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

Jean-Marie Pierrel, Christophe Parisse, Jérôme Blanchard, Etienne Petitjean, Frédéric Pierre

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

Jean-Marie Pierrel
University of Lorraine
CNRS
UMR ATILF
jean-marie.pierrel@atilf.fr

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

Jérôme Blanchard
CNRS
University of Lorraine
UMR ATILF
jerome.blanchard@atilf.fr

Etienne Petitjean
CNRS
University of Lorraine
UMR ATILF
etienne.petitjean@atilf.fr

Frédéric Pierre
CNRS
University of Lorraine
UMR ATILF
frederic.pierre@atilf.fr

Abstract

ORTOLANG [1] (Open Resources and Tools for Language: www.ortolang.fr) is a French infrastructure which aims to ensure the management, pooling and sharing, 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 for written and oral language forms. The ORTOLANG software platform is based on a new Digital Object Repository service. By combining a Service Oriented Architecture for high level services and a Software Component Architecture for its Repository Service, the platform seeks to build a robust and reliable Digital Object Repository that provides rich functionalities and a modern interface delivering excellent performances and the best optimization strategies. Thanks to its hardware and software architecture choices, the ORTOLANG platform ensures very flexible evolution possibilities to guarantee long-time support for the hosted resources.

1 Main characteristics of ORTOLANG

The ORTOLANG project is a French infrastructure implemented as part of the "Programme d'Investissement d'Avenir" (Investment program for the future) funded by the French Government.

This infrastructure aims to construct a network including a repository of language data (corpora, lexicons, dictionaries etc.) and readily available, well-documented tools for language processing. The repository was built following the guidelines of CLARIN repository centres, so that it could become a CLARIN centre if the opportunity arose and join the European effort to make language resources available. Following the decision of France to join CLARIN as an observer, we would like to rise to

[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).

Jean-Marie Pierrel, Christophe Parisse, Jérôme Blanchard, Etienne Petitjean and Frédéric Pierre 2017. ORTOLANG: a French infrastructure for Open Resources and TOols for LANGuage. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: 102–112.

the task and become a CLARIN B-Centre. The current paper presents the status of ORTOLANG as it is, before finalizing this process.

1.1 Strong emphasis on multidisciplinary openness

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

- linguistics with ATILF (*Analyse et Traitement Informatique de la Langue Française* – Computer Processing and Analysis of the French Language: www.atilf.fr), LPL (*Laboratoire Parole et Langage* – Speech and Language Laboratory: http://www.lpl-aix.fr), MoDyCo (*Modèles, Dynamiques, Corpus* - Models, Dynamics, Corpora : http://www.modyco.fr) and LLL (*Laboratoire Ligérien de Linguistique* – Loire Valley Linguistics Laboratory : www.lll.cnrs.fr);
- information technology with LORIA (Laboratoire lorrain de Recherche en Informatique et ses Applications - Lorraine Research Laboratory in Computer Science and its Applications: www.loria.fr) and INIST (Institut de l'information scientifique et technique - Institute for scientific and technical information www.inist.fr), 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 (*Centre National de Ressources Textuelles et Lexicales* - French National Centre for Textual and Lexical Resources: www.cnrtl.fr) [Pierrel and Petitjean 2007] and SLDR (Speech & Language Data Repository: http://www.sldr.org/) [Bel and Blache 2006].

Our aim is not only to combine expertise from different disciplines, but also to bring together – within this infrastructure for the sharing of language resources and tools – partners who represent the diversity of approaches to language study: constructing linguistic models, 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 centres (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 French heritage of languages;
- involvement in and coherence with TGIR Huma-Num (*Très Grande Infrastructure de Recherche* - Very Large Facility in Humanities and Social Sciences: www.huma-num.fr);
- 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 (*Délégation à la Langue Française et aux Langues de France* - General Delegation for the French language and languages of France: www.dglf.culture.gouv.fr) and BNF (*Bibliothèque Nationale de France* – National French Library: www.bnf.fr) concerning the heritage aspects of the languages of France.

1.2 An infrastructure that manages resources for the whole scientific community

The ORTOLANG platform is intended to be an infrastructure for the management, pooling and sharing, 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*,[2] 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.

