ORTOLANG: a French infrastructure for Open Resources and TOols for LANGuage

Jean-Marie Pierrel, Christophe Parisse

ORTOLANG: a French infrastructure for Open Resources and TOols for LANGuage

Jean-Marie Pierrel [1, 2]
[1] University of Lorraine
[2] CNRS
UMR ATILF
Jean-Marie.Pierrel@atilf.fr

Christophe Parisse [3, 4]
[3] INSERM
[4] University of Paris-Ouest Nanterre,
UMR MoDyCo
cparisse@u-paris10.fr

Abstract

ORTOLANG (Open Resources and Tools for Language: www.ortolang.fr) is a French infrastructure implemented in the framework of the "Programme d'Investissement d'Avenir" (PIA) funded by the "Investissements d'Avenir" French Government program. Based on the existing resource centers CNRTL (www.cnrtl.fr) and SLDR (http://sldr.org/), this infrastructure aims to ensure the management, mutualization, dissemination and long-term preservation of language resources such as corpora, lexicons, terminologies and language processing tools, with particular focus on the languages of France. It will be used as a technical language platform of written and oral language forms, to support the coordination actions of the TGIR HumaNum (http://www.huma-num.fr/).

1 Main characteristics of Ortolang

1.1 Strong emphasis on multidisciplinary openness

The Ortolang project is underpinned by a consortium of laboratories and resource centers with complementary expertise in the following fields:

- sciences of language with ATILF, LPL, MoDyCo and LLL;
- information technology with LORIA and INIST, but also partly with ATILF and LPL;
- data base management and management of access to scientific information, through INIST, and to linguistic resources, through CNRTL and SLDR.

Our aim is not only to combine expertise from different disciplines, but also to bring together – for this language resources and tools mutualization infrastructure – partners who represent the diversity of language study approaches: linguistic modelization, experimental and/or applied linguistics, language production and perception, diachronic studies, sociolinguistics, and the automatic processing of languages (written, oral and multimodal).

Ortolang draws on the wealth of experience gained by the teams supporting the infrastructure:

- the existing means of partners, resource centers (CNRTL and SLDR) and laboratories who offer a set of available resources and tools, and whose expertise covers the three main aspects targeted: oral language, written language and the preservation of the heritage of languages of France;
- involvement in and coherence with TGIR HumaNum;
- coherence with the European infrastructure CLARIN (we worked as part of CLARIN during the preliminary phase);
- and finally coherence with the efforts led by DGLFLF and BNF concerning the heritage aspects of the languages of France.

1. This research was funded by the French State program "Investissements d'Avenir" ORTOLANG managed by the Agence Nationale de la Recherche (grant reference: ANR-11-EQPX-0032).

1.2 An infrastructure that manages resources for the whole scientific community

The Ortolang platform is intended to be a mutualization infrastructure for the management, long-term preservation and dissemination of language corpora, lexicons, terminologies and tools, which of course remain the property of the depositors (researchers or laboratories). Access rights to these resources thus continue to be defined by their owners. On this point, however, Ortolang has made the following strong recommendations:

- compliance with the *Ethics & Big Data Charter*[1], drawn up through the collective efforts of several players engaged in the creation, dissemination and use of data;
- freedom of use for research, provided there is no commercial utilization;
- prior negotiation with the resource owners, whenever there is a desire for commercial exploitation.

With these points in mind, several operations have been set up with partners outside the ORTOLANG consortium that have deposited, or wish to deposit, their resources on ORTOLANG. They include:

- linguistics consortiums (HumaNum) – 'Corpus Ecrits', IRCOM and more recently CORLI - through common calls for projects for the finalization and standardization of corpora;
- the French linguistics research federations ILF (Institut de la Langue Française) and TUL (Typologie et Universaux Linguistiques). Ortolang is thus being used as a medium for the "French reference corpus"[2] initiative of ILF.

2 Objectives and missions of the infrastructure

The objectives and missions of Ortolang can be split into three complementary aspects: identification and preparation of data, long-term preservation of the resources and dissemination.

