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

Un équipement d'excellence de mutualisation de ressources et d'outils sur la langue et son traitement informatique

ORTOLANG bénéficie d'une aide de l'État au titre du programme « Investissements d'avenir » (ANR-11-EQPX-0032)

ORTOLANG un équipement d'excellence en réseau réunissant des compétences complémentaires

Positionnement institutionnel

- Un **équipement d'excellence** au service de l'ensemble de la communauté
  - en réseau, intégrant entre autres le CNRTL et le SLDR
  - géré par le CNRS,
  - financement assuré jusque fin 2019
- **Service spécialisé pour la langue**
  - complémentaire de l'offre généraliste proposée par Huma-Num
- **Une mission complémentaire de développement de ressources et d'outils de base**

Pourquoi un tel équipement ?

- Pour servir de **support aux travaux de recherche** :
  - La notion de corpus (annotés ou non) est aujourd'hui incontournable spécifiquement en linguistique et en TAL pour l'émergence et l'apprentissage de modèles
    - pour les approches stochastiques
    - pour les approches symboliques
- Pour la **valorisation des résultats de recherche** (corpus, lexique, dictionnaires et outils de traitement)
- La constitution et la normalisation de corpus de qualité est très couteuse
  - On ne peut pas se permettre de recréer des ressources qui existe déjà
  - Le passage à l'échelle supérieure impose de rassembler des corpus d'origines différentes

Etat actuel du projet

- **Une plateforme matérielle** s'appuyant sur
  - 6 serveurs bi-processeur avec 256/384 Go de RAM chaque.
  - 165 To utile de disques en Raid 6
  - 2 Librairies LTO6 et 50 slots de 300To
  - L'ensemble est hébergé à l'INIST avec un service 24h/24 et 7j/7
- **Une plateforme logicielle** comprenant
  - Un dépôt d'objet numérique fiable et robuste opensource
  - Une sécurisation des données
  - Une gestion des versions
  - Une authentification des utilisateurs permettant l'accès par serveur d'identité (Renater)
  - Une interface de dépôt et de consultation
  - Une création et diffusion de métadonnées

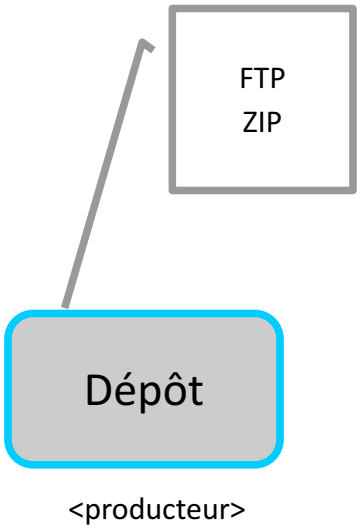
Contexte général

- Les utilisateurs émettent le souhait de **déposer une nouvelle ressource**
- **Ouverture d'un espace dans Ortolang**
- **Dépôt privé** d'**un ensemble de données** (fichiers) et de métadonnées. Ces données peuvent être organisées et formatées comme l'utilisateur le désire
  - Des discussions sur les meilleurs formats et présentations peuvent être ouvertes avec l'équipe Ortolang et la communauté
- Les données prêtes sont **publiées** avec un contrôle (formats, métadonnées, information) des administrateurs de la plateforme
- Les données sont **diffusées** en accord **avec les droits accordés par le déposant**. Des identifiants **pérennes** sont créés pour accéder aux données (qui ne changeront plus). Les données sont ouvertes à la consultation et au téléchargement.
- Les producteurs peuvent **compléter** leur travail et publier une **nouvelle version** de la même ressource.
- Les producteurs peuvent également soumettre leurs ressources pour un **archivage à long terme**.

Accompagnement pour toutes les phases de traitements des données

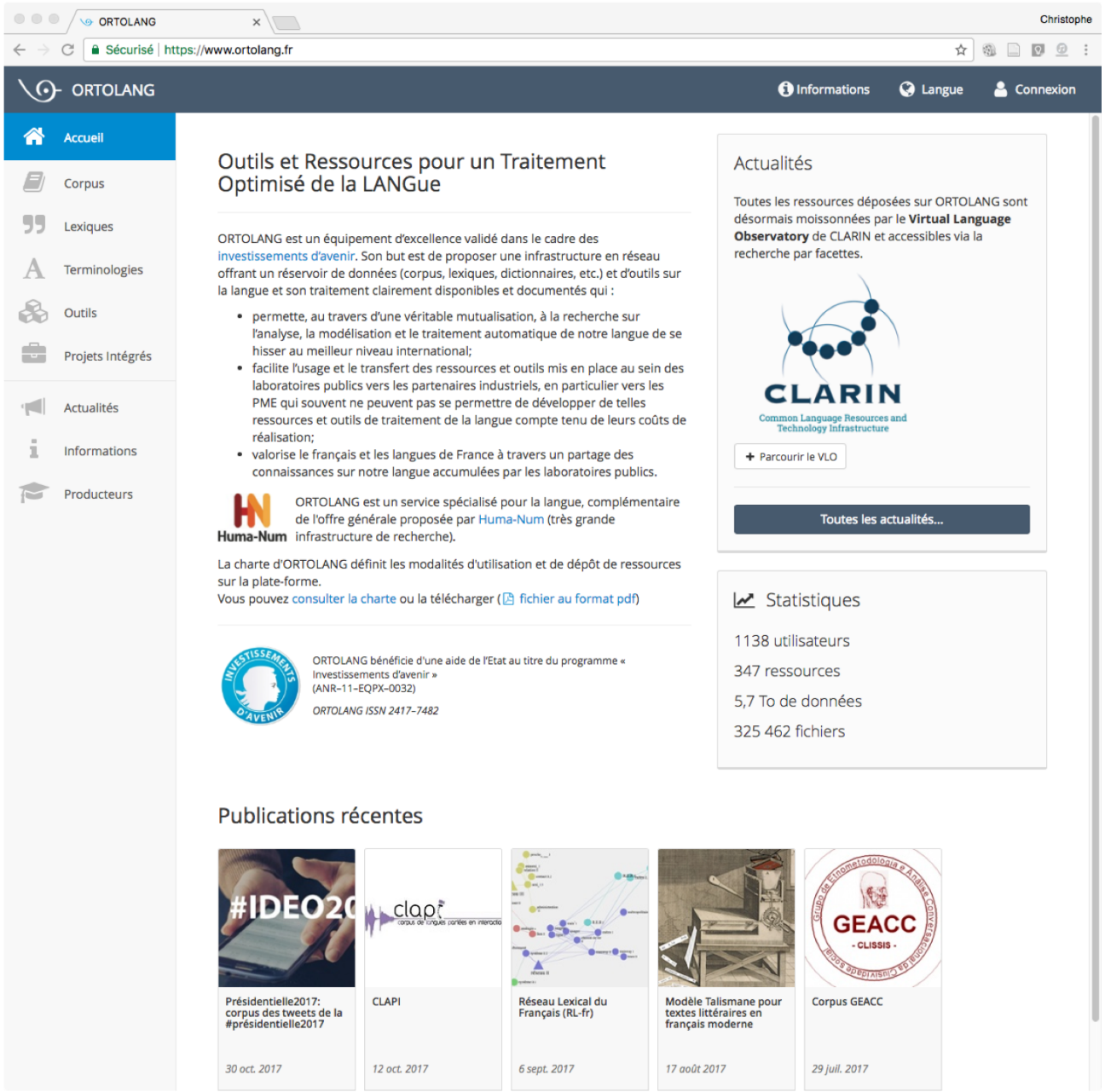
- Trois Centres de compétences thématiques
  - **Oral** : (SLDR & Modyco)
  - **Multi-modal** (SLDR & Modyco)
  - **Ecrit** (ATILF/CNRTL)
- Interactions directes et personnelles avec les membres de ces centres thématiques
  - pour le SLDR : resp. Gilles Pouchoulin
  - pour MoDyCo : resp. Christophe Parisse
  - Pour le CNRTL : resp. Etienne Petitjean, Bertrand Gaiffe
- Des informations de suivi régulier transmises aux déposants

