Assessment Information

CoreTrustSeal Requirements 2020–2023

Repository:	ORTOLANG
Website:	https://www.ortolang.fr
Certification period:	03 August 2023 - 02 August 2026
Requirements version:	CoreTrustSeal Requirements 2020-2022
This repository is owned by:	**CNRS**

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Repository Type

Please provide context for your repository. You can select one or multiple options.

Compliance level:

Not Applicable - 0

Response:

- Research project repository

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Description of Repository

Provide a short overview of the repository.

Compliance level:

Not Applicable - 0

Response:

ORTOLANG (Open Resources and Tools for Language: https://www.ortolang.fr) was a research project to build a national infrastructure for the deposition, curation, dissemination and long-term preservation of resources (corpora, lexicons, dictionaries, etc.) and tools created by the French research community in language sciences.

ORTOLANG started as a consortium of 6 public research laboratories:

- ATILF (Analyse et Traitement Informatique de la Langue Française)
- INIST (Institut de l'Information Scientifique et Technique)
- LLL (Laboratoire Ligérien de Linguistique)
- LORIA (Laboratoire Lorrain de Recherche en Informatique)
- LPL (Laboratoire Parole et Langage)
- MODYCO (Modèles, Dynamiques, Corpus)

Each member has contributed with a variety of technical and scholarly expertise to the project. The ORTOLANG platform is intended to be a mutualization infrastructure to promote the French language by sharing the knowledge acquired by public laboratories. Automatic processing of data is focused on French language. However, data deposited by all French public laboratories is accepted for all languages. Other services can be provided on a case by case principle.

The project was funded by the French government as part of “Programme d’Investissements d’Avenir” (PIA) investment plan from 2012 to 2019. Since 2020, four partners (ATILF, INIST, LPL and MODYCO) have agreed to provide staff to continue the maintenance and operation of the now fully functional repository and develop new functionalities and services for the research community.

We have joined the CLARIN (Common LAnguage Research INfrastructure) european network as a C centre. We intent to apply for B center status soon. Our technical infrastructure is already almost fully compatible with the requirements for the certification as a B Centre. Our repository has been harvested by the CLARIN VLO (Virtual Language Observatory) since 2016.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Designated Community

Provide a clear definition of the Designated Community

Compliance level:

Not Applicable - 0

Response:

The producers of resources deposited on ORTOLANG are the members of the Language Sciences research community. They publish their research data on the platform to ensure better dissemination and preservation of their work. ORTOLANG accepts all kinds of formats and data types but encourages users to follow the recommendations and best practices of the research community (see [1] and R8).

Potential users are any humanities, social sciences and language sciences researchers looking for linguistics resources to perform various tasks, e.g. research project, education material, computational linguistics, NLP (Natural Language Processing), etc. Open resources can also be used by SME*s which often cannot afford the cost and time of developing such resources and tools.

[1] https://www.ortolang.fr/en/help/data-formats/

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Level of Curation

Select all relevant types of curation.

- Content distributed as deposited
- Basic curation – e.g., brief checking, addition of basic metadata or documentation
- Enhanced curation – e.g., conversion to new formats, enhancement of documentation

- Data-level curation – as above, but with additional editing of deposited data for accuracy

Compliance level:

Not Applicable - 0

Response:

- B. Basic curation – e.g. brief checking; addition of basic metadata or documentation

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Level of Curation - explanation

Please add the description for your Level(s) of Curation.

Compliance level:

Not Applicable - 0

Response:

When users deposit resources in the repository, the first step is to create a workspace for their project. They can then upload their data.
The next step involves providing the metadata about their project. Some entry fields are optional, but most are mandatory (e.g. licenses).

We tried to ease as much as we could the burden of filling metadata by providing pre-filled menus for many fields (e.g. resource type, language, licenses, etc.).

The final step before publication is to define the access rights via a very simple web form. When the user submits his/her data for publication, the system automatically checks whether the metadata is completely filled in.

The platform uses a workflow engine to handle the publishing process. Reviewers receive a notification. They perform various tasks of data and metadata curation (error checks, missing information, URLs, data formats validation, etc.) [1].

The reviewers never edit the metadata and data provided by the producers. They check for errors and inconsistancies in the metadata (missing or incomplete information) and data (unicode errors, XML validation tools, etc.). If a problem is found, reviewers use an internal communication system to dialogue with the producer and report problems. This often leads to a back and forth exchange.

Once everything is correct, the reviewer accepts the resource for publishing.

[1] https://www.ortolang.fr/en/help/data-curation/

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Insource/Outsource Partners

If applicable, please list them.

Compliance level:

Not Applicable - 0

Response:

Huma-Num (Humanités Numériques) is a national research infrastructure (https://www.huma-num.fr/about-us) offering services (hosting, data deposition and preservation, computing grid, search engine, etc.) for the humanities in France. ORTOLANG is a specialized partner of Huma-Num for the language sciences.

CINES (National Computing Center for Higher Education) is a national service (https://www.cines.fr/en/) for the long-term preservation of data and the partner of the French National Archive System for digital data. ORTOLANG uses the services of CINES to provide long-term preservation for data available on the platform.

RENATER (REseau NAtional de Télécommunications pour la technologie, l'Enseignement et la Recherche) is the national research infrastructure (https://www.renater.fr/) providing network connectivity to all research institutions in France. It also provides national services like an Identity Federation (part of the european federation GEANT) to provide secure authentication for all research personnel.

