Metadata for Web-Crawling

Kevin Scannell
Saint Louis University

WAC Panel Discussion
16 September 2007
Louvain-la-Neuve, Belgium

Primary Goal

- A single, open source tool for accurate web corpus creation in any language
  - This includes the entire process from locating appropriate documents on the web, downloading, converting to text, cleaning, tagging, etc.
  - The problem: a great deal of language-specific information is needed to do this all effectively
  - Most of the data I use for An Crúbadán have been gathered in an *ad hoc* way from native speakers and are not easily available
  - Also, the data are incomplete, even among the existing 416 Crúbadán languages

Examples

- Tokenization rules
- Character inventories
- Stopwords
- 3-gram data for language recognition
- More generally, sample texts for other training purposes
- Information on competing orthographies (e.g. Cornish), scripts (Serbian), dialects (Occitan, Ladin), *ad hoc* character encodings and undocumented or proprietary fonts

Secondary Goal

- Get all of this data off of my hard-drive!
  - The Crúbadán project has become the *de facto* source of most minority language corpora
  - By combining existing, easy-to-use, open-source crawlers like BootCat with the Crúbadán data we can enable people to build their own corpora directly
  - We would also like to make the data available via a scheme that allows them to be used from other online NLP applications

Database Requirements

- Freely available
- Structured according to open standards
- Centralized repository
- Available on the web
- Comprehensive (many languages)
- Complete (includes all of the data needed for web crawling as described above, and extensible to allow new kinds of data)
- Easily queried and parsed by computer programs (top priority)
- Browsable by humans (lower priority)

Possible Solution

- XNL-RDF, created by Oliver Streiter and Mathias Stuflesser
- RDF means “Resource Description Framework”, simply a scheme for expressing metadata, of central importance in the Semantic Web
- Most commonly serialized as XML
- XNL-RDF is a flavor of RDF designed with exactly our requirements in mind

Other Features

- A great deal of data already exist in their database
- Browsable through a web interface
- Exportable to RDF-XML
- Focus is on “scripts”, so orthographic issues are handled by design
- Supports all of the metadata mentioned earlier, plus sentence segmentation rules, number formats, function words, alternate language names (B. Hughes' work on very endangered languages), and much more...

Plans

- I will make Crúbadán data available in this format, including URL lists for all 416 corpora
- Others are encouraged to do the same
- When writing WAC applications consider modularizing in such a way that a language-independent engine interfaces with these data via (free, easy-to-use) RDF parsers
- Collaborate with Ethnologue, OLAC, Rosetta?
- What other language metadata would be useful for WAC applications?