Le dépôt

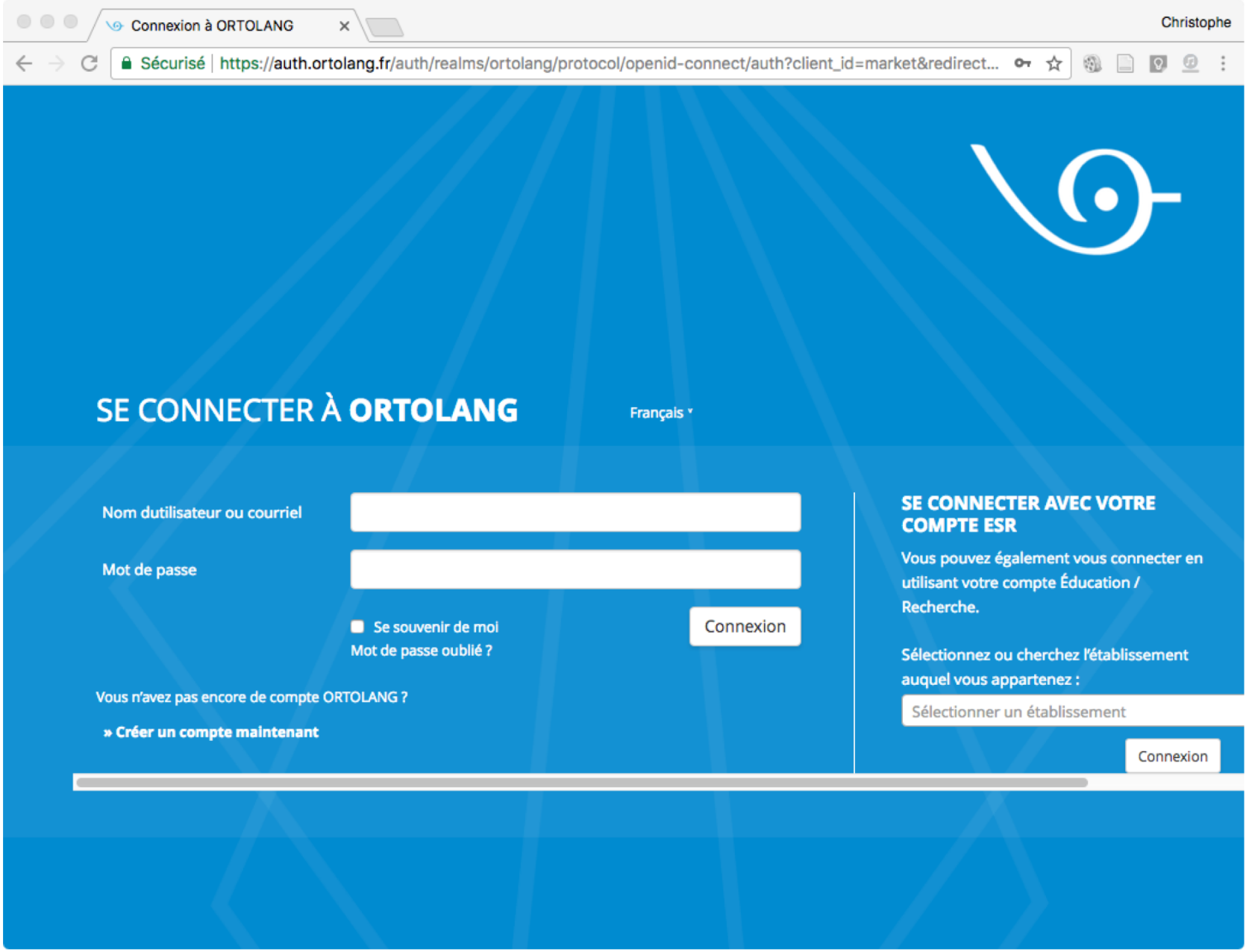


- Le producteur dispose d'un moyen **simple** de **dépôt de ses données** même si ces données sont encore impropres à la publication.
- Différentes méthodes sont proposées pour le dépôt ou téléchargement :
  - via **FTP**
  - via une **interface web**
  - par téléchargement de **fichiers compressés**
- Aussitôt déposées, les ressources sont **sécurisées** par l'utilisation de supports fiables (redondance) et par la réalisation de sauvegardes quotidiennes sur bande.