CORLI (Corpus, Langues, Interactions) is a Huma-Num consortium (https://corli.huma-num.fr/en/) and a CLARIN K Centre (https://corli.huma-num.fr/en/kcentre) bringing together researchers and teaching researchers in linguistics. Its objective is a federation of teams, laboratories, researchers, and teaching researchers* engaged in the production and analysis of digital, oral, and written corpora, regardless of the language and/or the writing system. ORTOLANG is working closely with CORLI to follow its recommendations and best practices (legal, metadata, etc.).

VLO (https://vlo.clarin.eu) and ISIDORE (https://isidore.science/) are search engines specialized in humanities. ORTOLANG's metadata is harvested by these two search engines to improve visibility of research data available in the repository.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Significant Changes

Summary of Significant Changes Since Last Application if applicable.

Compliance level:

Not Applicable - 0

Response:

-

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Other Relevant Information

You may provide other relevant information that is not covered by the requirements.

Compliance level:

Not Applicable - 0

Response:

During the development phase from 2013 to 2019, the project hired temporary staff (10 people) to develop the platform and provide quality language resources. Since 2020, the platform has been maintained and improved thanks to three full-time equivalents (FTEs) (see R5).

The repository currently hosts almost 600 resources, which amounts to 15TB. As of October 2022, we have 2732 registered users and the number keeps growing.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Organizational Infrastructure

R1 Mission/Scope

The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance level:

The repository is in the implementation phase - 3

Response:

ORTOLANG was created as part of an EQUIPEX grant awarded as a result of a nation-wide funding program (aka Investissements d'avenir) for innovative and creative research projects. Its main goal is to offer a fully functional repository for deposition, dissemination and preservation of all types of linguistic data.

The mission of the platform is to accept data coming from all language research fields in France and also about all the languages of France. Datasets from other origins are also accepted whenever their volume fits within ORTOLANG capacities and especially when the data come from countries where no public repositories like ORTOLANG are available.

ORTOLANG is based on a consortium of four laboratories (ATILF, INIST, LPL, and MODYCO) which are all permanent CNRS laboratories. CNRS (National Centre for Scientific Research) is the largest public research operating organization in France. ORTOLANG has strong links with the research community in corpus linguistics through two closely related channels. For all technical and scientific aspects, ORTOLANG works in collaboration with CORLI (see R0).

ORTOLANG is also one of the services provided by Huma-Num (https://www.huma-num.fr), a very large national infrastructure for the digital humanities (see R0).

ORTOLANG currently offers medium-term preservation of deposited data by using a highly secure storage environment.
Although the research data are stored using several security layers, ORTOLANG's goal is to provide long-term preservation for the data deposited by researchers.

In order to do so, we have signed a Service Level Agreement with Huma-Num [1]. This agreement allows us to use the services of the CINES to store all the research data deposited on the platform. CINES has a strategic partnership with Huma-Num and is one of the national facilities certified by the French government [2] to provide long-term preservation.

The ORTOLANG platform has been fully functional since 2016. A detailed document describes ORTOLANG's general policy and user policy [3].

[1] https://www.ortolang.fr/wp-content/uploads/2022/10/Charte_Archivage.pdf
[2] https://francearchives.fr/fr/article/26287437
[3] https://www.ortolang.fr/en/home/charter

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

My last comment was the following:
Very good to see that the LTP workflow with CINES is now implemented. I am still a bit doubtful as to the content of the contract between ORTOLANG and the LTP providing party CINES. The contract is only available in French, but as far as I understand not all data will be get for the long term? Ity would be good to have a summary in English that catches the essence with respect to the retention policy. A further complication to my mind is that CINES does not currently have a certified TDR.

Furthermore, I earlier understood that only part of the data in Ortolang will be ingested in the CINES long-term archive. This would be a choice to be made by the depositor. Is this still the case?
Because of both of these reasons I still believe a compliance level 3 is the best fit, until I have more information on the two points above.

I cannot see what the exact changes in the evidence in this version are. Based on the text and the links (in French), I am inclined to stay with a level 3 here and would expect to see improvement over the next three years.

R2 Licenses

The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

There are four access levels for resources available:

- Public: the resource is open access and freely available/downloadable to the general public (anonymous users)
- Identified: the resource can only be accessed by identified users, but there is no constraint about the community the user is a member of
- Restricted to the research community: the resource is protected and researchers need to authenticate themselves on the national research identity provider RENATER [1] to be allowed to access and download the resource
- Private: the resource is protected and can only be accessed and downloaded by selected members.

Those levels are the only access levels available on the ORTOLANG platform.
When they deposit their resources, researchers can specify the type of associated license. Currently, we offer the following licensing schemes:

- Creative Commons 3.0
- Creative Commons 4.0
- GNU LGPL
- CeCILL 2 [2]
- CeCILL-C

The goal of ORTOLANG is to promote data shared within the research community so we do not promote any type of license that challenges the use of data. We can provide specific licenses whenever required by the origin of the data, but we do our best to accept only licenses that do not restrain access.

In case a suitable license cannot be found in already existing licenses [3], submitters can contact us with a request to create a custom license but Creative Common licenses are the preferred choice on the platform.

A user interface guides the user through the access level options for each file and for each folder, and the license he/she chooses for a whole corpus.

