

Working Group 5: Web Services

Gary Simons (facilitator)
Reinhard Hiss (rapporteur)
Terry Langendoen
Will Lewis
Ljuba Veselinova

Vision statement

- ◆ The linguistic cyberinfrastructure (eLinguistics) will be built primarily from web services—not desktop applications.
- ◆ A web service is implemented just once.
- ◆ It is then used by many other web apps and many online desktop apps.
- ◆ The single implementation is always the most up-to-date version—no need to update copies.

Current examples

- ◆ OLAC aggregator and search engines
- ◆ ODIN: Online Database of Interlinear Text
- ◆ Resnik's Linguist's Search Engine
- ◆ Huffman's Acquaintance Information Page
- ◆ Altavista's Babelfish

Kinds of service (1)

- ◆ Infrastructure service
 - A set of core functions used by implementers of other services (see below).
- ◆ Repository service
 - A service that supports archival storage of language resources with automated discovery and access.
- ◆ Value-added service
 - An application that builds on other services to add value for a particular subdomain.

Kinds of service (2)

◆ Closed service

- Only people inside the service know how to place new resources into the service.

◆ Open service

- The specifications for adding to the service are published and people outside the service can meet those specs.

Kinds of service (3)

◆ Standalone service

- A service that fails to implement the resource discovery protocol of the cyberinfrastructure.

◆ Integrated service

- A service that implements the resource discovery protocol and supports access to resources by other services.

Possible infrastructure services

1. Resource Discovery Service

- Return all language resources matching the given search criteria.

2. Language Information Service

- Given a language name, return its code.
- Given a language code, return related languages or nearby languages or areas where used.

3. Language Identification Service

- Given a sample of text, return the language it is most likely to be in.

Possible infrastructure services (2)

4. Concept Identification Service

- Given a selection of text, return the GOLD concepts that it appears to be about.

5. Markup Conversion Service

- Given a resource in one markup format, return the equivalent in another—probably different services for different data types.

6. Encoding Conversion Service

- Convert from one encoding to another (limiting conversion to particular markup elements).

Possible infrastructure services (3)

7. Digital Rights Management Service

- For protecting and controlling access to resources that have intellectual property restrictions.

8. Identity Management Service

- For establishing user identity and credentials between institutions in order to access protected materials.

Possible value-added services

9. Add language precision to library and book catalogs.

10. Harvest and regularize resources into an interoperable database—separate services for different types: IGT, wordlists, lexicons, treebanks, grammar paragraphs.

11. Match a WAV file of a sound or sequence against a database of utterances from the world's languages.

12. Perform OCR to a Unicode text stream from a TIFF input (with optional language code).

Language Identification Service

- ◆ An open integrated infrastructure service.
- ◆ Recognized by the community as the single authoritative source.
- ◆ Used by humans and by services.
- ◆ Aims for comprehensive coverage of all languages and encodings for which a written sample is available.
- ◆ Coverage would grow as community adds more training samples.

Use cases

- ◆ Baden Hughes is spidering the 3,000,000,000 documents indexed by Google to build corpora for “low-density” languages.
- ◆ Doug Whalen has been given a whole bunch of legacy data text files; he uploads some to learn what languages they are in.
- ◆ Östen Dahl has found documents in a mysterious language in his drawers; he types in the first few sentences to learn what language it is.

Roles and functionality (1)

◆ Client user

- Accesses service through a form-based UI.
- Enters a text presumed to be in one language (through a text box or a browse button or a URL).
- Service returns a ranked list of languages (with encodings) above a specified threshold of certainty.
- Clicks through to a page giving details (e.g., source and samples) for a result, with feedback form for reporting errors.

Roles and functionality (2)

◆ Client service

- Accesses service through HTTP request (URL plus parameters).
- Passes text sample as a string or as a URL.
- Service returns ranked results as an XML document.

Roles and functionality (3)

◆ Supplier

- Registers to gain credentials as a trusted supplier.
- Logs in to create a new training set or update one previously entered.
- Supplies training sample, ISO 639-3 code, encoding (from open pick list), orthography identification, documents or URLs for testing samples.
- Accepted if trained model gives right answer against all testing samples.

Roles and functionality (4)

◆ Custodian

- Manages registration of suppliers.
- Manages the pick list of encodings.
- Reviews metadata with new submissions and updates as needed.
- Removes troublesome submissions.

More functions

- ◆ Web site lists all suppliers with a page on each listing their contributions.
- ◆ Have parameters to limit search by family or by area.
- ◆ Allow download of models to support offline operation.

Markup Conversion Service for IGT (Interlinear Glossed Text)

- ◆ A closed integrated infrastructure service.
- ◆ But open to new conversions through participation in open source development.
- ◆ Used by humans and by services.
- ◆ Aims for comprehensive coverage of all IGT formats in common use.
- ◆ Coverage would grow as developers add new format options.

Use cases

- ◆ Arienne Dwyer created an IGT in Microsoft Word (to make a point) but then realized she needed to share it with the world in a best practice format; she uploads the document to the converter and gets an ELAN XML document in return.
- ◆ Will Lewis gets tired of implementing IGT conversion routines for more and more formats encountered by ODIN until it dawns on him to use the community's IGT markup conversion web service and save all that work.

Roles and functionality (1)

◆ Client user

- Accesses service through a form-based UI.
- Pastes an IGT into a text box or uploads an IGT file through a browse button or URL.
- Chooses the input format and the desired output format.
- Service returns a document in desired format with equivalent content.
- Uses provided form to report cases (with inputs and outputs) where result not satisfactory.

Roles and functionality (2)

◆ Client service

- Accesses service through HTTP request (URL plus parameters).
- Passes IGT as a string or as a URL, with selected input and output forms
- Service returns a document in desired format with equivalent content.