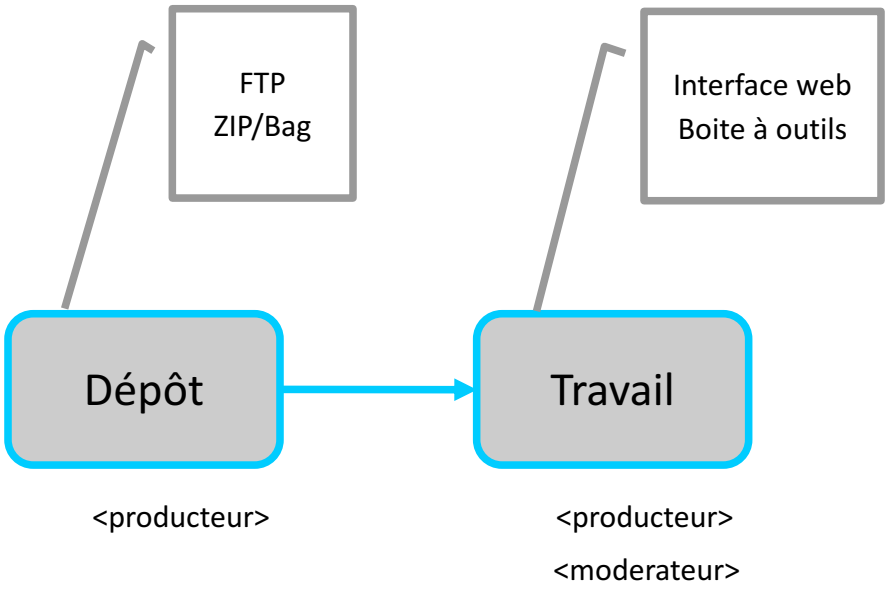
Accès


ORTOLANG
Christophe
Sécurisé | https://www.ortolang.fr
ORTOLANG
Informations
Langue
Connexion
Accueil
Corpus
Lexiques
Terminologies
Outils
Projets Intégrés
Actualités
Informations
Producteurs
Outils et Ressources pour un Traitement Optimisé de la LANGue
ORTOLANG est un équipement d'excellence validé dans le cadre des investissements d'avenir. Son but est de proposer une infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement clairement disponibles et documentés qui :
permette, au travers d'une véritable mutualisation, à la recherche sur l'analyse, la modélisation et le traitement automatique de notre langue de se hisser au meilleur niveau international;
facilite l'usage et le transfert des ressources et outils mis en place au sein des laboratoires publics vers les partenaires industriels, en particulier vers les PME qui souvent ne peuvent pas se permettre de développer de telles ressources et outils de traitement de la langue compte tenu de leurs coûts de réalisation;
valorise le français et les langues de France à travers un partage des connaissances sur notre langue accumulées par les laboratoires publics.
Huma-Num
ORTOLANG est un service spécialisé pour la langue, complémentaire de l'offre générale proposée par Huma-Num (très grande infrastructure de recherche).
La charte d'ORTOLANG définit les modalités d'utilisation et de dépôt de ressources sur la plate-forme.
Vous pouvez consulter la charte ou la télécharger (fichier au format pdf)
ORTOLANG bénéficie d'une aide de l'Etat au titre du programme « Investissements d'avenir » (ANR-11-EQPX-0032)
ORTOLANG ISSN 2417-7482
Actualités
Toutes les ressources déposées sur ORTOLANG sont désormais moissonnées par le Virtual Language Observatory de CLARIN et accessibles via la recherche par facettes.
CLARIN
Common Language Resources and Technology Infrastructure
+ Parcourir le VLO
Toutes les actualités...
Statistiques
1138 utilisateurs
347 ressources
5,7 To de données
325 462 fichiers
Publications récentes
Présidentielle2017: corpus des tweets de la #présidentielle2017
30 oct. 2017
CLAPI
12 oct. 2017
Réseau Lexical du Français (RL-fr)
6 sept. 2017
Modèle Talismane pour textes littéraires en français moderne
17 août 2017
Corpus GEACC
29 juil. 2017