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

All the data deposited in ORTOLANG must be made available to the whole public research community. This is mandatory as ORTOLANG is a free public service for research and visibility of the data is requested in exchange for this free service. However, data can be deposited and stored in private workspaces (see part 3.4) for the duration of a project, for example a PhD Thesis or any funded project. This duration is limited in time and cannot be extended more than one year after the end of the thesis or the project. Exceptions to this rule can be made in very special situations, following the guidelines of the linguistic consortia from Huma-Num. In this case, the data will be made public according to the principles of the French public archiving system.

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

- linguistics consortia (Huma-Num) – 'Corpus Ecrits', IRCOM (Corpus Oral and Multimodal**)** and more recently CORLI (Corpus Languages and Interactions: https://corli.huma-num.fr/) - through common calls for projects for the finalization and standardization of corpora;
- the French linguistics research federations ILF (*Institut de la Langue Française* : www.ilf.cnrs.fr) and TUL (*Typologie et Universaux Linguistiques* : www.typologie.cnrs.fr). ORTOLANG is thus being used as a medium for the "French reference corpus"[3] 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 present, 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 pooling and sharing. This action is being carried out in close collaboration with the consortia Corpus Ecrits, IRCOM and now CORLI of TGIR Huma-Num. To generate this kind of sharing momentum and extend it to teams outside the consortium, we have set up funding through calls for common projects with the linguistic consortia of Huma-Num 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, metadata, norms and international recommendations, such as XML (Extensible Markup Language), TEI (Text encoding Initiative: www.tei-c.org), LMF (Lexical Markup Framework: www.lexicalmarkupframework.org).

2.2 Long-term preservation of resources

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

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

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

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

[5] Centre Informatique National de l'Enseignement Supérieur : https://www.cines.fr/

2.3 Dissemination

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

3 Hardware and software architecture

3.1 Hardware

ORTOLANG has a hardware architecture specifically designed for the purposes of this project (cf. https://dev.ortolang.fr/doc.infrastructure.html) and recently upgraded. The ORTOLANG hardware architecture is based on a dedicated cluster of computing servers, a SAN (Storage Area Network) and an automated tape backup system (LTO6). The entire software platform is hosted on a virtualized environment solution (VMWare) in order to provide complete flexibility (CPU, RAM and storage dynamic allocations) to better suit each service need. Internet connectivity is ensured using two redundant connections (10Gb/s and 1Gb/s) and two firewalls. The internal network connectivity between servers, Storage Area Network and backup system uses up to 12 Fiber Channel links. The whole infrastructure is hosted and operated by INIST, one of the members of the ORTOLANG consortium.

Our current hardware equipment is composed of:

- a cluster of six servers: three DELL R620 servers (48 cores – 768 GB of RAM) and three DELL 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.

3.2 Software architecture

The goal of the ORTOLANG Diffusion Service is to build a robust and reliable Digital Object Repository. It is based on a Service Oriented Architecture for high level services and a Software Component Architecture for its repository service. This diffusion service will fully comply with the recommendations of CLARIN for the resource centres in the near future. The service is connected directly to the website www.ortolang.fr, enabling users to browse through resources or to select resources via metadata requests. The software architecture of this platform is described below in section 4. ORTOLANG is accessible via various Application Programing Interfaces (APIs): REST [Richardson & Ruby 2007], OAI-PMH,[6] Handle Persistent Identifier, FTP.[7] 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 such as an FTP connection on workspaces in order to upload very large files or numerous files at once. We also manage an OAI-PMH interface of published resources and the Handle Persistent Identifier on each file that is published. Our implementation is free and open-source (LGPLv3[8]). The source code is available online from an open source software repository (see https://www.openhub.net/p/ortolang).

3.3 A CLARIN-compatible dissemination centre

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

[6] Open Archives Initiative - Protocol for Metadata Harvesting: https://www.openarchives.org

[7] File Transfer Protocol: https://www.w3.org/Protocols/rfc959/

[8] https://www.gnu.org/licenses/lgpl-3.0.fr.html

- identification of each resource by means of a Handle Persistent Identifier;
- proof of integrity of the data associated with a Handle by means of a checksum linked to the Handle;
- metadata: OAI-PMH, OAI Dublin-Core, OLAC,[9] CMDI,[10] RDF;[11]
- 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 (Réseau National de télécommunications pour la Technologie l'Enseignement et la Recherche: National telecommunication network for technology, teaching, and research) to authenticate users requesting access to restricted data.

3.4 A user-friendly interface for depositing and consulting resources

Special efforts have been made to offer an interface and workspaces 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.

