Ortolang

Open Resources and Tools for LANGuage
Outils et Ressources pour un Traitement
Optimisé de la LANGue

INVESTISSEMENTS D'AVENIR

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

contact@atilf.fr

ORTOLANG was funded by the French State program "Investissements d'Avenir" managed by the ANR (ANR-11-EQPX-0032)

Main characteristics of Ortolang

- underpinned by a consortium of laboratories and resource centers
  - 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

Experience of the teams supporting the infrastructure

- existing means of partners, resource centers (CNRTL and SLDR) and laboratories
  - set of available resources and tools
  - expertise in oral language, written language and the preservation of the heritage of languages of France;
- involvement in and coherence with TGIR HumaNum;
- experience with the European infrastructure CLARIN;
- coherence with the efforts led by DGLFLF and BNF concerning the heritage aspects of the languages of France.

An infrastructure that manages resources for the whole scientific community

- compliance with the *Ethics & Big Data Charter*, 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

- linguistics consortiums (HumaNum) – more recently CORLI
  - common calls for projects for the finalization and standardization of corpora;
- 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” initiative of ILF

Objectives and missions of the infrastructure

- identification and preparation of data
  - finalization and standardization of existing resources and tools, with a view to their mutualization
  - control and validation of resources and tools
  - enrichment of resources and tools
- long-term preservation of the resources
  - curating the resources and tools;
  - secure storage and maintenance of resources;
  - long-term archiving, using the solution set up by TGIR HumaNum in conjunction with CINES
- dissemination
  - aid and support to exploit the mutualized resources and tools by drawing