Identification

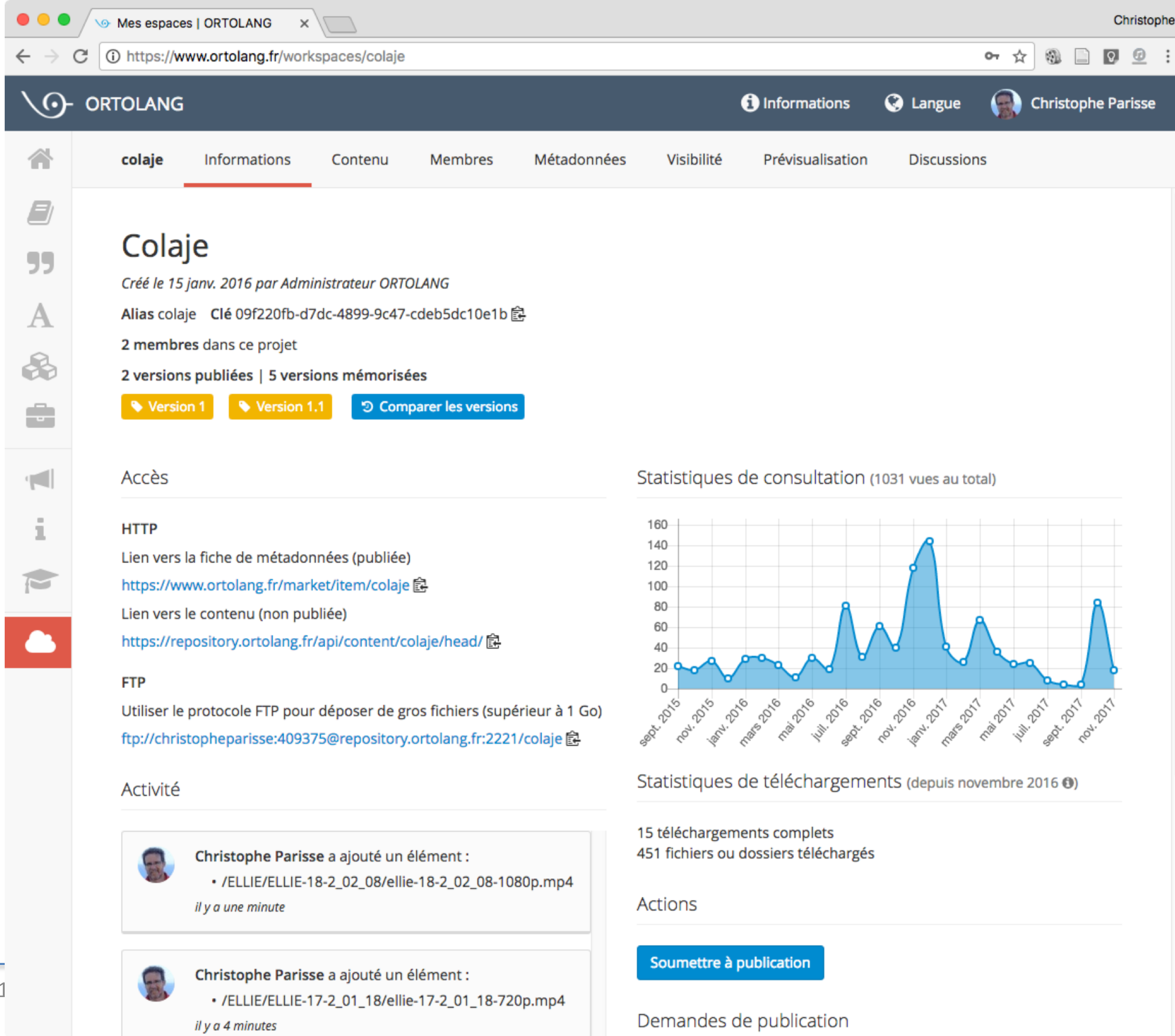


Espace de travail sécurisé

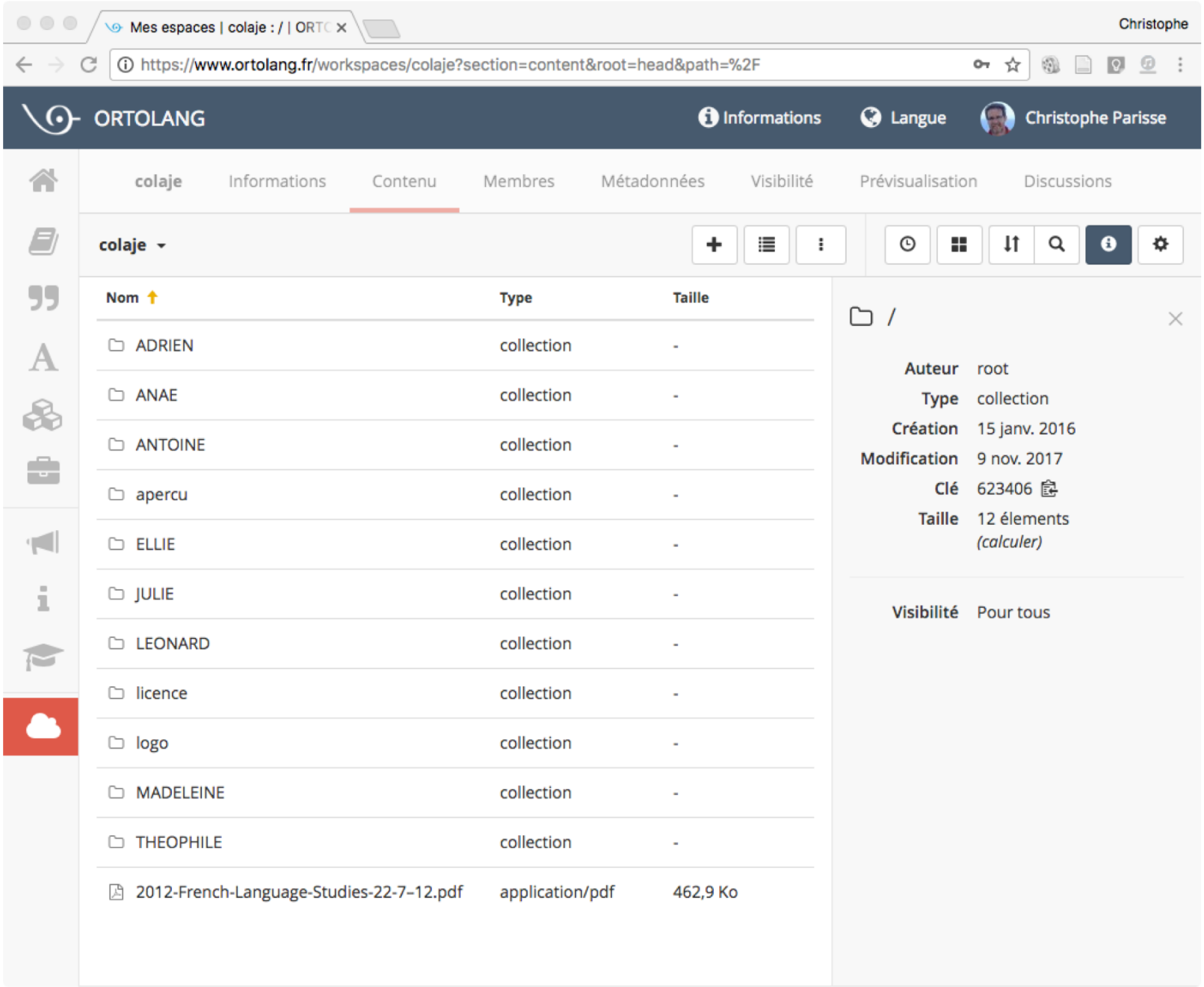


- Via un espace de travail en ligne, un partage de fichiers et une API, le producteur peut **travailler** sur ses données.
- Cet **espace de travail est sécurisé** (sauvegarde journalière)
- L'**accès** aux données est **contrôlé** et elles ne sont visibles que par les membres de l'espace de travail et les administrateurs de la plate forme
- Une mise en page HTML des données est possible.
- *Il bénéficie en outre d'outils en ligne spécifiques qui vont lui permettre d'enrichir son travail (alignement, annotation, …)*