Website users can view the access level and the applicable license. They have to accept them explicitly by clicking a check box stating that they agree with the license and any specific requirements. Access control is provided automatically through registration to the website and user identification. Users cannot download items that they aren't allowed to access. Available items are visible to users, unless a folder is fully protected. In this case, users are only aware of the existence of the folder.

[1] https://services.renater.fr/federation/en/index
[2] See https://en.wikipedia.org/wiki/CeCILL for CECILL licensing schemes
[3] https://www.ortolang.fr/en/help/licenses/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R3 Continuity of access

The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance level:

The repository is in the implementation phase - 3

Response:

ORTOLANG was funded from 2012 to 2019 but is guaranteed to operate in the foreseeable future thanks to the renewed commitment of fours of its members (ATILF, INIST, LPL and MODYCO), four long-established CNRS laboratories [1]. The 2012-2019 project has made it possible to construct the present ORTOLANG repository.

This is made possible by the specific organization of research in France, where budgets for permanent human staff are kept separate from those for time-limited funding and temporary staff. Human resources are assigned to laboratories on a long-term basis, while special equipment investments and manpower can be earmarked in separate project funding budgets.

ATILF, INIST, LPL and MODYCO have commited to provide staff (3 FTE) to keep maintaining the software platform and continue to improve the services and provide new functionalities to the research community in language sciences. ORTOLANG is also proactive to find additional fundings through its

involvement in regional, national or european projects.

Each member of ORTOLANG has the following responsibilities:

- ATILF is the maintainer of the software architecture.
- INIST is the maintainer of the hardware architecture and provides systems administration / supervision.
- MODYCO and LPL provide support and advice to the users and promote the platform.

The stability of the French laboratory research system can be demonstrated by the longevity of the four members of ORTOLANG. LPL is 50 years old, INIST is 35 years old, and ATILF and MODYCO are 21 years. Moreover, these laboratories are following the steps of their predecesors, which means that the creation of a laboratory does not stop them from the preservation of past work from their staff.

The implication of the four French laboratories covers the medium term (and even longer) continuity. It also guaranties that total lack of funding is unlikely and transition to other support for the long term will be possible.

At the same time, ORTOLANG took measures to preserve data access even in case of some unexpected situations:

- The source code for the platform is open source and available in public repositories.
- All the technologies used are open source and well documented to facilitate future transfers.
- The software architecture can be easily provisioned in any cloud infrastructure.
- All the data can be easily downloaded by their contributors and hosted elsewhere.

Finally, all the resources deposited by contributors will be archived (when their format is compatible) by the CINES. The CINES has the mission to provide long term archiving the official French community of Higher Education and Research. The data stored in this institute has a long-term preservation lifetime (see https://www.cines.fr/en/long-term-preservation/).

[1] https://www.ortolang.fr/en/home/continuity-of-services/

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Response Text: See below
Thanks to the applicant for answering my earlier questions around: what will happen in the case of cessation of funding, which could be through an unexpected withdrawal of funding, a planned ending of funding for a time-limited project repository, or a shift of host institution interests?

IThis certainly provides more clarity. However, there does not seem to be a written agreement with the four host organisations that guarantees that they will take over the responsibility in the case of a service discontinuity.

With this being the case the level cannot be more than 3.

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

R4 Confidentiality/Ethics

The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository includes resources provided by all the research laboratories and researchers producing language resources for the French language. The resources available in our repository contain sufficient metadata (name of producers, research papers, etc.) for users to assess the scientific quality of the research data.

We rely on a policy (https://www.ortolang.fr/en/home/charter/) defining the terms of use of the service and submission rules to achieve the best possible quality of deposited resources in compliance with the linguistic field ethics and norms.

Authentication is required for data involving disclosure risks. Moreover, data with levels of disclosure risk are archived but not distributed, and all resources deposited by researchers also go through a dedicated review workflow before publication [1].

ORTOLANG follows the recommendations of the ethics and legal workgroup of CORLI (https://corli.huma-num.fr/en/group-projects/gp4/), which represents the linguistic scientific community and informs its members about data privacy requirements. ORTOLANG also follows CORLI's scientific recommendations, especially about the content of the metadata and suggestion about the format of the deposited data.

The ORTOLANG team is in charge of reviewing the quality of metadata and data provided by producers, following the recommendations above. A built-in message board is included for direct contact between reviewers and producers. The ORTOLANG team does its best to make sure that the deposited resources comply with our guidelines and policies to offer the best available quality to the research community.

Users are informed about data privacy requirements before creating an account and at any time on the website [2] to comply with the General Data Protection Regulation (GPDR). The Data Protection Officer of the CNRS conducted a full review of our platform in 2022 in order to guarantee our compliance to the GPDR.

[1] https://www.ortolang.fr/en/help/publishing-workflow/
[2] https://www.ortolang.fr/en/home/privacy-policy/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R5 Organizational infrastructure

The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

ORTOLANG is a consortium of CNRS public research laboratories. Since 2020, four partner laboratories have confirmed their commitment to maintaining and perpetuating the platform: ATILF, INIST, LPL and MODYCO.

ATILF has been a major architect of the ORTOLANG repository and its success. The laboratory coordinated all the administrative and financial management of the repository. ATILF also contributed all its experience in terms of linguistic resources on written language as well as its expertise in IT development and the use of research data. Its engineers developed the entire software infrastructure of the ORTOLANG platform. Now, the laboratory maintains its unfailing commitment to the sustainability of the repository through the work of its engineers.

