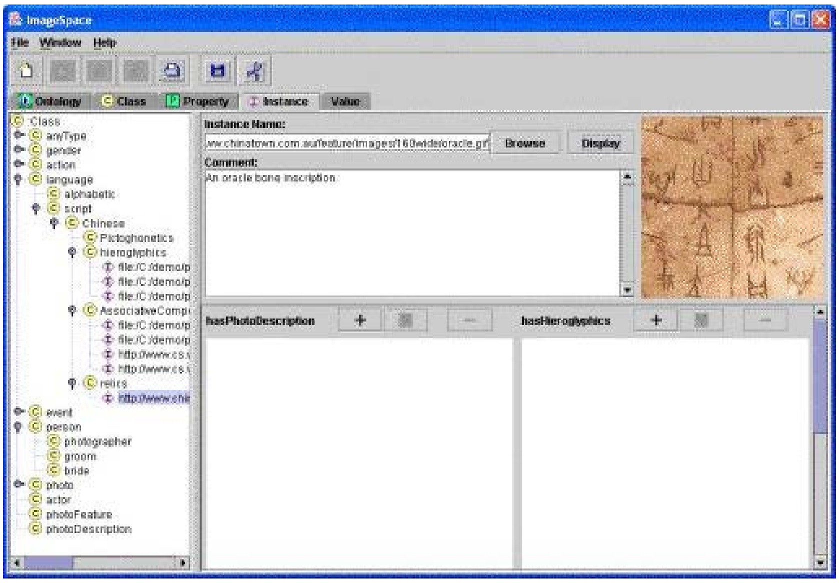


Event
Pictures
R006
FamilyAlbum

Restrictions Name	Property Name	Min	Max
R006	PicturePlace	1	1
R008	url	1	1
R0027	PictureDate	1	1
R0029	PictureDescrip...	1	1
R0035	PicturePersons	0	

Figure 3. Creating an ontology using ImageSpace

Figure 4 displays a snapshot of annotating an oracle bone inscription using *ImageSpace*. The left frame shows the class hierarchy and instances (shown by I-icons) associated with the classes to which they belong. The interface on the right is ontology-driven; that is, for different ontologies and different classes, the interface will be generated dynamically based on the properties and cardinality constraints specified in the ontology.



Instance Name: ww.chinatown.com.au/feature/images/160wide/oracle.gif

Comment: An oracle bone inscription

Figure 4. A snapshot of annotating an oracle bone inscription

A detailed description of *ImageSpace* is beyond the scope of this paper. Interested readers are referred to Huang (2004) for the description of how to create an image ontology and annotate images based on the created ontology using *ImageSpace*, and to Lu (2004) for the description of our experiences of annotating linguistic data with *ImageSpace*. Detailed design and implementation issues as well as other features of *ImageSpace* are available in Huang's master's thesis (2003).

Research plan and methodologies

We will carry out our research project in three years. In the following, we sketch the methodologies and evaluations for each year.

Year 1. In the first year, we will focus on identifying endangered language data sources and developing wrappers for each data source. In particular, we will:

- Design and develop an XML global schema for the proposed virtual warehouse.
- Develop a Web service framework within which Web crawlers will be able to detect, access, and monitor new and old Web sites that serve as reference points for the targeted languages.
- Identify and discover various endangered language data sources and design and develop needed wrappers for the integration of various linguistic data sources. A good starting point of a list of such data sources is the Language Archives Website.
- Enhance FIELD to allow for storage of audio and visual materials.
- Test and evaluate the system.

Year 2. In the second year, we will focus on query design and the development of the Web integration system, as well as the development of the annotation management system. In particular, we will:

- Develop Web-based program to display linguistic documentation.
- Develop queries and their associated Web-based query interface for our digital library.
- Develop content analysis and retrieval techniques for the language engineering domain.
- Develop the annotation management system.
- Conduct user acceptance testing and evaluation.