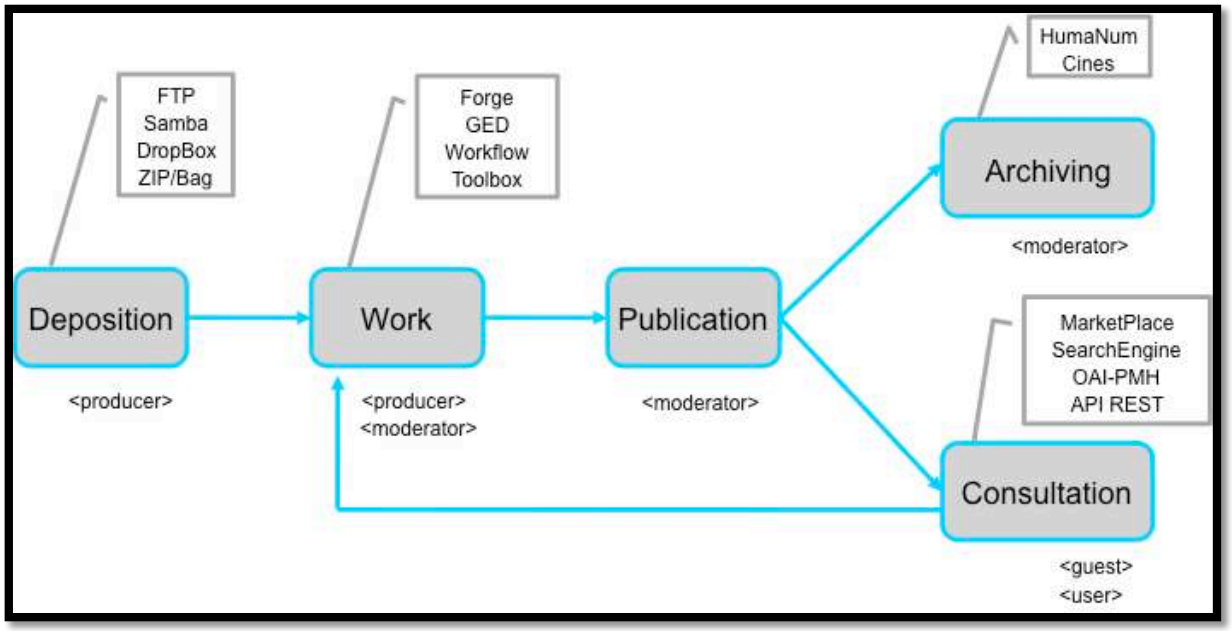

Figure 1: ORTOLANG deposition workflow chart

With this objective in mind, we propose a 5-stage workflow (see Figure 1):

- Depositing: 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 secured by the use of reliable media (redundancy) and by daily back-up copies on tape.
- Working in a secure workspace: The producer is provided with specific online tools so that metadata can be edited in a user-friendly way and the work can be enriched (alignment, annotation, etc.). Metadata can be edited at the general level (information about the whole project) and can also be further specified at the document level (information and rights about specific documents, for example). Metadata can be described in French and in English. 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

[9] http://www.language-archives.org/

[10] https://www.clarin.eu/content/component-metadata

[11] https://www.w3.org/RDF/

advantage of support from three expertcentres specializing in written (ATILF/CNRTL), oral (SLDR & Modyco), and multi-modal (SLDR & Modyco) data.

- Publishing: 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: ORTOLANG is not responsible for the archiving process itself. This is handled by Huma-Num and CINES. ORTOLANG will submit the published data for archiving. Automatic data enrichment during earlier phases means that the data are "clean" and the archiving format has been checked. All the data that conform to the archiving conditions of the CINES will be forwarded for public archiving to Huma-Num and CINES.
- Consulting and reusing: Data can be consulted in various ways: via a Web interface that lists all the published resources, which are split into categories and described with detailed metadata. Online browsing through the content of resources is also possible. References can be added to published data in a new workspace.

4 Software organisation

4.1 Initial requirements

Building on the experimental work we carried out in 2011 in CLARIN, we analyzed the characteristics of the CLARIN network of centres,[12] looked more specifically at specific centres such as LINDAT/CLARIN,[13] and at the CLARIN B Centre checklist.[14] Besides, we analyzed the needs of the French research community and defined use-case scenarios for such a platform, as described in the previous section. Our requirements also included using open source and free third-party software, performing data de-duplication, setting fine-grained access control rules, giving scalability to millions of objects and ensuring a fast response time.