LPL has made a major contribution to the construction of the ORTOLANG platform by providing its initial version with a collection of more than 200 datasets (data, metadata, etc.) and tools. Currently, the LPL remains an active partner in the ORTOLANG repository by participating in decisions and policies related to the organization and operation of the repository, and bringing technical and scientific backing to an infrastructure for pooling written and oral linguistic resources.

MODYCO has extensive experience in the creation of oral and multimodal language corpora. It has brought to ORTOLANG large corpora as well as tools and techniques for oral language processing. Modyco has also been involved in consortia of linguistics (CORLIs) since their creation.

INIST hosts a large infrastructure of platforms developed with its partners, such as ORTOLANG. It has an IT Department with a dedicated team in charge of infrastructure and IT production. This team ensures continuity of service all year round by using advanced tools to provide supervision and control of all operating platforms.

The ORTOLANG staff is composed of:

- Experienced researchers in language sciences (oral, written and multimedia)
- A senior research engineer with 20+ years of experience in software development
- A senior research engineer with 20+ years of experience in repository management and infrastructures
- Highly skilled/experienced software engineers to maintain the platform

The staff is appointed by the directors of the laboratories, according to the personnel available and technically competent to do the work in the permanent personnel of these laboratories. This is the standard practice in France for work that does not rely on funding by an external source or a time limited project.

The governance of the consortium is provided by a Director and two Deputy Directors. The members of the consortium are part of an orientation committee in charge of steering the technical and scientific monitoring of the repository. The committee meets annually to discuss the evolution of the repository, its financing and its long-term visibility.

The staff also regularly participates in national and international conferences (EOSC, CLARIN, etc) and events to acquire or improve knowledge and to stay in touch with the latest information in humanities, software development and research.

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R6 Expert guidance

The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

ORTOLANG relies on two types of guidance:

- Internal:

Through consortium and regular staff meetings, ORTOLANG’s orientation committee receives feedback and guidance from its members to improve and set priorities for the enhancement and development of the platform.

- External:

ORTOLANG works in close collaboration with Huma-Num as we share the same goal. We are especially closely involved with CORLI, a French linguistic consortium (https://corli.huma-num.fr/en) and a CLARIN K Centre (https://corli.huma-num.fr/en/kcentre) working on a wide range of topics such as data formats, tools, legal issues, etc. CORLI represents the community of users and they can provide us feedback from this community. They also can follow its evolution and change in the demands of the users. Also, CORLI provides us information about recommended formats for data and metadata.

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Digital Object Management

R7 Data integrity and authenticity

The repository guarantees the integrity and authenticity of the data.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

Producers have three possibilities to upload files to the repository:
- Using the Web interface (drag & drop).
- Using a secure FTP connection.
- Using the REST API.

Integrity:

When files are uploaded by a producer, the repository checks the integrity of the data stream and generates a checksum (SHA-1) which is stored in the database as an internal identifier [1]. When a user asks for the data stream of a digital object, the checksum is used to retrieve the file stored in the filesystem. The repository has only the capacity to read and write the file, but not to alter it. Since a checksum is generated for each data file, the name of the file located on the filesystem is the checksum. That way, only one version of the same (equal checksums) file is really stored on the filesystem.

There are three types of digital objects:
- A workspace which contains a root collection and several snapshots (collections locked by the publishing process)
- A collection which may include metadata files and contains elements (collection or data object)
- A data object which contains a data stream (files) with metadata.

Metadata are processed the same way as the data. Clients write the metadata and send them to the repository after completion. A digital object can contain a dataset and several metadata. All changes to a digital object are recorded in the database (table event). And these events can be viewed by the users and the reviewers.

A workspace is a set of digital objects. Before publication, digital objects are not available publicly (only for members of the workspace). The user can do anything he/she wants (create, read, update, delete). No versioning exists before publication. The user can decide to publish a snapshot of his/her workspace in order to make the data available (with access control). Once a snapshot is published, it becomes immutable and cannot be changed or deleted. And a workspace knows all its snapshots.

Authenticity:

Only registered users can deposit resources to the repository. Non-academic users can create an account by providing a few information: first-name, last-name, email, username and a password. Academic users do not need to register and can login using their academic account on the national identity federation RENATER [3]. The authentification service of ORTOLANG uses RENATER to provide an easy authentication mechanism (Single Sign On [4]) to producers when they log in on the platform. Only authenticated members of a workspace can modify its content. Each data ingested into the repository is assigned to an owner (registered in the database), and each event related to a dataset can be reported with its associated owner. ORTOLANG is compliant with the GDPR [5] for user data privacy, and the user is informed during the creation of the account and on the website [6].

Versioning:

ORTOLANG fully supports versioning of the data. Everytime a producer wishes to publish a new version of a workspace, the publishing workflow creates a new version for all digital objects (if the data or metadata have changed since the last version). When the new version is accepted for publication by reviewers, all the files of the workspace gain a new PID [2] and the new version is made available online. All the previous versions are still accessible using the static versioning in PIDs, for example: https://hdl.handle.net/11403/morphalou/3.0

[1] https://www.ortolang.fr/en/help/data-integrity/
[2] https://www.ortolang.fr/en/help/persistent-identifiers/
[3] https://services.renater.fr/federation/en/index
[4] https://en.wikipedia.org/wiki/Shibboleth_Single_Sign-on_architecture
[5] https://en.wikipedia.org/wiki/General_Data_Protection_Regulation
[6] https://www.ortolang.fr/en/home/privacy-policy/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R8 Appraisal

The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance level:

The repository is in the implementation phase - 3

Response:

Users can deposit four types of resources: corpora, lexica, tools and terminology. For these resources, any type of files can be ingested into the repository. ORTOLANG provides public guidelines for data submission [1] and encourages [2] users to follow the recommendations of CLARIN which gives an overview of relevant standards in the Language Resources and Technology domain [3].