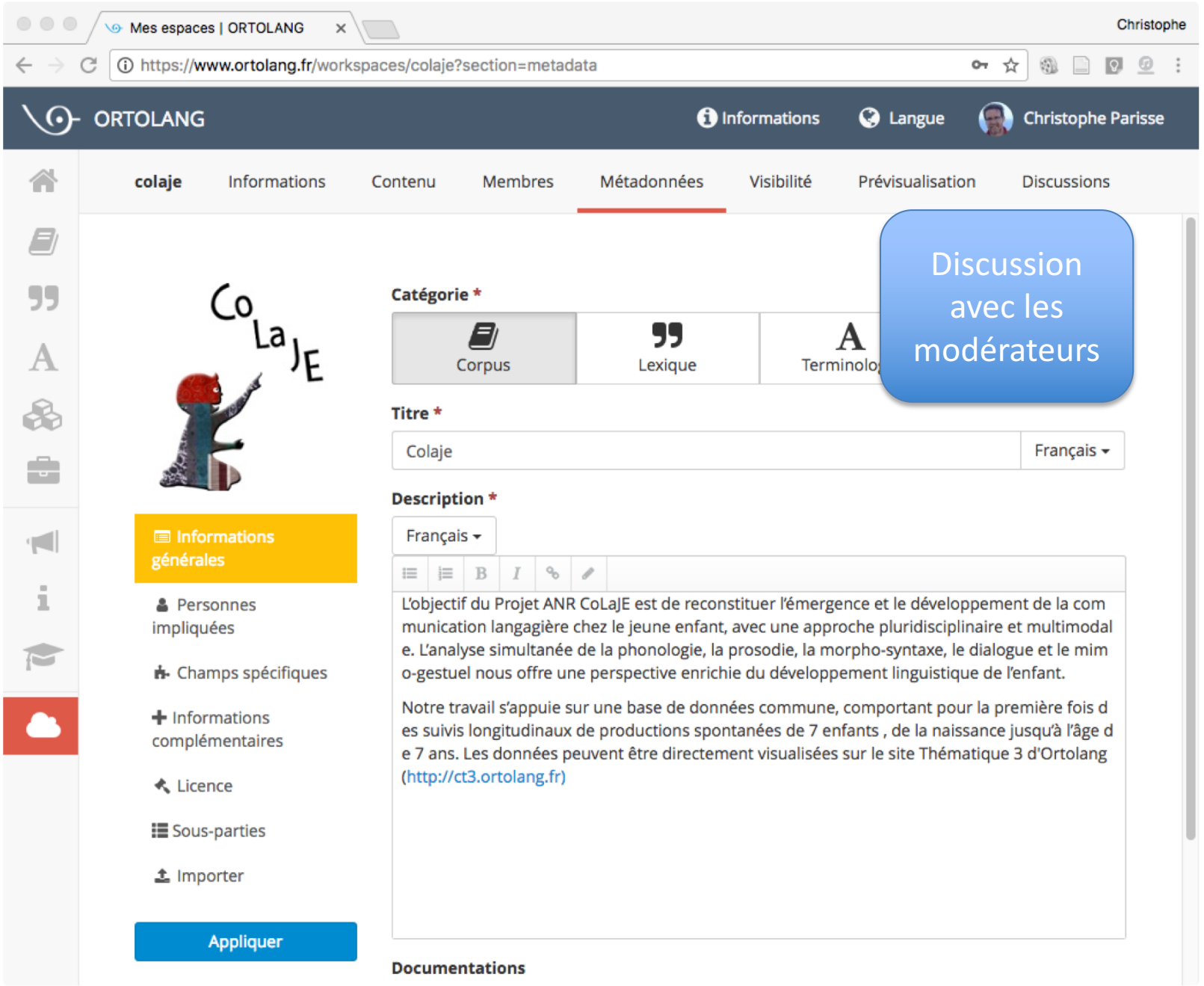
Général


Mes espaces | ORTOLANG
Christophe
https://www.ortolang.fr/workspaces/colaje
ORTOLANG
Informations
Langue
Christophe Parisse
colaje
Informations
Contenu
Membres
Métadonnées
Visibilité
Prévisualisation
Discussions
Colaje
Créé le 15 janv. 2016 par Administrateur ORTOLANG
Alias colaje Clé 09f220fb-d7dc-4899-9c47-cdeb5dc10e1b
2 membres dans ce projet
2 versions publiées | 5 versions mémorisées
Version 1
Version 1.1
Comparer les versions
Accès
HTTP
Lien vers la fiche de métadonnées (publiée)
https://www.ortolang.fr/market/item/colaje
Lien vers le contenu (non publiée)
https://repository.ortolang.fr/api/content/colaje/head/
FTP
Utiliser le protocole FTP pour déposer de gros fichiers (supérieur à 1 Go)
ftp://christopheparisse:409375@repository.ortolang.fr:2221/colaje
Activité
Christophe Parisse a ajouté un élément :
• /ELLIE/ELLIE-18-2_02_08/ellie-18-2_02_08-1080p.mp4
il y a une minute
Christophe Parisse a ajouté un élément :
• /ELLIE/ELLIE-17-2_01_18/ellie-17-2_01_18-720p.mp4
il y a 4 minutes
Statistiques de consultation (1031 vues au total)
160
140
120
100
80
60
40
20
0
sept. 2015
nov. 2015
janv. 2016
mars 2016
mai 2016
juil. 2016
sept. 2016
nov. 2016
janv. 2017
mars 2017
mai 2017
juil. 2017
sept. 2017
nov. 2017
Statistiques de téléchargements (depuis novembre 2016)
15 téléchargements complets
451 fichiers ou dossiers téléchargés
Actions
Soumettre à publication
Demandes de publication