2.1 Identification and preparation of data

At the present time, one of the difficulties faced in identifying and accessing resources (corpora, lexicons, terminologies and processing tools) stems from their considerable dispersion and the great disparities between them, particularly in terms of coding. Furthermore, over the last twenty years, many language resources of high quality, developed for research projects or theses, have been lost because of a failure to rigorously manage this heritage. This is why the primary objectives are:

- the finalization and standardization of existing resources and tools, with a view to their mutualization. This action is being carried out in close collaboration with the consortiums Corpus Ecrits, IRCOM and now CORLI of TGIR HumaNum. To generate this kind of mutualization momentum in a process that is widely opened up to teams outside the consortium, we have set up funding through calls for common projects with the linguistic consortiums of HumaNum so as to support the necessary work on the standardization of resources that teams outside the consortium wish to deposit on the Ortolang platform;
- the control and validation of resources and tools, including in particular support for the authors of resources about current standards, norms and international recommendations, such as XML, TEI, LMF and SYNAF[3];
- the enrichment of resources and tools. This action is underpinned by the teams involved in Ortolang and is concerned, among other things, with the development of a concordancer that processes large volumes and can be used on any written language corpus, the enrichment of a French morphosyntactic lexicon, the improvement of the time-frame coverage of a French lemmatizer which will then be made available in the form of a Web Service, the development of multilingual sentence segmentation tools, the development of an oral corpus transcription aid tool, the development of plugins to enable interoperability between the various editing and annotation tools, the development of a covering French

[1] http://wiki.ethique-big-data.org.

[2] http://www.ilf.cnrs.fr/spip.php?rubrique95.

[3] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37329

grammar, and the standardization of various corpora including COLAJE, EMERGRAM, L'Est Républicain, ESLO, PFC, and TCOF.

2.2 Long-term preservation of resources

To ensure the long-term preservation of the resources, we have implemented three types of actions:

- the curating of the resources and tools;
- secure storage and maintenance of resources;
- long-term archiving, using the solution set up by TGIR HumaNum[4] in conjunction with CINES[5].

2.3 Dissemination

Finally, to ensure the necessary dissemination and exploitation of the resources, we offer aid and support to users for the setting up of procedures enabling platform users to exploit the mutualized resources and tools by drawing on the prior experience of the resource centers CNRTL and SLDR (which are set to be fully merged into Ortolang).

3 Hardware and software architecture

Ortolang has a hardware architecture set up for the purposes of this project (cf. https://dev.ortolang.fr/doc.infrastructure.html) and recently upgraded. It is based on:

- a cluster of six servers: three R620 servers (48 cores – 768 GB of RAM) and three R630 servers (60 cores – 1152 GB of RAM);
- 165 useful Tb of disks in Raid 6
- A back-up system based on a Quantum library with two LTO6 readers and fifty 300Tb slots.

ORTOLANG Diffusion Service, by mixing a Service Oriented Architecture for high level services and a Software Component Architecture for its Repository Service succeeds in building a robust and reliable Digital Object Repository. This diffusion service is to be made compatible with the recommendations of the CLARIN project for these resource centers, on to which is connected directly the website www.ortolang.fr enabling users to browse through the resources or to select resources via metadata requests. In order to ensure maximum flexibility and maintenance, we choose an SOA (Service Oriented Architecture) pattern to design the software architecture. The software architecture of this platform is described in (Blanchard et al. 2016). Ortolang is accessible via various APIs (REST, OAI-PMH, Handle, FTP): Some components are accessible via multiple interfaces. We provide a REST interface for most of the operations on workspaces and other components of the platform. We provide more specific interfaces like an FTP connection on workspaces in order to upload very large files or numerous files at once. We also manage OAI-PMH interface of published resources and Handle Persistent Identifier on each key that is published. The whole of the code developed is available in Open-source (see https://www.openhub.net/p/ortolang).

3.1 A CLARIN-compatible dissemination centre

The lower layer of the software architecture of Ortolang (the dissemination centre) complies with the constraints of quality of service (maximum availability) and document management meeting DSA (Data Seal of Approval) requirements. The infrastructure (OAI-PMH warehouse), which is largely invisible to users, is a reliable data warehouse (corpora, lexicons, terminologies and language processing tools) incorporating the following functions:

- identification of each resource by means of a Handle;
- proof of integrity of the data associated with a Handle by means of a checksum linked to the Handle;
- metadata: OAI-PMH, OLAC, RDF

[4] http://www.huma-num.fr/services-et-outils/archiver

[5] https://www.cines.fr/

- version management: any modification of data leads to a new version;
- authentication of users via a Single Sign On mechanism, using the Education-Research federation of Renater in the consultation of restricted-access data.

3.2 A user-friendly interface for depositing and consulting resources

At the time of writing, the resource access website (cf. www.ortolang.fr) enables access to a constantly growing set of resources with possibilities of searching for a resource on the basis of standardized metadata (resource type, language, rights of use, source, coding format and annotation types). At the end of September 2016 it offers more than one hundred corpora, twelve lexicons, eleven terminologies, seventeen processing tools and several integrated project, as the lexical portal CNRTL (http://www.cnrtl.fr/portail/) to which the order of 600,000 queries per day in the French lexicon (http://www.cnrtl.fr/aide/stat/)

Special efforts have been made to offer an interface and work spaces that provide depositors with a flexible procedure that is as user-friendly as possible, to enable non-IT specialists to easily deposit their resources and draw attention to them (cf. Figure 1).