In order to publish, the users have to fill a form which contains mandatory fields like a title and a description. This information is saved in a metadata object to the repository. ORTOLANG provides a guideline to fill this form [4]. During this process, the user is guided by the user interface and by pointers to help files. The principle of the user interface tool is to provide only the information necessary (because this was necessary for compliance with previous data deposited) and to enrich the information whenever this appears to be necessary. Information about the author is always required, although technically non mandatory.

When the user asks for the publication of a workspace, a reviewer checks the completeness and understandability of the metadata. Some understandability checks may be carried out on files, but they are performed manually (see R11). The workspace contains a section for discussion groups, which allows interaction between producers and reviewers.

Request for publication can be made without filling all the metadata but only the metadata that appears to be relevant for the user. Metadata required for long-term preservation and full identification of the data can be requested during the reviewing process before publication. Missing metadata is also requested at this point and the editing process can be resumed by the user to do so.

There is no list of preferred formats but during the review process we check the consistency of the format used and suggest improvements of the data whenever necessary or more helpful for data dissemination (for example using open formats or formats that allow reuse of the data). People asking for information about the data are directed to the guidelines used in the scientific community [5].

Even if the long-term preservation is in progress, the platform can already check the formats compatible for long-term preservation and notify the producer through the web interface if files are not compliant with the supported archiving formats [6].

At the moment, reviewers may help in converting and transferring data files to the repository. But in the future, we plan to host external tools and make them directly available from the repository. Some of tools can convert data files into a preferred format, for example from an old PDF version to a newer one.

A set of internal metadata is used to describe and validate a resource (corpus, tool, etc.) according to a schema, each time the user changes it. This metadata complies with ISO-639-3 standard and use OLAC code [7] (discourse type vocabulary, linguistic data type, linguistic subject vocabulary, role). It is exported to Dublin Core, OLAC and CMDI/CLARIN (ISO-CD 24622-1) [8] formats and exposed by the OAI-PMH protocol.

[1] https://www.ortolang.fr/fr/deposer/
[2] https://www.ortolang.fr/en/help/data-formats/
[3] https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
[4] https://www.ortolang.fr/en/help/metadata/
[5] https://corli.huma-num.fr/en
[6] https://www.ortolang.fr/en/help/long-term-archiving/
[7] http://www.language-archives.org/
[8] https://www.clarin.eu/cmdi1.2-specification

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Response Text: See below
I still feel that only 3 mandatory metadata fields is very limited in the context of the re-usability of data.
A distinction between preferred and other accepted formats would ideally be combined with different levels of LTP guarantees.
It would be great if these elements would be taken forward in the coming three years before the renewal of the seal.

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

R9 Documented storage procedures

The repository applies documented processes and procedures in managing archival storage of the data.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The software and hardware architecture of ORTOLANG provide highly available storage, secure backup and disaster recovery procedures to restore software and data if case of a fatal event. Backup tapes are stored off-site on a regular basis to increase protection.

The infrastructure, hosted by INIST, uses NAS storage hardware and automatic monitoring to quickly detect failures and replace hard drives. We carried out crash simulations of the platform to test the recovery procedures of our partner INIST. The infrastructure and backups are further described in the Technical infrastructure (R15) and Security (R16) section.

ORTOLANG helps producers to publish their data by providing a workspace where they can upload data, metadata and documentations files. They can access this workspace at any time to improve their materials or deposit a new version. The data storage workflow is the following:

- The data producer deposits files.
- For each file, a SHA fingerprint is computed and linked to the file into the database.
- The SHA fingerprint is used as the access key to the file.
- Once a file is stored in the platform, it becomes immutable. All new versions of a file will generate a unique SHA fingerprint.

The integrity of data is ensured by using multiple technologies (SMART, RAID) and tools (hardware and software monitoring). All the technologies of the ORTOLANG infrastructure allow us to offer data producers a robust platform with strong security.

But the infrastructure has not been designed to provide real long-term preservation. Our close relationship with Huma-Num allow us to use the services of the CINES to do so (R10).

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R10 Preservation plan

The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance level:

The repository is in the implementation phase - 3

Response:

For the long-term preservation [1] and archival of the research data stored on the repository, we will use the services of an external partner, the CINES, which is OAIS compliant [2]. When producers wish to publish a new version of their research data, we will help them to check the compliance of their data with long-term preservation using tools provided by the CINES.

ORTOLANG leaves the producers the responsibility of complying with the compatible formats and do not undertake any conversion of data. To be compatible, a file format must be published, widely used and standardized [3]. A list of formats compatible for preservation is available [4] and should be consulted by producers before any deposit. Even if we do not enforce the use of compatible and open formats, we suggest strongly during the curation process to the users to include such formats, which they do most of the time because it is in their own interests. The quantity of data in unsuported format not accompagnied by a compatible format is actually quite small in the whole repository.