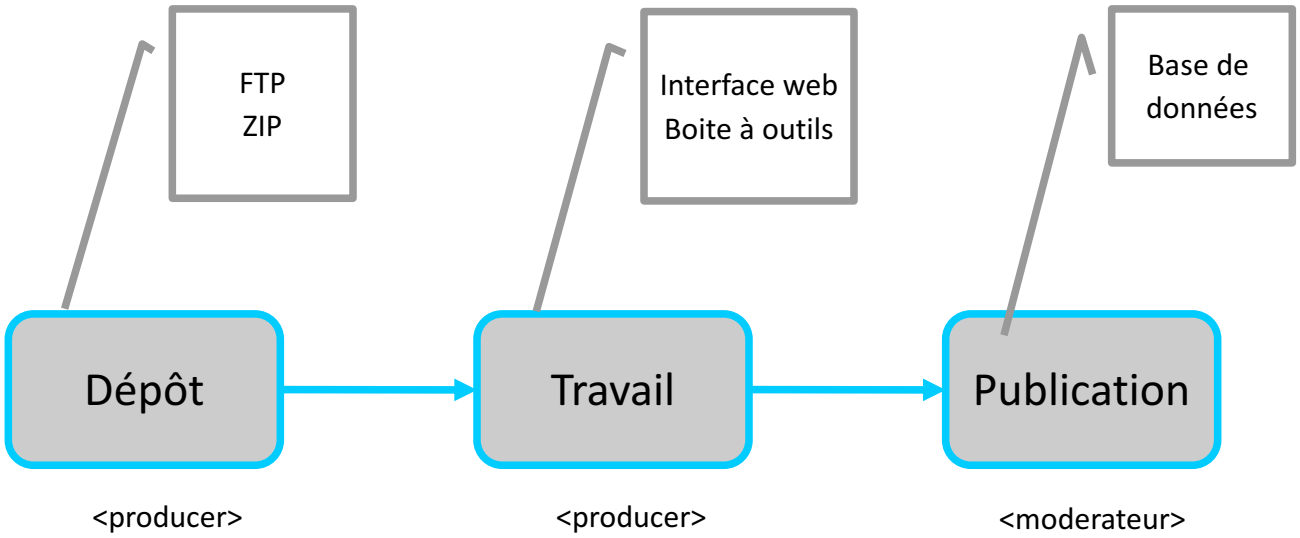
Données


Mes espaces | colaje : / | ORTO
Christophe
https://www.ortolang.fr/workspaces/colaje?section=content&root=head&path=%2F
ORTOLANG
Informations
Langue
Christophe Parisse
colaje
Informations
Contenu
Membres
Métadonnées
Visibilité
Prévisualisation
Discussions
colaje
Nom
Type
Taille
ADRIEN collection -
ANAE collection -
ANTOINE collection -
apercu collection -
ELLIE collection -
JULIE collection -
LEONARD collection -
licence collection -
logo collection -
MADELEINE collection -
THEOPHILE collection -
2012-French-Language-Studies-22-7-12.pdf application/pdf 462,9 Ko
/
Auteur root
Type collection
Création 15 janv. 2016
Modification 9 nov. 2017
Clé 623406
Taille 12 élements
(calculer)
Visibilité Pour tous