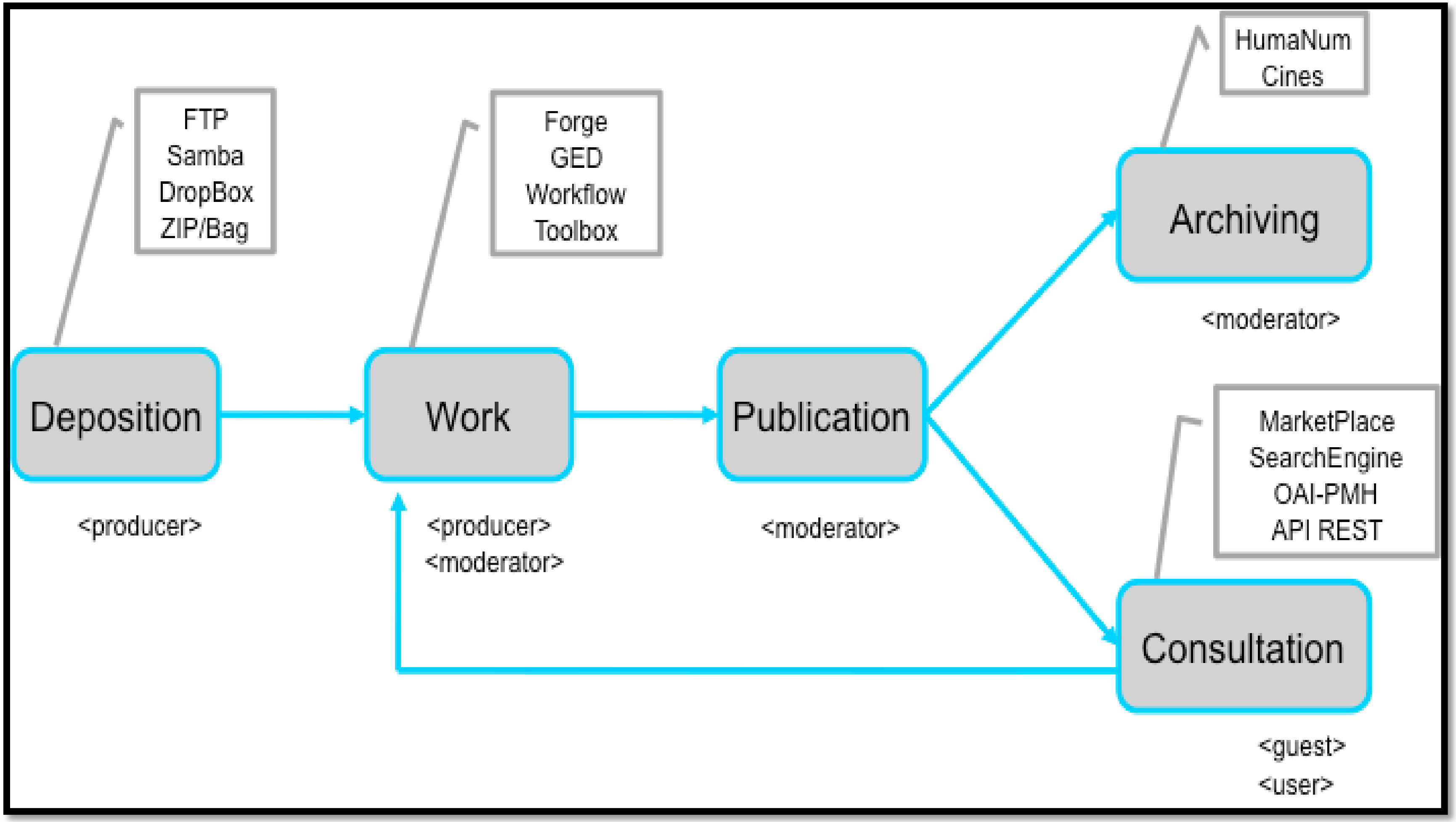
Hardware and software architecture

- hardware architecture set up for the purposes of this project
  - a cluster of six servers: three R620 servers and three R630 servers
  - 165 useful Tb of disks in Raid 6
  - back-up system based on a Quantum library with two LTO6 readers and fifty 300Tb slots
- Ortolang is accessible via various APIs (REST, OAI-PMH, Handle, FTP)
  - code available in Open-source