In 2011, when we first started work on a digital object repository, we took the time to analyze and test existing solutions to find a platform that would meet our requirements. Back then, we tested two different software solutions. The first one, Fedora,[15] was an interesting platform and we started developing a proof of concept demo that would lay the groundwork for a larger digital repository. Unfortunately, we encountered many technical challenges that were hard to overcome. The underlying software lacked flexibility and would not scale up to the requirements we had set ourselves. The second platform we analyzed was D-Space.[16] But at the time, no serious work appeared to be ongoing and the software development stalled.

These reasons drove us to the conclusion that to meet our requirements, we needed to start developing our own software architecture that would create a solid Digital Object Repository.

4.2 Software architecture

In order to ensure maximum flexibility and maintenance, we chose a Service Oriented Architecture pattern to design the software architecture. The application relies on six Top Level Services and a farm of dedicated business specific services. The main repository service (ORTOLANG Diffusion) is designed using a Software Component Architecture and implemented using JEE (Java Enterprise Edition) Technologies.

4.3 High level services

All six top level services and the tools farm are hosted independently in the cluster and are based on different kinds of software. These services are connected using very simple protocols (mostly HTTP[17]) to ensure loose coupling, easy deployment, load balancing and scalability.

[12] https://www.clarin.eu/content/overview-clarin-centres
[13] https://lindat.mff.cuni.cz/en
[14] hdl:1839/00-DOCS.CLARIN.EU-78
[15] http://fedorarepository.org/
[16] https://www.dspace.com
[17] Hypertext Transfer Protocol: https://www.w3.org/Protocols/

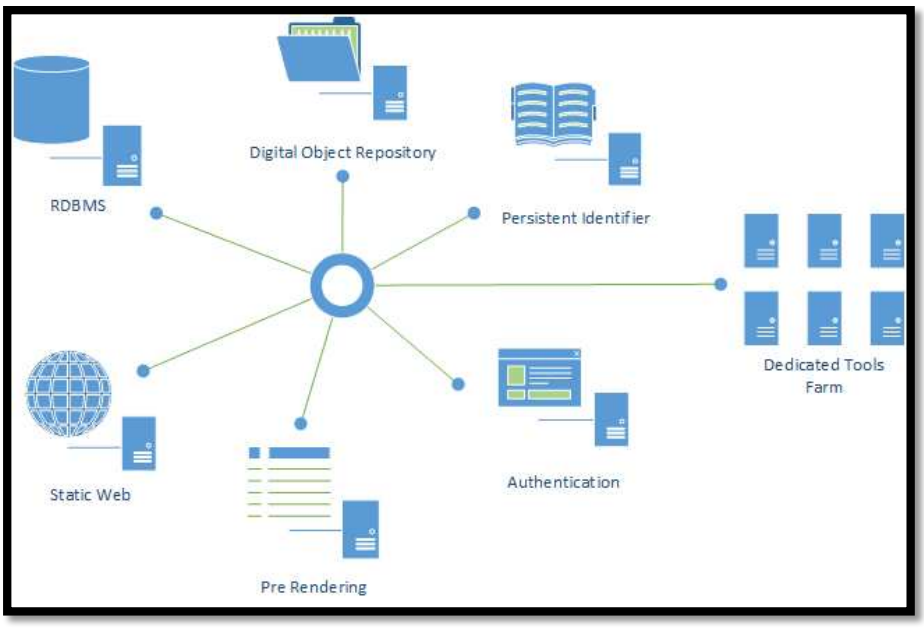

Figure 2: ORTOLANG High Level Services

Relational database management system: We use a relational database management system (PostgreSQL[18]) to store all the service data that require transaction isolation. Because of the very large volume of binary data, we avoid storing binary content in the database but only critical data that require ACID (Atomicity, Consistency, Isolation, Durability) properties [Haerder and Reuter 1983]. This service is hosted in its own virtual machine and can be scaled either by increasing VM capacity or by using clustering strategy.

Authentication: As we need an external authentication mechanism, we chose the OAuth protocol.[19] This protocol allows authentication (Direct Grant) without the need for a user interface.