Métadonnées

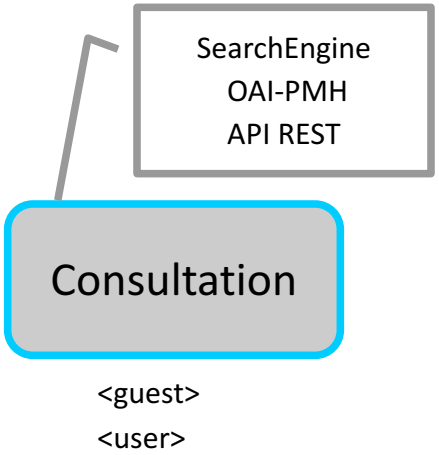


Publication



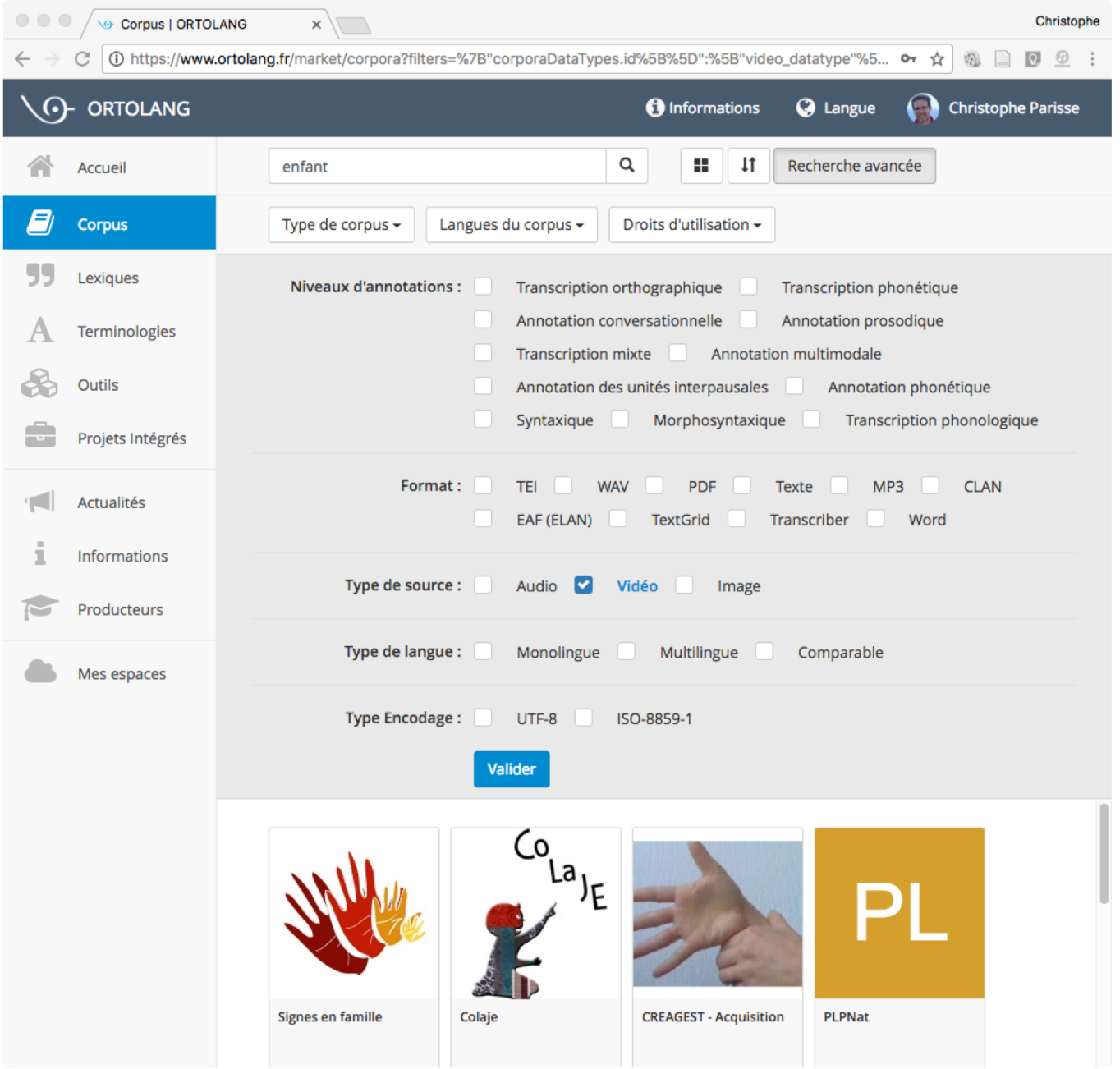
- Une fois que les données sont prêtes, le producteur peut soumettre son travail pour publication.
- Il peut alors suivre l'état de ses demandes et, en collaboration avec les administrateurs, aboutir à une version stable de sa ressource.
- Plusieurs formes de publications sont possibles (tous, identifié, recherche, privé) – chaque élément de la ressource peut avoir des droits différents