If all the files of a workspace are compliant, a Submission Information Package (SIP) is automatically created by the repository, sent to the CINES for archival and then deleted from our repository. The CINES receives the SIP, uses data curators and tools to ensure its consistency and quality.

If the package is approved, an Archival Information Package is created and stored at the CINES and we receive an Archival Resource Key (ARK) which is a unique identifier to the archived package. The ARK identifier is stored in ORTOLANG's database, and the producers are informed the archival was successful. They can use the ARK in the future to ask the CINES to retrieve the archived data.

If a package is not approved, we receive a notification and inform the producer. ORTOLANG will never modify a file format to fill the requirements. It is the responbility of the producer to update its resources to comply with the archiving formats.

The CINES preservation plan relies on migration between formats once obsolete as explained in their website.

It should be noted that, at the present time, all resources are eligible for long-term preservation if they comply with the format requirements of CINES.

[1] https://www.ortolang.fr/en/help/long-term-archiving/
[2] https://www.cines.fr/en/long-term-preservation/a-concept-problems-2/reference-model-oais/
[3] https://www.cines.fr/en/long-term-preservation/expertises/file-format/selection/
[4] https://facile.cines.fr/

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

R11 Data quality

The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality- related evaluations.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The ORTOLANG platform follows the requirements of the users of our repository, as laid down by the linguistics community in the provisions of CORLI, a CLARIN K-Centre (https://corli.huma-num.fr/en/kcentre). The reviewing process is performed by the 3 FTE of ORTOLANG, which are all specialists of the domain of linguistics, or computer linguistics.

We accept a very wide range of data, as it is impossible in any type of cutting-edge research to foresee the exact type of data that people might develop in their original work. We also accept a very broad range of data types: corpora, lexica, terminologies, tools, etc.

However, concurrently, we encourage users to deposit data in open formats [1] [2] that will be accepted for long-term preservation [3]. We encourage them during our review process to convert their data when this was not done and to provide alternative forms in open formats, like TEI, XML, Unicode text, and adequate formats for audio, video, and raw data.

The quality of the data is enhanced by the quality of the metadata, and this makes it possible, for example, to come to comply with currently accepted recommendations, like the FAIR principles. Therefore, the ORTOLANG platform follows a systematic procedure to achieve the best data and metadata quality level before publication [4] :

- The producer has to fill a form containing mandatory [5] and auto-generated fields (like the publication date)
- The data and metadata are reviewed by scientific & technical reviewers
- The reviewers analyze the data and metadata
- The reviewers can interact with the producer for inquiries or to request modifications through a private discussion channel in the workspace of the website.

The review process follows several guidelines, with no absolute criteria as some specific situation might arise that justify specific content of the data deposited. The basic principle is to have a discussion with the users and to explain the problems that might happen. When the resource and metadata do not meet the required quality, the publication can be rejected by reviewers, which makes it possible for the users to modify the data deposited and resubmit. We do not have cases of total rejection from ORTOLANG, although it could happen if severe problem were to appear (for example, plagiat, data of dubious content). This has never happened yet.

The main guidelines are to control that sufficient metadata is provided and that the formats used are interoperable, if this is possible to our knowledge. Although the minimum metadata is about three (type, title and description), we encourage the producers to fill about a dozen fields (contributors, sponsors, corpora language) [4] by asking them whether this extra information can be provided. Moreover, the license information is mandatory. The information about the authors is controled so that future redistribution of data will always be possible.

We control that the data deposited is at least readable and, whenever this is possible, is of correct format. When data is deposited with proprietary format, we always ask the user to convert (if this is technically possible) the data to an open format. We allow the user to keep the proprietary format alongside the open format. Our goal is that the data can be used by people other than the original user. Also, we encourage people to deposit documentation and publications as well as the data.

When a resource is published on the platform, a persistent identifier (Handle) is automatically assigned to it. We also automatically generate a citation with this PID to make it easier for researchers' reference needs. We strongly advise users to cite the publications where they use the resources.

Since organizations like CLARIN VLO [6] and ISIDORE [7] harvest our repository, we receive feedback about metadata issues, which helps us in constantly improving our metadata.

[1] https://www.ortolang.fr/en/help/data-formats/
[2] https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
[3] https://facile.cines.fr/
[4] https://www.ortolang.fr/en/help/publishing-workflow/
[5] https://www.ortolang.fr/en/help/metadata/
[6] https://vlo.clarin.eu/about
[7] https://isidore.science/about

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R12 Workflows

Archiving takes place according to defined workflows from ingest to dissemination.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The ORTOLANG platform includes a publication workflow [1].

When a user wants to publish a resource on the platform, the workflow includes the following steps:

- Log in to the platform.
- Upload the resource data.
- Fill a metadata form containing mandatory fields, like description and license. The user can find help in the user guide [2].
- Set the visibility of folders/files.
- Request publication of the resource.

When a publication request is sent, the workflow automatically notifies reviewers. At the present time, there are three reviewers : one for checking the metadata fields and two for checking and talking about the data format and commenting or improving the metadata from a scientific point of view. Tasks are presented to all reviewers and the reviewers assign tasks to themself according to their schedule and competence. The first step is to check the description is understandable and if the metadata could be more detailed.

Of course, we don't force anyone to fill out all the metadata fields because it is time consuming and it cannot be always applied. The second step consists in checking the content and nature of the data. No automatic format control is done because in most cases automatic analysis is difficult. Manual checking is done randomly on large datasets.