To do so, we use Keycloak,[20] an open source identity management component developed by Jboss.[21] Keycloak provides connectors for most of Java Application Servers but also for Javascript applications. To handle RENATER integration, we developed a small application that plays the role of the Shibboleth Discovery Service.[22]

For users who are not registered in the French authentication system (or in the future European system), it is possible to access the data that are free (which are a large part of the ORTOLANG data) or to use a classical identification with username and password. This identification does not provide the same rights as the institutional authentication.

Repository: The digital repository service aims to manage user resources from the first deposit to the final publication providing specific features for each phase of the resource life cycle. This service has been developed internally in order to meet our requirements but relies on some low-level software components that have been embedded in the repository application to propose functionalities such as File Storage, Version Configuration, Content Management.

Persistent identifier: We provide persistent resource identifiers based on the Handle system. We have packaged the Handle.net server software as a platform service but in read-only mode. Thus, it is up to the repository application to create and update handle entries in the database to ensure consistency. This is achieved by including Handle write operations in a global transaction in the repository service.

[18] https://www.postgresql.org/

[19] https://oauth.net/

[20] http://www.keycloak.org

[21] http://www.jboss.org/

[22] http://shibboleth.net/

Web server: Client side applications (repository browser and administration) are developed using HTML5/Javascript and use the AngularJS framework.[23] These applications only need to be served as static web content. We use Nginx[24] to serve these applications but also to do a specific routing of requests coming from search engine indexers to a dedicated service.

Pre-rendering: In order to ensure the best search engine indexation, this service provides static versions of the website pages. It acts like a Client Web Application browser and stores the rendered pages in order to serve them directly to the search engine indexer bot, thus avoiding Javascript interpretation side effects.

Tools farm: Specific applications can interact with the repository using its REST API. This allows anybody to develop their own application (for example a specific file format conversion application) and submit this application as a tool for ORTOLANG. These applications are self-contained and must provide their own user interface. Authentication is done using OAuth and applications can access user information and data only if authorized by the user. Using this mechanism ensures that each application has permissions granted by a user to be able to access data in the repository.

4.4 Repository service architecture

The digital object repository service (ORTOLANG Diffusion) business logic is complex and is the result of merging the logic of many existing components. It provides a virtual online versioned file-system (like DropBox) for each resource: a workspace. Around this workspace, we provide the ability to enrich content by setting some metadata using the format provided for all types of content (files, folders, and resources). These metadata will be indexed to populate an internal search engine and to give visibility to the published workspace in a kind of market place that will present all the published resources. Collaborative functionalities and publication processes allow a group of people to work on the same resource before and after its publication. All the business logic is defined in dedicated components that are exposed through interfaces providing a consistent repository service.

Implementation is done in a Java JEE application using EJB (Enterprise JavaBeans), JPA (Java Persistence API) and JMS (Java Message Service**)** to ensure the robustness and stability of the platform. Some EJB components wrap subsystems that are completely embedded in the platform such as a BPEL (Business Process Execution Language**)** Engine,[25] a NoSQL Database[26] or a Lucene index base.[27] This component wrapping avoids coupling between platform components and embedded ones making it possible to switch to any other implementation. That's the key point of an SCA Architecture. Each component is also testable independently using mock strategy.

Main principles: In order to avoid coupling between software components, we have based the identification of all objects in the repository on a unique key managed by a registry. All operations are performed using the key of an object. The registry maintains the association between a key and the concrete object in the database (group, workspace, collection, process) by mapping an object identifier to its registry key. An object identifier is composed of its service name, its type name and its internal id. The registry manages some common aspects of all concrete objects e.g. the state, lock, author, properties, history, etc.

