Co-funded by the European Community Programme eContent*plus*

europeana connect

Language Resources Repository: draft specification and latest updates

Europeana Connect Plenary Meeting : WP2 session

Berlin, May 12th, 2010

Aaron Kaplan (Xerox)

Coordinated by the Austrian National Library Österreichische Nationalbibliothek

Resource types

- Stop word lists
- Language guesser(s)
- Morphological analyzers (stemmers, lemmatizers, decompounders)
- Named entity recognizers
- Bilingual dictionaries
- ~~Corpora~~

Consumers

- T2.2 (repository): Month 1-12
- T2.3 (controlled vocabulary mapping): Month 3-21
- T2.4 (query translation): Month 12-24
- (WP3/5 for named entity extraction?)
- The world…

Processes supported by repository

- Store data in repository with associated code, documentation
  - Original format could be text, executable, web service info
- Write code to expose data via common API
  - May involve manually or automatically transforming data to different format
- Make corrections, additions
- Integrate new versions
  - Rerun transformations
  - Merge local changes into new version
- Deploy resources into production

Infrastructure possibilities

- File hierarchy
  - Simple to implement
  - Appropriate for all kinds of data
- SVN
  - Convenient history tracking, particularly for line-oriented files
  - Inefficient for very large files
- JCR
  - (?)

Proposal: SVN for all resources but corpora

- Private (password-protected) europeanalabs.eu server for proprietary resources
- Public (world-accessible) europeanalabs.eu server for open source resources

Directory Structure

```
trunk
  <resource type>
        <resource group>
                README
                common
                        pom.xml
                        src
                                main
                                        orig
                                        transformed
                                        java
                <resource>
                        pom.xml
                        src
                                ...
```

APIs

- Stop words: no API, just simple text files
- `LanguageGuesser`
- `TranslationDictionary`
- `LinguisticAnalyzer`
  - Covers tokenization, lemmatization, decompounding, multi-word detection, part of speech tagging, named entity recognition

Standards

- ISO/DIS 24612: Language resource management -- Linguistic annotation framework (LAF)
- ISO/DIS 24611: Language resource management -- Morpho-syntactic annotation framework

What about corpora?

- Potentially several orders of magnitude larger than other types of resources
- Most important corpus for Europeana is Europeana
- Only one usage (query translation) currently foreseen
- Plan:
  - For now, no corpora in repository
  - If needed, develop separate infrastructure

Web services

- Implementation can be a local call or a web service invocation—transparent to caller.
- May need to add batch invocation possibilities for web services.