A first set of integrated web services

Milestone D2.R7.1

2010-02-12   Version: 2

Editors: Marc Kemps-Snijders

www.clarin.eu
A first set of integrated web services

CLARIN-EB-1/2010

EC FP7 project no. FP7-RI-2122230

Milestone: D2.R71 - Deadline: T24

Responsible: Peter Wittenburg

Contributing Partners: MPI, ICS PAS, ILC, ILSP, RACAI, ULisbon, UPF, UTübingen, WROCUT, UFSD

Contributing Members: BBAW, ULeipzig, UStuttgart

© all rights reserved by MPI for Psycholinguistics on behalf of CLARIN
Scope of the Document

This document describes a first set of integrated web services that could be used by all CLARIN members and beyond, i.e. functioning systems that could be used by other communities as well.

This document will be discussed in the appropriate working groups and in the Executive Board. It will be subject of regular adaptations dependent on the progress in CLARIN.

CLARIN References

- Centers Types CLARIN-2008-1 February 2009
- Persistent and Unique Identifiers CLARIN-2008-2 February 2009
- Centers CLARIN-2008-3 February 2009
- Language Resource and Technology Federation CLARIN-2008-4 February 2009
- Metadata Infrastructure for Language Resource and Technology CLARIN-2008-5 February 2009
- Report on web services CLARIN-2008-6 March 2009
- Integration Linguistic processing chains as Web Services: Initial linguistic considerations CLARIN-2009-D5R-3a December 2009
Contents
1. Scope of this document.......................................................................................................................6
2. Terminology........................................................................................................................................7
  2.1. Definitions...................................................................................................................................7
  2.2 Acronyms .....................................................................................................................................8
  2.2 Related Documents ....................................................................................................................10
3. Introduction.......................................................................................................................................11
4. Processing chains and individual web services ................................................................................11
  4.1 WebLicht....................................................................................................................................11
  4.2 IULA web services ....................................................................................................................11
  4.3 ILSP Text Processing Chain(ILSP TPC)...................................................................................11
  4.4 LXService ..................................................................................................................................12
  4.5 RACAI Services.........................................................................................................................12
  4.6 WS-LexPl...................................................................................................................................12
  4.7 LEXUS service ..........................................................................................................................12
  4.8 WROCUT/ICS PAS services......................................................................................................12
  4.10 GATE services .........................................................................................................................12
  4.11 ISOcat services..........................................................................................................................12
1. Scope of this document

This document concentrates on an overview of a number of web services implementing serving a variety of linguistic needs. Of particular interest are of course NLP processing chains and their main technical properties relevant for service integration that present a coherent suite of functionalities. Therefore most of the web service suites described in this document cover the fundamental NLP oriented workflow process as assessed by WP 5 (See figure below). Other web services central for the CLARIN infrastructure are mentioned as well. This document will be discussed in the appropriate working groups and in the Executive Board. It will be subject of regular adaptations dependent on the progress in CLARIN.

![Figure 1: Core workflow](image-url)
2. Terminology

2.1. Definitions

**AAI** [Stanica 2006]
**Authentication and Authorization infrastructure**
An infrastructure that provides Authentication and Authorization Services. The minimum service components include Identity and Privilege Management with respect to users and resources.

**Archive** [CiTER]
repository dedicated to the long-term preservation of the associated data

**IP** [Stanica 2006]
**Identity provider**
An entity in an AAI that performs Identity Management.

**Metadata** [Guenter 2004]
structured information that describes, explains, locates, and otherwise makes it easier to retrieve and use an information resource.

**Metadata registry** [Guenter 2004]
registry
a formal system for the documentation of the element sets, descriptions, semantics, and syntax of one or more metadata schemes

**Provenance data**
Information that provides a traceable record of the origin and source of a resource

**Resource** [Berners-Lee 2005]
The term "resource" is used in a general sense for whatever might be identified by a URI. Familiar examples include an electronic document, an image, a source of information with a consistent purpose (e.g., "today's weather report for Los Angeles"), a service (e.g., an HTTP-to-SMS gateway), and a collection of other resources. A resource is not necessarily accessible via the Internet; e.g., human beings, corporations, and bound books in a library can also be resources. Likewise, abstract concepts can be resources, such as the operators and operands of a mathematical equation, the types of a relationship (e.g., "parent" or "employee"), or numeric values (e.g., zero, one, and infinity).