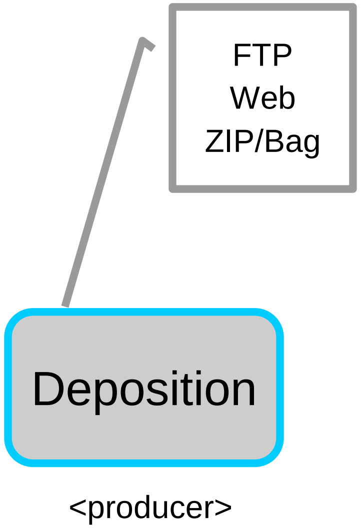
CLARIN-compatible dissemination centre

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

A 5-stage workflow

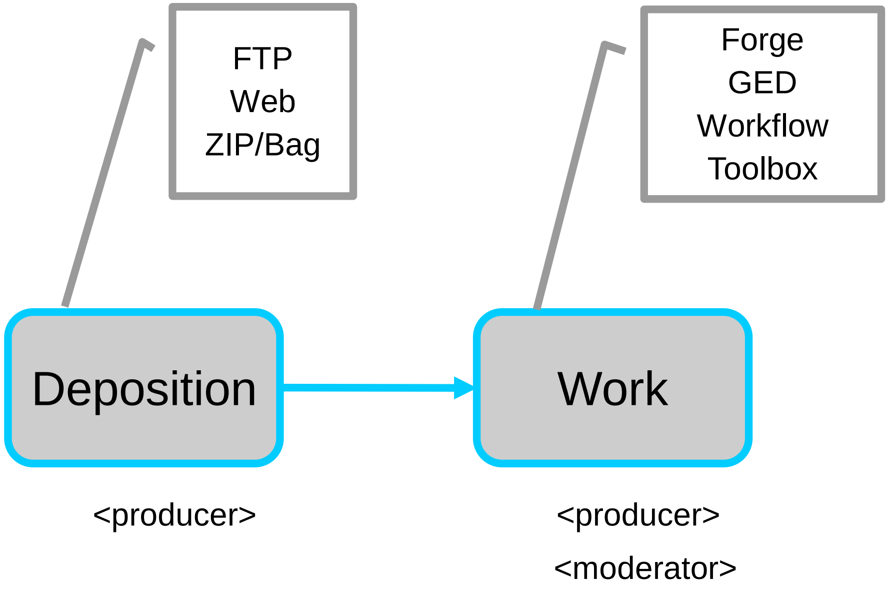


Deposition



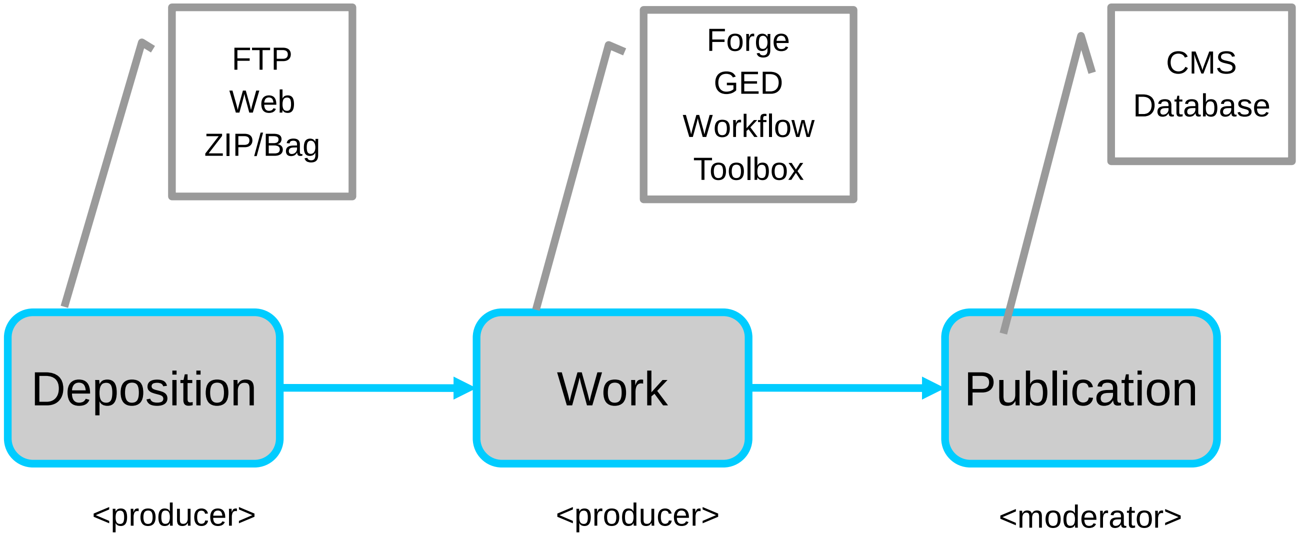
- After opening an online workspace, the producer is provided with a simple means of depositing the data, even if they are not yet ready for publication.
- Different methods :
  - via FTP
  - via web interface
  - via uploading compressed files.