Year 3. In the third year, we will focus on the development of linguistic analysis and cross-linguistic comparison, language data display, and the testing and evaluation of the whole system. In particular, we will:

- Develop modules for analysis and cross-linguistic comparison.
- Develop additional modules for online translation of linguistic data from one encoding to another.
- Develop content adaptation techniques to transform linguistic data to various client devices such as Personal Digital Assistants with different display capabilities.
- Export functionality of our digital library in terms of standard Web services.
- Test and evaluate the whole system.

Conclusions and future work

In this paper, we have described the effort undertaken at Wayne State University to preserve endangered languages using the state-of-the-art information technologies. In particular, we discuss the issues involved in such an effort, and present architecture of a distributed digital library for endangered languages which will contain various data of endangered languages in the forms of text, image, video, audio and include advanced tools for intelligent cataloguing, indexing, searching and browsing information on languages and language analysis. We use various Semantic Web technologies such as XML, OLAC, and ontologies so that our digital library becomes a useful linguistic resource on the Semantic Web.

Lack of training has been a primary factor impeding progress in the documentation of endangered languages in general. Therefore, it is important to interest young linguists in the techniques of language documentation and train them in fieldwork techniques. Currently, we are in the process of developing a language engineering programme which will provide basic training in the fundamentals of language engineering, including elementary linguistic methodology and the basics of natural language database manipulation and querying. We expect that our digital library for language engineering will become an important tool for this programme.

Note

1. Ergative: a term used of a grammatical case marking the subject of a transitive verb in languages such as Eskimo, Basque, and some others. (Oxford English Dictionary).

References

- Bechhofer, S., Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P. and Stein, L. (2003). *OWL Web ontology language reference. W3C Proposed Recommendation*. World Wide Web Consortium (W3C.org) Retrieved 11 January, 2004 from http://www.w3.org/TR/owl-ref/
- Berreby D. (2003, May 27). Fading species and dying tongues: when the two part ways. *New York Times*, p. F.3
- Bobaljik, J., Pensalfini, R. and Storto, L. (Eds.). (1996). *Papers on language endangerment and the maintenance of linguistic diversity*. Cambridge, MA: Massachusetts Institute of Technology. (MIT Working Papers in Linguistics, 28)
- Bray, T., Paoli, J. and Sperberg-McQueen, C.M. (1998). *Extensible markup language (XML) 1.0. W3C Recommendation 10-February-1998*. World Wide Web Consortium (W3C.org) Retrieved 8 August, 2003 from http://www.w3.org/TR/1998/REC-xml-19980210
- Brenzinger, M. (Ed.). (1992). *Language death: factual and theoretical explorations with special reference to East Africa*. Berlin: Mouton de Gruyter.
- Dixon, R. (1997). *The rise and fall of languages*. Cambridge: Cambridge University Press
- Florescu, D., Levy, A. and Mendelzon, A. (1998). Database techniques for the World-Wide Web: a survey. *SIGMOD Record*, **27**(3), 59-

74. Retrieved 8 August, 2003 from http://citeseer.nj.nec.com/cache/papers/cs/15808/http:zSzzSzwww.dsi.unive.itzSz%7EsmmzSzdocszSzflorescu.pdf/florescu98database.pdf

- Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J. and Widom, J. (1997). The TSIMMIS approach to mediation: data models and languages. *Journal of Intelligent Information Systems*, **8**(2), 17-132. Retrieved 8 August, 2003 from http://citeseer.ist.psu.edu/rd/25367591,43273,1,0.25,Download/http%3AqSqqSqwww-db.stanford.eduqSqpubqSqgarciaqSq1995qSqtsimmis-models-languages.ps
- Grenoble, L. and Whaley, L. (Eds.). (1998). *Endangered languages*. Cambridge: Cambridge University Press.
- Harmelen V. F. and Horrocks I. (Eds.). (2001). Reference description of the DAML+OIL ontology markup language. *DAML+OIL reference description*. Retrieved 8 August, 2003 from http://www.daml.org/2000/12/reference.html
- Huang, R. (2003). *ImageSpace: a DAML+OIL based image ontology and annotation tool.* Unpublished master's dissertation, Wayne State University, Detroit, Michigan, USA.
- Huang,R., Lu, S. and Fotouhi, F. (2004). *ImageSpace: an image ontology creation and annotation tool.* Paper presented at the 19th International Conference on Computers and Their Applications (CATA'2004), Seattle, WA, USA, March 18-20, 2004.
- Krauss, M. (1992). The world's languages in crisis. *Language*, **68**(1), 6-10
- Levy, Y.A., Rajaraman, A., and Ordile, J.J. (1996). Querying heterogeneous information sources using source descriptions. In T. M. Vijayaraman, Alejandro P. Buchmann, C. Mohan, Nandlal L. Sarda. (Eds.) *VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases, September 3-6, 1996, Mumbai (Bombay), India.*. (pp. 251-262) San Francisco, CA: Morgan Kaufman. Retrieved 9 March, 2004 from http://www.acm.org/sigmod/vldb/conf/1996/P251.PDF
- Lu, S., Dong, M. and Fotouhi, F. (2002). The semantic Web: opportunities and challenges for next-generation Web applications. *Information Research*, **7**(4), paper 134. Retrieved 9 March, 2004 from http://informationr.net/ir/7-4/paper134.html
- Lu, S., Sun, Y., Atay, M. and Fotouhi, F. (2003b). A new inlining algorithm for mapping XML DTDs to relational schemas. *In Proc. of of the First International Workshop on XML Schema and Data Management (XSDM'03), in conjunction with the 22nd ACM International conference on Conceptual Modeling (ER'2003)* (pp. 366-377) Heidelberg: Springer-Verlag. (Lecture Notes in Computer Science, 2814) Retrieved 11 January, 2003 from
- Lu, S., Huang, R., and Fotouhi, F. (2004). *Annotating linguistic data with ImageSpace for the preservation of endangered languages.* Paper presented at the 19th International Conference on Computers and Their Applications (CATA'2004), Seattle, WA, USA, March 18-20, 2004.
- Matsumura, K. (Ed.). (1998). *Studies in endangered languages*. Tokyo: Hituzi Syobo.
- May, W. and Lausen G. (2000). *Information extraction from the Web*. Freiburg: Universitat Freiburg, Institut fur Informatik. (Technical Report 136) Retrieved 9 March, 2004 from http://www.informatik.uni-freiburg.de/~dbis/Publications/2K/TR136-InfoExtr.ps
- Robins, R.H., and Uhlenbeck, E. (Eds.). (1991). *Endangered languages*. Oxford: Berg.
- Santini, S., Gupta, A. and Jain, R. (2001). Emergent semantics through interaction in image databases. *IEEE Transactions on Knowledge and Data Engineering*, **13**(3), 337-351. Retrieved 9 March, 2004 from http://www.sdsc.edu/%7Egupta/publications/kde-sp-01.pdf
- Schmidt, A. (1990).*The loss of Australia's aboriginal language heritage*. Canberra: Aboriginal Studies Press.
- Sutherland, W. (2003). Parallel extinction risk and global distribution of languages and species. *Nature*, **423**(6937), 276-279
- Vailaya, A., Figueiredo M.A.T., Jain, A.K. and Zhang, H. (2001). Image classification for content based image retrieval. *IEEE Transactions on Image Processing*, **10**(1), 117-129
- Wurm, S (2001). *Atlas of the world's languages in danger of disappearing*. (2nd ed.). Canberra: UNESCO Publishing.

Scholar Search **Find other papers on this subject** Google Search

How to cite this paper:

Shiyong Lu, et al. (2004) "Language engineering for the Semantic Web: a digital library for endangered languages" *Information Research*, **9**(3) paper 176 [Available at http://InformationR.net/ir/9-3/paper176.html]

Check for citations, using Google Scholar

- Contents |
- Author index |
- Subject index |
- Search |
- Home