[23] https://angularjs.org/

[24] http://nginx.org/en/

[25] http://www.activiti.org

[26] http://orientdb.com

[27] https://lucene.apache.org/

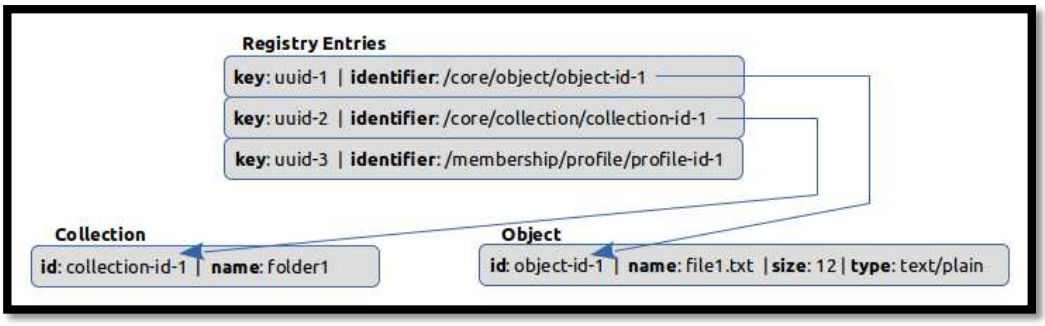

Figure 3: Registry Service Mapping

Data granularity choices: We have opted for fine-grained objects, thus the smallest object in the system is a simple file. Objects can also be larger, and as big as a complete part of a file system. This has a very large side effect because each object that is referenced in the registry can have its own history, owner, security rules, index entry, publication state, etc. We chose to place granularity at the lowest level possible but we also defined a structure to organize files into folders (collections) and folders into workspaces using a one to many relation.

Workspace and version control: Most of the repository business logic is organized around a structure that we call a workspace and that holds the content of a complete resource. A workspace, like a versioned file-system, manages folders and files and keeps track of their modifications; it also manages metadata to describe the resource and its content. Once a producer assumes that his workspace content has reached a sufficient level of maturity or needs to be visible by external users, the content can be submitted to the publication process. Multiple versions of the same resource can be published to follow the resource's life cycle.

Binary content aspects: Binary content associated with files is not stored as is. A dedicated component is used to generate a hash (SHA-1: Secure Hash Algorithm 1[28]) for each binary stream stored and organizes the physical storage on an underlying file system that can be shared in multiple volumes. Using a hash SHA-1 as identifier for stored streams means that de-duplication of identical binary content can be performed.

Metadata: It is possible to set metadata on any object in a workspace using a one to many named relation. Metadata are typed, and for some particular types, content is structured using a schema.

We use the JSON (JavaScript Object Notation) data format[29] for structured metadata. These structured metadata are indexed in a dedicated NoSQL database (OrientDB[30]) in order to produce an enriched database that allows to search for objects based on their characteristics and using a rich query language. The metadata stored in the system can be converted to an adequate format for metadata harvesting such as OLAC Dublin Core used by Isidore (a French search platform allowing access to digital data in the Humanities and Social Sciences: https://www.rechercheisidore.fr/) or CMDI used by CLARIN.

Security concerns: A set of security rules is defined for each registry key and enables specific permissions to be stored on each object. Security is enforced by a dedicated component that can be queried for a specific permission (read, update, delete) on a particular key.

Asynchronous treatments: We use Java Messaging Service in order to handle asynchronous jobs. We have dedicated a topic in which all platform events are fired. Some listeners are in charge of triggering actions on event reception (indexing a file, extracting metadata, notifying, and logging).

Process management: In order to manage the publication and review processes, we have defined a runtime component that can handle Business Process Model and Notation scripts.[31] This component is a wrapper over a well-known BPEL Engine (Business Process Execution Language**)**: Activiti. BPEL

[28] DOI: http://dx.doi.org/10.6028/NIST.FIPS.180-4

[29] http://www.json.org/json-fr.html

[30] http://orientdb.com/orientdb/

[31] http://www.bpmn.org/

processes are injected into the runtime component which allows transactional processing and human tasks management.

Search engine: We use asynchronous indexers for search functionalities. A key/value base (Apache Lucene[32]) is designed for full text searches whereas the NoSQL base allows more specific queries and faceted results. All search results are security filtered to ensure privacy.

API (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 such as an FTP connection on workspaces in order to upload very large files or numerous files at once. We also manage the OAI-PMH interface of published resources and Handle Persistent Identifier on each key that is published.

Performances: We have carried out performance tests on the repository that show that we are able to store more than 25 million objects without significant response time modification.

5 Tools farm ecosystem

We plan to open a tools farm to provide specific processing for resources. Some specific language treatment tools (tokenizer, text extraction, speech and text alignment, etc.) will be integrated as external applications and, relying on the OAuth grant permission mechanism, will access ORTOLANG resource files in a secure way, even before publication of the content. These tools will help resource providers to work on their files before publication but will also allow users to apply some treatments on a selected set of ORTOLANG resources.

We have already produced a proof of concept by integrating TreeTagger [Schmid 1994] as an external tool for ORTOLANG. We are about to release a file conversion tool (avconv[33]) and a concordancer allowing indexation of a specific set of files for dedicated search.