Secure workspace



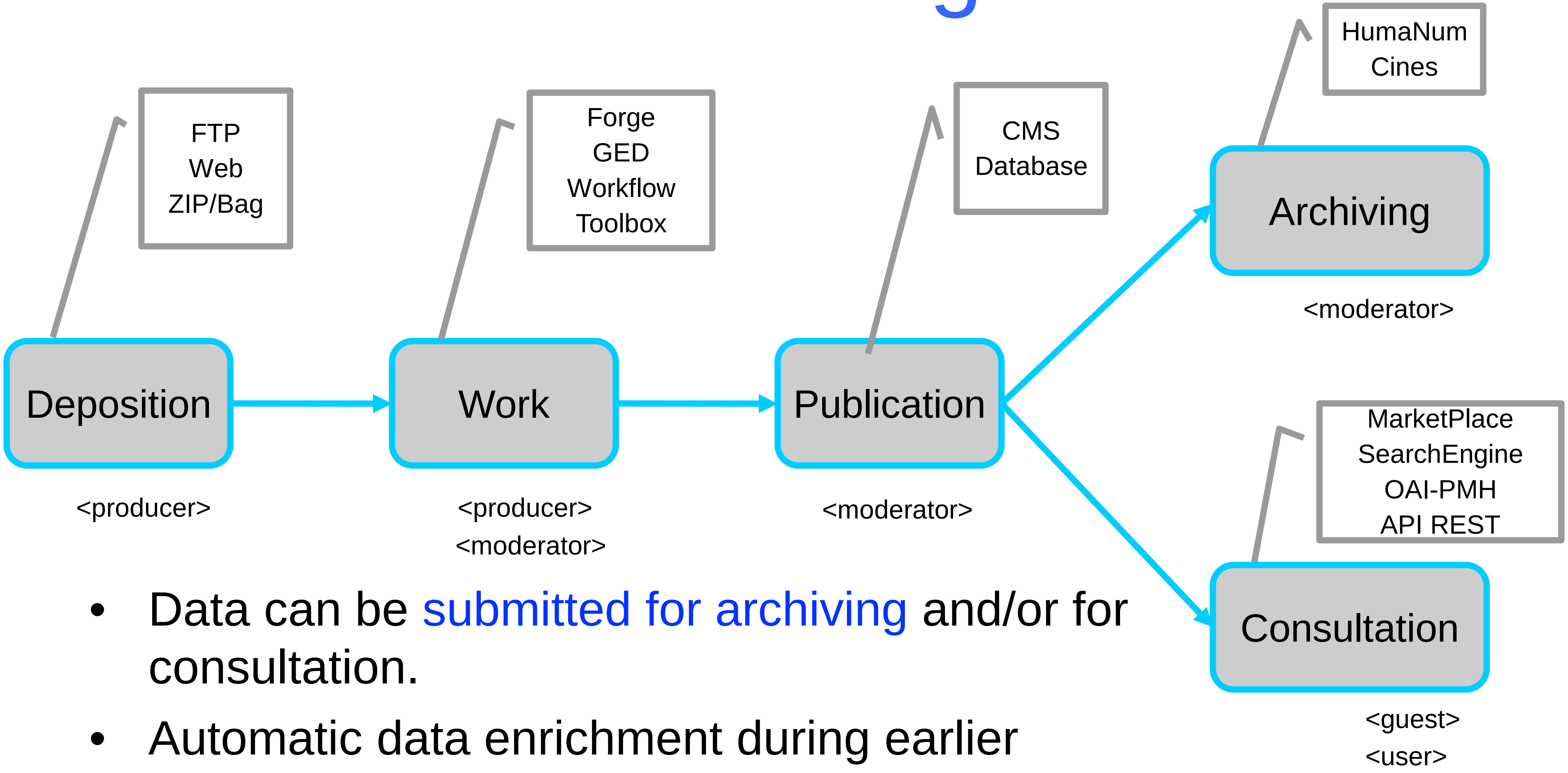
- Online interface, modification enabled.
- Daily backup of all data
- Online tools
- Access to data controlled (members of workgroup only)

Publication



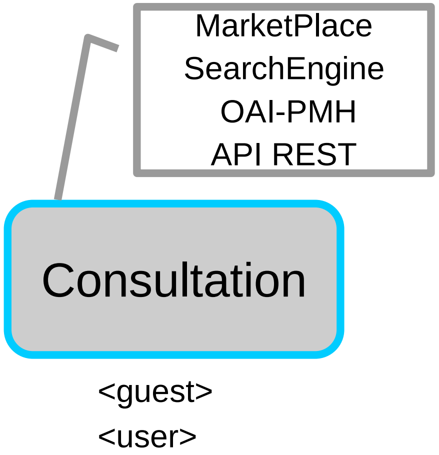
- Once the data are ready, the producer can submit the work for publication
- The producer can then monitor the status of his/her requests.
- Several level of authorization are available
- Control and support of the three centers expertise: Written (ATILF/CNRTL), Oral (SLDR & Modyco), and Multi-modal (SLDR & Modyco)

Archiving



- 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.

Browsing and reuse



MarketPlace
SearchEngine
OAI-PMH
API REST

Consultation

<guest>
<user>

- Data can be consulted in various ways:
  - via a Web interface (metadata information).
  - online browsing of the content of resources
  - OAI-PMH
  - REST interface
- Handle for citation
- Reuse following the licence conditions.