The reviewers use the platform tools to check the metadata and data of the resource. The time necessary for the reviewing process is variable, according to the size of the data and the nature of the files. Once the reviewers have completed their work, the final step is to decide whether the resource is published or not. Interaction with the producer is possible at any moment using a private messaging system. The producer receives notifications when:

- He/she receives a new message from reviewers.
- The publication has been accepted/rejected. When a publication request is rejected, the producer receives a message from the reviewer explaining the reason and asking the producer to fix and resubmit.

In addition, the user can make changes to metadata/data in his/her workspace at any time and submit a new version, which is a snapshot of his/her workspace. New PIDs are provided to the resource and all new/changed files. The platform maintains one PID which links to the last version and produces as many PID as there are versions of the resource.

The repository also provides a REST API [3] for advanced features like listing all files in a folder, copying a local folder to the remote repository, searching file metadata, etc.

[1] https://www.ortolang.fr/en/help/publishing-workflow/
[2] https://www.ortolang.fr/en/help/metadata/
[3] https://www.ortolang.fr/en/rest-api/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R13 Data discovery and identification

The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The ORTOLANG platform offers extensive browse and search features.

The platform website (https://www.ortolang.fr) shows all the published resources broken down into four categories (Corpora, Lexica, Terminologies and Tools).

Users can use advanced search filtering resources by type, language or license status, reflected in in facets. We provide metadata for OAI-PMH harvesting in different formats:

- Dublin Core
- OLAC
- CMDI

Our platform is currently harvested by:

- The CLARIN VLO (https://vlo.clarin.eu)
- ISIDORE (a French humanities search engine) (https://www.rechercheisidore.fr).

We are currently working to add OpenAIRE support.

ORTOLANG uses Handle persistent identifiers to generate unique URLs for each deposited file and resource. Each published resource has its own resource page (eg. https://www.ortolang.fr/market/corpora/scienceshumaines) that, among other things, contains information for data citations (in text and bibtex formats).

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R14 Data reuse

The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository uses a user-friendly interface to help users describe easily as much metadata information as possible (but not all metadata field are mandatory).

During the review process, we suggest producers who are about to submit their dataset for publication to use open formats officially validated by the CINES, or to convert it from a proprietary to an open format and to deposit both versions. We encourage users to use open formats [1].

During this process, producers are also invited to deposit a description and a free sample of the data. Our website makes compliance with this part of the deposition process easy. All data, including samples and descriptions, is included in the snapshot corresponding to the data published. As a result, a description will always be attached to the data itself.

Finally, for all data deposits, a reference to the laboratory responsible for it is required. This will promote better long-term dialog with producers.

The format and the details of the metadata is not shown to the user. We add as much information possible in the website so as to guide the user, but not overwhelm them with too much information. As soon as the metadata is edited by the user, the metadata is included in the system [2] and provides all information necessary to authorize and enable the future use of the dataset by providing clear open-license and mime-type information in this respect.

We have implemented our own metadata format (JSON) that is validated by a schema before being saved in the repository. It is extensible so that new needs can be added. The compatibility with all the versions is preserved.

The linguistics community uses the OLAC metadata format which is generated by the platform for each resource based on our own metadata format (JSON). CMDI is another metadata format which has been in use more recently in the community. It is also computed from our own metadata format.

[1] https://www.ortolang.fr/en/help/data-formats/
[2] https://www.ortolang.fr/en/help/metadata/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Technology

R15 Technical infrastructure

The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

Software Architecture:

The platform developed for ORTOLANG was created from scratch by our engineers [1]. Our software relies on many community-supported components:

- PostgreSQL - WildFly
- KeyCloak
- Angular JS
- ElasticSearch - Docker Swarm

We continually improve our software infrastructure by refactoring, improving reliability, fixing bugs and adding new functionalities (e.g.: rewriting the metadata indexing by using ElasticSearch, package linguistic tools to process data online, etc.). All the code developed for the platform is stored in git repositories. We use Gitlab to manage all our projects. The code is available on our server.

The ORTOLANG platform was implemented through a close collaboration between the ATILF development team and the INIST production team. During the development phase of the project (2013-2017), many releases were made (on a monthly basis). Since the exploitation phase, updates have been regularly carried out to either add new functionalities, to replace or update a component of the application or to upgrade the middleware (e.g. Ubuntu 16.04 to 18.04, Docker deployment, etc.).

There are three distinct environments:

- Development
- Pre-production
- Production

Software installation and qualification is performed using different deployment solutions such as Puppet, Ansible or Docker.

Hardware Architecture:

Our infrastructure is hosted at INIST.

The ORTOLANG architecture [2] is made up of 6 dedicated Dell R630 and R640 physical servers running Ubuntu. The maintenance of these servers has already been secured until 2025 (R630) and 2027 (R640).

These servers are aggregated using the VmWare hypervisor to provide a dedicated cluster. In addition to the dynamic allocation of resources (CPU, RAM), VmWare's functionalities make it possible to transfer the virtual machines (VMs) constituting the ORTOLANG application from one physical server to another and secure continuity of service if a physical server crashes, or if it is shut down for maintenance.

All the platform's equipment (server, storage, backup) is interconnected by very high-bandwith switches (8 and 16 Gbits/s) with redundant physical paths.