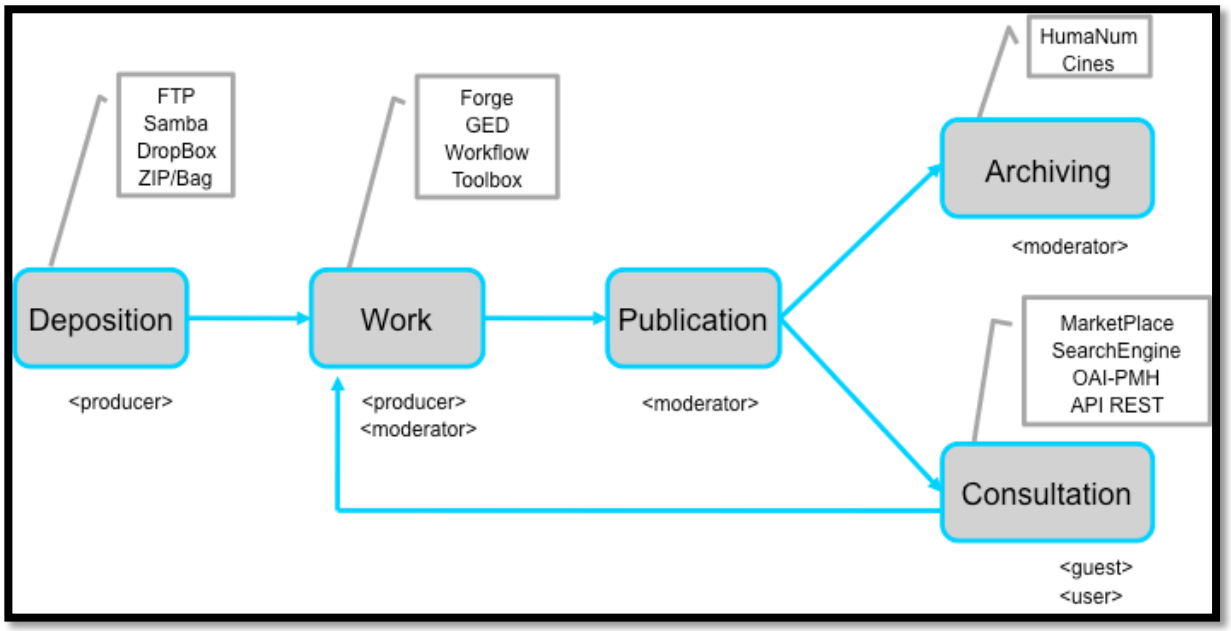

Figure 1: ORTOLANG deposition workflow chart

With this aim in mind, we propose a 5-stage workflow:

- Deposition: After opening an online workspace, the producer is provided with a simple means of depositing the data, even if they are not yet ready for publication. Various methods are proposed for deposition or uploading: via FTP, via a Web interface, or by uploading compressed files. As soon as resources are deposited, they are made secure by the use of reliable media (redundancy) and by daily back-up copies on tape.
- Work in a secure workspace: The producer is provided with specific online tools so that the work can be enriched (alignment, annotation, etc.). During this work phase, access to data is controlled, and data are only visible to the workspace members and platform administrators. Furthermore, resource producers can take advantage of support from three centers of expertise: Written (ATILF/CNRTL), Oral (SLDR & Modyco), and Multi-modal (SLDR & Modyco).
- Publication: once the data are ready, the producer can submit the work for publication. The producer can then monitor the status of his/her requests, and – in collaboration with the administrators – achieve a stable version of the resource.

- Archiving: The published data can be submitted for archiving and/or for consultation. Automatic data enrichment during earlier phases means that the data are “clean” and the archiving format has been checked.
- Consultation and reuse: Data can be consulted in various ways: via a Web interface that sets out all the resources hosted, split into categories and described with a detailed data sheet. Online browsing through the content of resources is also possible. References can be added to published data in a new workspace.

4 Conclusion

Following the French State’s decision to join the CLARIN ERIC with observer status, we suggest that we should study with the CLARIN partners the possibility of ORTOLANG joining the network of CLARIN centers (http://clarin.eu/content/centres).

Reference

Blanchard Jérôme, Pierre Fréderic and Petitjean, Etienne. (2016) ORTOLANG Diffusion - A Component Based Digital Object Repository, Poster presentation, 5th CLARIN Annual Conference, Aix-en-Provence, France, 26-28 October, 2016.