**Repository** [CiTER]
facility that provides reliable access to managed digital resources

**SOA** [Mackenzie 2006]
**Service Oriented architecture**
A paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains. It provides a uniform means to offer, discover, interact with and use capabilities to produce desired effects consistent with measurable preconditions and expectations.

**SP** [Stanica 2006]
**Service provider**
An entity that provides access to a service based on federated authentication.

**Web service** [Brown 2004]
A web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format.
Workflow [Wulong 2001]
Workflow is a term used to describe the tasks, procedural steps, organizations or people involved, required input and output information, and tools needed for each step in a business process.

2.2 Acronyms

<table>
<thead>
<tr>
<th>Reference</th>
<th>Abbreviation of</th>
<th>Link</th>
<th>Reference</th>
<th>Abbreviation of</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ELDA UC]</td>
<td>Universal Catalogue</td>
<td></td>
<td>[ENABLER]</td>
<td>European Science Foundation Second Learner Study</td>
<td><a href="http://books.google.de/books?id=g292tXMX4tgC&amp;pg=PA1&amp;lpg=PA1&amp;dq=Second+learner+perdue&amp;sourc=bl&amp;ots=Wk13GUQP6&amp;sig=n7QSWy3StXy06nMAzY7GBm9w&amp;hl=de&amp;sa=X&amp;oi=book_result&amp;resnum=3&amp;ct=result#PPP1,M1">http://books.google.de/books?id=g292tXMX4tgC&amp;pg=PA1&amp;lpg=PA1&amp;dq=Second+learner+perdue&amp;sourc=bl&amp;ots=Wk13GUQP6&amp;sig=n7QSWy3StXy06nMAzY7GBm9w&amp;hl=de&amp;sa=X&amp;oi=book_result&amp;resnum=3&amp;ct=result#PPP1,M1</a></td>
</tr>
<tr>
<td>[ISOcat]</td>
<td></td>
<td></td>
<td>[ISOcat]</td>
<td></td>
<td><a href="http://www.isocat.org">http://www.isocat.org</a></td>
</tr>
</tbody>
</table>
Common Language Resources and Technology Infrastructure

[LAF]  Linguistic Annotation Framework
[LIRICS]  Linguistic Infrastructure for Interoperable Resources and Systems  http://lirics.loria.fr/
[LMF]  Lexical Markup Framework  http://www.lexicalmarkupframework.org
[NLSR]  http://registry.dfki.de/
[SIMPLESAML]  SimpleSAMLphp  http://rnd.feide.no/simplesamlphp
[SRU]  Search/Retrieve via URL  http://www.loc.gov/standards/sru/
[SYNAF]  Syntactic Annotation Framework
[UDDI]  Universal Description Discovery and Integration  http://www.oasis-open.org/committees/uddi-spec/doc/ucspecs.htm
[WSDL]  Web Services Description Language  http://www.w3.org/TR/wsd120
2.2 Related Documents

[CLARIN_WS_NOTE] CLARIN note on web services
[D-SPIN_PRES] D-SPIN workshop report and Presentations

[CLARIN_MD_SHRT] CLARIN Component Metadata Shortguide

[CLARIN 2008-1] Centers Types
CLARIN-2008-1
February 2009

[CLARIN-2008-2] Persistent and Unique Identifiers
CLARIN-2008-2
February 2009

[CLARIN-2008-3] Centers
CLARIN-2008-3
February 2009

[CLARIN-2008-4] Language Resource and Technology Federation
CLARIN-2008-4
February 2009

CLARIN-2008-5
February 2009

CLARIN-2008-6
March 2009

[CLARIN-2009-D5R-3a] Linguistic processing chains As Web Services: Initial Linguistic Considerations
http://www.sk.let.uu.nl/u/D5R-3a.pdf
3. Introduction
This document focuses on a number of web services provided in the area of LRTs (Language Resources and Technologies) that have the potential of providing useful functionality available to the CLARIN end users. Almost all of these can be seen as a silo with respect to formats and tag sets used and conversion between them is generally not possible without the use of convertors although in some cases already excellent work has been done to bring certain services together. From the Amsterdam CLARIN Demonstrator workshop\(^1\) it became clear that many organizations supplying web services are willing to make these services interoperable amongst each other as well. At the moment, the scale of this is limited, but this is definitely seen as the direction in which CLARIN must be heading. Further convergence on formats and tag sets and tag set convergence is however still needed here.

