Proposal for an ad-hoc language resource inventory

joint effort of WP2/5/7

Dieter Van Uytvanck, Andreas Witt, Daan Broeder
2008-05-19

One of the goals of CLARIN is to setup a full-fledged registry to which one can add any collection of language resources and services. However, this task obviously takes some time. Until this registry has fully matured, it is necessary to have an intermediary solution to keep track of the amount and kind of language resources and technology that users would like to add to such a registry. Therefore, we will relize a tentative registration mechanism for linguistic data and tools that will be available in a simple way via the Drupal organized web-site.

This tentative inventory serves a number of functions:

- serve as anchor point for an overview of language resources and tools (WP5)
- serve as an information source about IPR issues related with LRT (WP7)
- start with a registration of data resources and services which CLARIN members can offer (WP2)

Due to its mixed functions on the one hand and the necessity to have a low threshold for entering resources by the CLARIN colleagues a user interface will be suggested that requires only minimal metadata at the beginning. Sub-pages will allow users to add more detailed descriptions. The work package leaders will enter as much as descriptions as they can be based on overviews that have already been carried out at earlier moments.
moments.

1. General Fields

The central part will have a limited number of fields so that the effort is minimal:

- ResourceType (Media, Text, Annotation, Lexicon, KnowledgeSource, Tool, Web-Service, other)
- Name short name of the resource
- Languages languages the resource is about
- Description short clarifying description of the resource
- Country countries the resource was created
- Institute institutes which were responsible in resource creation
- Creator names of the main creators of the resource
- Year year the resource was published/offered
- Format description of the format the resource is in or the software is written in
- MetadataLink link to the metadata description if there is any
- ReferenceLink link to the resource, i.e. to access the data resource, tool or service

The underlined fields have a limited vocabulary, the others are unlimited fields. In case of multiple entries in the restricted fields these should be separated by comma. All these fields should be filled in by the user (or WP leader).

The user interface will offer these fields directly and offer a number of extra buttons which allow people to make more specific entries that might be useful already now in the tentative solution.

2. Data Resources Special Info

We will further on make a distinction between several types of data resources:

2.1 Catalogue Type

Scenario: a user wants to register a complete corpus without adding it as a whole to the archive.

Additional fields:

- WorkingLanguages languages that are used as working languages

- Location — more detailed description than country (place etc)
- Content Type — a description of the subject/purpose
- FormatDetailed — detailed format specifications
- Quality — quality specification
- Applications — area of application
- Project — project within which the resource was created
- Size — size of the resource
- Distribution Form — form in which the resource is distributed
- Access — short description of access terms

2.2 Annotated Media/Text Types

Scenario: a wants to register one or more resources which are either media or text files with or without annotations.

Additional fields:

- Date — date of creation
- Content Type — a description of the subject/purpose
- FormatDetailed — detailed format specifications
- Location — more detailed description than country (place etc)
- Project — project within which the resource was created
- Access — short description of access terms

2.3 Lexicon and Knowledge Source Type

Scenario: a user has a lexicon or knowledge source that needs to be made publicly visible

Additional fields:

- Date — date of creation
- Type — type of lexicon or knowledge source
- FormatDetailed — detailed format specifications
- SchemaRef — either schema reference or short note about schema used
- size — short note about number of entries, RDF triples etc
- WorkingLanguages — languages that are used as working languages
- Access — short description of access terms

3. Applications and Web Services

3.1 Application Special Info (Nuria/Marc)

Scenario: a user has a tool or application that needs to be made publicly visible

Additional fields:

- Date — date of creation/offer
- Type — categorization (stand alone tool, web application, etc)
- Task — task characterization (see e.g. DFKI classification)
- Input — short characterization of input requirements
- ISchemaReference — reference to input schema if available
- Output — short characterization of output generated
- OSchemaReference — reference to output schema if available
- DevDescription — description of relevant aspects of software development aspects
- Environment — description of environment in which software can run
- Application location — where the application can be accessed/retrieved
- Access — short description of access terms

3.2 Services Special Info (Nuria/Marc)

Scenario: a user has a web service that needs to be made publicly visible

Additional fields:

- Date — date of creation/offer
- Task — task characterization (see e.g. DFKI classification)
- Location Webservice — URI where the webservice can be accessed
- InterfaceReference — reference of the interface specification
- Input — short characterization of input requirements
- ISchemaReference — reference to input schema if available
- Output — short characterization of output generated
- OSchemaReference — reference to output schema if available
- DevDescription — description of relevant aspects of software development aspects
- Access — short description of access terms

4. IPR Special Info (Kimmo)

This part serves to create an overview about IPR issues, license types etc as they are relevant for WP7.

- LegalReference — reference to the corresponding legal statement
- EthicalReference — reference to the corresponding ethical statement (code of conduct etc)
- LicenseType — characterization of licensing aspects
- Description — other IPR related descriptions
- Contact Person — person to contact for further information on IPR

5. User Interface Aspects

It is important to offer users to view resourced in table format.

There should be three main orders/filters:

- language
- country
- resource type

Input will be done by forms that will offer typo help.