The connection to the RENATER National Research Network is via a 1Gbits/s link that can be easily upgraded if necessary. A dual active/passive high availability firewall is used to filter access. The ORTOLANG servers are not accessible directly from the Internet but through a proxy.

The supervision of the whole system relies on the open-source software Shinken and on vCenter for the whole VmWare part (dynamic resource allocation, stop/start/move VMs). The support (maintenance) VmWare/vCenter has already been secured until 2026.

Infrastructure Roadmap:

The ORTOLANG infrastructure is currently in its third version:

- A 2013 version was used during the development phase (2013-2016)
- A 2016 version for the initial phase of production (2016-2019)
- A 2020 version to cover the period 2020-2025. This version was acquired and implemented in late 2019 (new Dell R640 servers, new Brocade B6505 switches, new Quantum DXi 4800 backup system, new Dell SC 5020 storage disk array).

ORTOLANG also relies on the INIST pooled hardware resources which are updated on a yearly basis (firewall replaced in 2019, backup server replaced in 2019, etc.). All INIST systems are under a enterprise-type support scheme (maintenance operation within 4 hours).

[1] https://www.ortolang.fr/en/help/sotfware-architecture/
[2] https://www.ortolang.fr/en/help/hardware-architecture/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R16 Security

The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The whole ORTOLANG infrastructure is hosted at INIST, which also houses other large-scale national infrastructures. A dedicated team of system engineers is in charge of its IT system security and they use software like Shinken for failure monitoring and notification [1].

Our repository software architecture relies on multiple VMWare virtual machines. These machines are provisioned using Puppet models and can be rebuilt from scratch in case of a failure. The repository data is backed up daily on LTO drives.

Building Safety

There are 2 server rooms of 100 square meters each in INIST's premises. This two-room configuration makes it possible to distribute the different equipment (physical servers, switch), so the ORTOLANG platform hardware is distributed and interconnected between the 2 rooms. The ORTOLANG infrastructure can operate if one of the 2 rooms is no longer available.

The power supply is provided by 2 inverters operating redundantly and connected to a coupling cabinet. A single inverter is capable of providing the entire power supply to INIST site. Batteries are capable of providing power for 20 minutes in the event of a power cut. In addition, a diesel generator set starts

automatically within a few minutes and provides the auxiliary power source with an autonomy of several days in case of a power outage.

There are 2 air conditioners in each room, 1 air conditioner being enough to cool each room. The production of chilled water is provided by 2 chillers, located in technical premises outside the computer building. A single unit can produce enough chilled water needed to cool the computer rooms and the inverter room.

Fire Safety

A central detection and automatic neutral-gas extinguishing system covers fire safety in addition to the machine-room layout, which is designed to provide fire protection for more than one hour. The four rooms are monitored separately (2 machine rooms, inverter, console) with multiple sensors. The extinguishing gas is of the ARGO 55 type, a natural inert gas composed of 50% nitrogen and 50% argon. CO2-type extinguishers were also installed in the rooms.

Fire drills are carried out regularly, several times a year.

Intrusion Detection

Access controls are systematic for entering INIST campus buildings.
In addition, there are reinforced access checks to the server rooms (restricted to INIST IT technical staff and senior management).

A security officer is permanently present on site outside working hours. Unauthorized entries, technical and fire alarms are centralized and transmitted to this officer, who is consequently able to react very promptly. A BMS (Building Management System) also makes it possible to monitor the various operating situations.

Data security

Three backup solutions are currently being implemented:

- Tape backups with HP Dataprotector software coupled with a dual robotic Quantum Scalar I6000 (LTO7) and Scalar i80 (LTO6);
- Disk backups with Veeam Backup software;
- Snapshots (instant captures of stored data) on the primary storage arrays.

For Linux application environments, like ORTOLANG, 2 types of backup (Dataprotector and Veeam) are performed following the same principle in each case: one weekly full backup and one incremental backup on the other days with a 6-week data storage retention.
Veeam backups are supported by Quantum (DXi) appliances, which allows on-the-fly compression and deduplication.

All data is hosted at INIST in both server rooms. Offline archives (tapes) can be created for storage in fireproof vaults, one of which is located in a separate building. Hosting on a remote site of the duplicated data is one possible way of future improvement.

Crash tests

We tried out two crash-test strategies successfully before service launch:

- Restoring an entire ORTOLANG infrastructure from backups on empty physical servers.
- Restoring an entire ORTOLANG infrastructure by rebuilding the entire application using deployment tools.

We are currently in talk with the Universite de Lorraine to use their infrastructure as a second backup site for all the data available on ORTOLANG. The volume of the data and the security requirements present a few issues we need to solve in the coming months.

A continuous improvement policy

INIST has been committed to an improvement effort for many years, but has not gone as far as certifying its data centre yet. It has called upon specialized companies for this effort. We can mention the following:

- A study for a disaster recovery plan (PRAI) by the Ares firm in 2008. - A safety audit by Ernst & Young in 2018.

[1] https://www.ortolang.fr/en/help/security/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: A back-up at a different location would be an important next step

Applicant Feedback

R17 Applicant Feedback

We welcome feedback on the CoreTrustSeal Requirements and the Certification procedure.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

-

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Board comment: For recertification in 3 years the board expects that the Requirements 1, 3, 8 and 10 will improve to compliance level 4.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Board comment: For recertification in 3 years the board expects that the Requirements 1, 3, 8 and 10 will improve to compliance level 4.