4. Processing chains and individual web services

4.1 WebLicht
WebLicht (Web based Linguistic Chaining Tool) is a pilot project by the German D-Spin initiative which aims at providing a platform for constructing and executing NLP web service processing chains to produce incrementally annotated resources. Approximately 25 web services are offered which include tokenizers, sentence boundary detection, POS taggers, NE recognition, lemmatization, constituent parsing, cooccurrence annotation and semantic annotation. It also provides several format convertors to convert from MS Word, PDF, RTF and plain text into the TCF internal format. Services may be integrated by implementing a wrapper that allows the service to be used in the processing chain. It uses a web service registry to select available and suitable web services during processing chain construction. At this moment, CMDI metadata is not yet used to describe text resources and web services and the connection to ISOcat has not been implemented. Both extensions are planned for the future. Weblicht uses a proprietary format and the UPenn and STTS tag sets.

4.2 IULA web services
The IULA suite of web services offers statistical web services, corpus analysis, automatic acquisition of lexical information, Freeling and upload and transformation services. Some services are language independent, language specific ones are capable of handling English, Spanish, Catalan, Galician, Italian, Welsh, Portuguese, and Asturian. It is made available by Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (IULA-UPF), Barcelona, Spain. These services cover tokenization, sentence boundary detection, POS tagging, NE recognition, lemmatization, parsing, cooccurrence annotation, collocation extraction, frequency analysis, association measures and semantic annotations. IULA services are based on the XCES format and uses the EAGLES/PAROLE tag sets. The IULA suite is used to implement one of the exemplary use cases identified in WP 5.

4.3 ILSP Text Processing Chain (ILSP TPC)
The ILSP TPC covers tokenization, sentence boundary detection, POS tagging, lemmatization, chunking and dependency parsing for Greek. It is made available by Institute for Language and Speech Processing (ILSP), Athens, Greece. Access can only be granted through ILSP. ILSP TPC is implemented using UIMA, functionality is not accessible through web services. The format is based on XCES using the EAGLES/PAROLE and Treebank Prague Dependency tag sets.

---

\(^1\) January 25/26, 2010, Meertens Institute, Amsterdam, the Netherlands
4.4 LXService
LXService offers tokenization, chunking and POS tagging for Portuguese. It is made available by University of Lisbon, Department of Informatics, Natural Language and Speech Group (NLX), Lisbon, Portugal and is publically available upon authorization by the providing organization. It uses a proprietary XML format and a proprietary tag set.

4.5 RACAI Services
Racai service offers language identification, tokenization, chunking, POS tagging, parsing and machine translation for English and Romanian. It is made publically available by the Research Institute for Artificial Intelligence, Romanian Academy of Sciences (RACAI), Bucharest, Romania. The format is partly XCES based with proprietary extensions using the EAGLES/PAROLE tag set.

4.6 WS-LexPI
WS-LexPI offers Italian lexicon access. It is made available through the Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale (CNR-ILC), Pisa, Italy. Access is restricted through x509 certificates. The format is proprietary using the SIMPLE tag set.

4.7 LEXUS service
The LEXUS service offers lexicon access. It is made available through the max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands. Access is restricted to authorized users. The format is proprietary. Tag sets may be selected from ISOcat or be proprietary in nature.

4.8 WROCUT/ICS PAS services
WROCUT/ICS PAS services provide tokenization, chunking, POS tagging, lemmatization, parsing, coocurrence annotation, collocation extraction, frequency analysis and association measures for Polish. It is made available under a GPL license by the Institute of Informatics, Wroclaw University of Technology (WROCUT), Wroclaw, Poland and Institute of Computer Science, Polish Academy of Sciences (ICS PAS), Warsaw, Poland. The format is based on XCES with proprietary extensions using the ICS PAL tag set.

4.10 GATE services
GATE offers a full-lifecycle open source solution for text processing. The GATE team has made their ANNIE pipeline (tokenizer, sentence splitter, POS tagger, named entity recogniser), chunker, lemmatizer and POSS tagger for English, a POS tagger for Bulgarian and a POS tagger for Dutch available to CLARIN. These are provided by the University of Sheffield, Department of Computer Science. The produced format is either GATE XML or MAF/SYNAF.

4.11 ISOcat services
ISOcat\(^2\) offers a standard set of web services that provide search and lookup functionality to external applications. These services may be integrated into existing applications to provide referencing information to data categories stored in the Data Category Registry. ISOcat is made available through the Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands which serves as the Registration Authority on behalf of ISO. The exchange format of the services is based on the DCIF format which is the standard interchange format for data category information as described in the ISO 12620:2009 standard.

\(^2\) [www.isocat.org](http://www.isocat.org)