At the same time, we are working on a sample web application that will make new tool integration easy by customizing this application connected to the ORTOLANG Repository. In the best case, the customization will consist in writing a simple HTML form and a mapping between this form's values and a shell command line.

All the tools will be hosted in their own server, either on the ORTOLANG cluster or on an external server and will use the ORTOLANG REST API.

6 Achievement of the project

As of January 2017, the first phase of the project is finished. Improvements to the infrastructure, especially the user interface, are still ongoing, to take into account the actual use of the infrastructure by the researchers. Some software developments are still in the final phase and will be included in the project in 2017.

New developments are scheduled and should be finalized during the second phase of the EQUIPEX project. Our target concerns mostly the enrichment of resources and tools. The goals include 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 development of an oral corpus transcription aid tool, the development of plugins to enable interoperability between the various editing and annotation tools, and the standardization of various corpora including COLAJE [Morgenstern and Parisse 2012], L'Est Républicain [ATILF 2011], ESLO [Eshkol-Taravella et al. 2012], PFC [Durand et al. 2009], and TCOF [ATILF 2017]. Some tools are already available as independent tools in the "Tools" section of the ORTOLANG main repository.

At the time of writing, the website (www.ortolang.fr) offers access to a constantly growing set of resources with possibilities of searching for a resource based on standardized metadata (resource type, language, rights of use, source, coding format and annotation types). At the end of March 2017, the platform hosted almost 189 corpora, 14 lexicons, 21 terminologies, 27 processing tools and several integrated projects, such as the CNRTL lexical portal (http://www.cnrtl.fr/portail/) serving more than 600,000 queries a day (http://www.cnrtl.fr/aide/stat/). This corresponds to a total 4.9 To of data and more than 300,000 files.

[32] https://lucene.apache.org/

[33] https://libav.org/avconv.html

7 Conclusion

After more than two years of effort, we have deployed a dedicated hardware cluster that hosts the ORTOLANG platform, and have developed a new Digital Object Repository in accordance with our needs and requirements. ORTOLANG already acts as a robust choice to deposit resources and metadata.
The final objectives of our work are to comply with the guidelines of the Data Seal of Approval and to comply with the technical requirements defined for CLARIN centres.

As France has joined the CLARIN ERIC with observer status, our current goal is to become a CLARIN B-Centre (http://clarin.eu/content/centres) by the end of 2017.

References

[ATILF 2011] ATILF 2011. Corpus journalistique issu de l'Est Républicain [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/est_republicain/v1, https://hdl.handle.net/11403/est_republicain/v1

[ATILF2017] ATILF 2017. TCOF: Traitement de Corpus Oraux en Français [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/tcof/v1.

[Bel and Blache 2006] Bernard Bel and Philippe Blache. 2006. Le Centre de Ressources pour la Description de l'Oral (CRDO). Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence (TIPA), vol. 25. 2006, p. 13-18.

[Durand and al. 2009] Jacques Durand, Bernard Laks & Chantal Lyche. 2009. Le projet PFC: une source de données primaires structurées. In J. Durand, B. Laks et C. Lyche (eds)(2009) Phonologie, variation et accents du français. Paris: Hermès. pp. 19-61, https://hdl.handle.net/11403/pfc/v1.

[Eshkol-Taravella et al. 2012] Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C., Tellier I. 2012, Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012., in Ressources linguistiques libres, Traitement Automatique des Langues. Volume 52 – n° 3/2011, 17-46, https://hdl.handle.net/11403/eslo/v1

[Haerder and Reuter 1983] Theo Haerder, Andreas Reuter, A. (1983). "Principles of transaction-oriented database recovery". ACM Computing Surveys. 15 (4): 287

[Morgenstern and Parisse 2012] Aliyah Morgenstern and Christophe Parisse 2012. The Paris Corpus, *Journal of French Language Studies,* 22/01, 7–12, https://hdl.handle.net/11403/colaje/v1.1.

[Pierrel and Petitjean 2007] Jean-Marie Pierrel et Etienne Petitjean. 2007. Valorisation et exploitation scientifiques de documents numériques pour la recherche en linguistique : l'exemple du CNRTL, *Actes de CIDE 2007 Congrès International sur le Document Numérique*, Nancy, 2-4 juillet 2007, p13-24, Europia 2007, ISBN 978-2-909285-38-2

[Richardson and Ruby 2007] Leonard Richardson and Sam Ruby. 2007. RESTful Web Services, O'Reilly Media, 454 p.

[Schmid 1994] Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.