La consultation

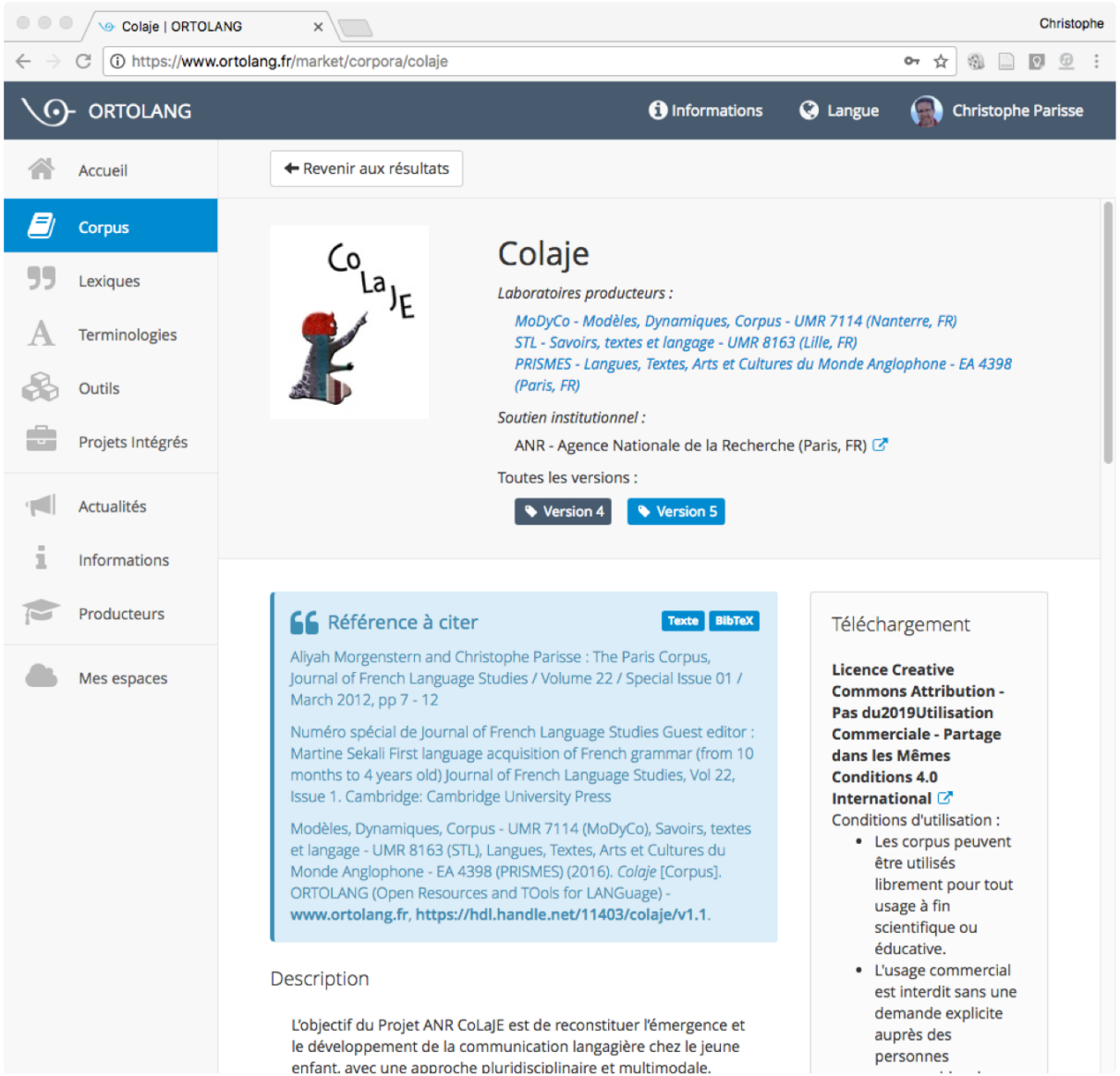


- La **consultation** des données peut se faire de plusieurs manières :
  - via une **interface web** qui présente toutes les ressources hébergées organisées par catégories et décrites par une fiche détaillée.
  - une **navigation** dans le contenu des ressources est également possible en ligne
  - un **moteur d'indexation**
  - un connecteur **OAI-PMH**
    - **ISIDORE, VLO CLARIN**
  - une interface **REST**
- Un téléchargement aisé **de la totalité ou de parties des corpus**
  - Filtrage par noms de fichiers
  - Téléchargement de certains répertoires

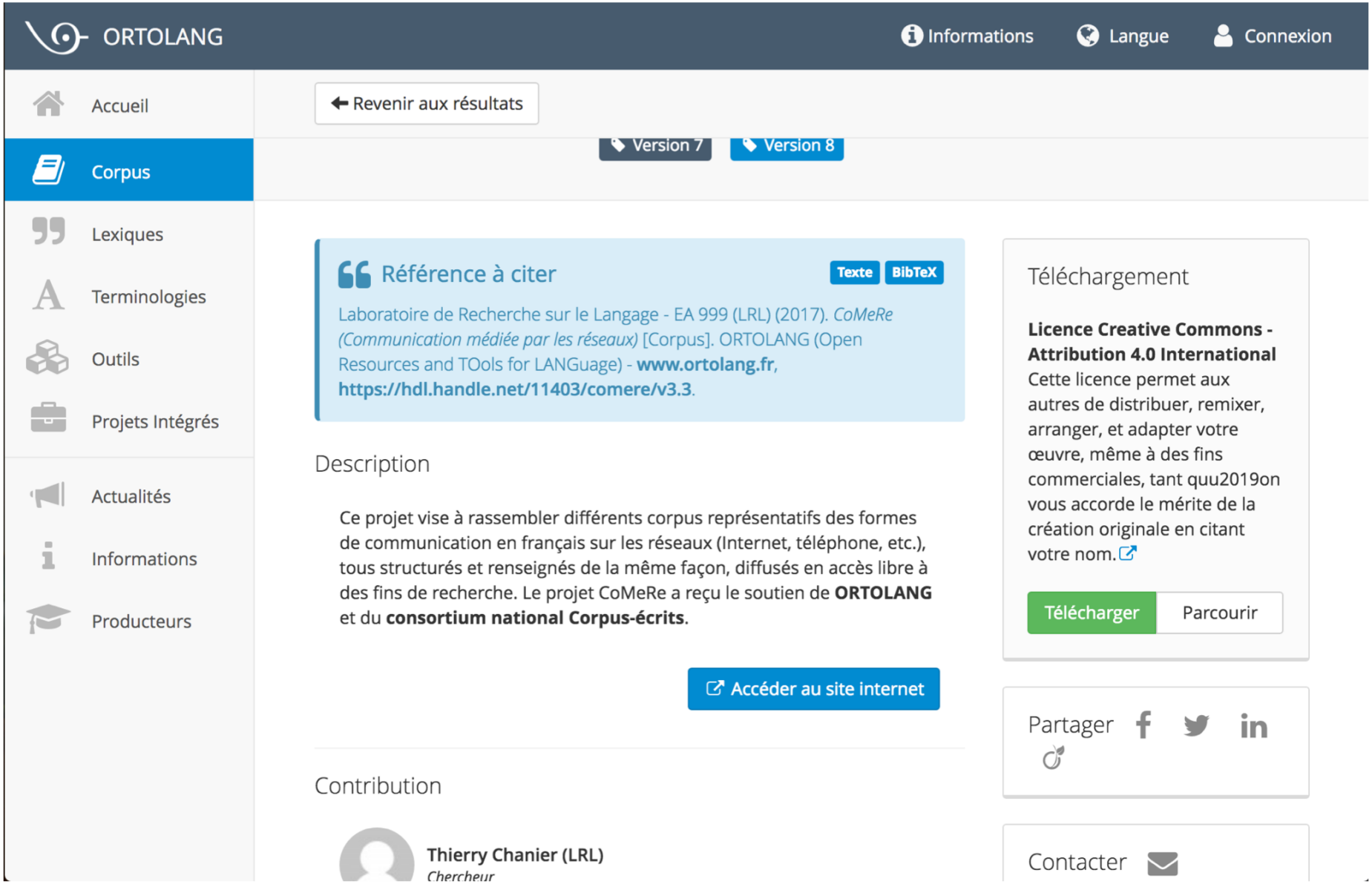
Interrogation



Référencement / accès



Référencement / accès



Présentation HTML

https://repository.ortolang.fr/api/content/comere/v3.3/comere.html

ORTOLANG

Open Resources and TOols for LANGuage

This site: https://hdl.handle.net/11403/comere

CoMeRe Repository: Corpora of Computer-Mediated Communication in French

This repository includes corpora of mono or multimodal interactions mediated through networks (Internet, Phone, etc.). Three fundamental principles underlie CoMeRe: variety, standards, openness.

- Variety: interactions stemming from networks, as well as mono and multimodal, synchronous and asynchronous communications, eg: Email, forums, textchat, Tweets, SMS, Wiki discussions, video-audio conference systems, etc.
- Standards: corpora are structured and referred to in a uniform way, i.e. in TEI - Text Enconding Initiative- with a specific extension dedicated to CMC and elaborated within the European TEI-SIGand openness.
- Openness: all corpora are open and free access. They can be fully downloaded. CoMeRe follows the OpenData recommandations for research data.

Reference : Chanier,T., Poudat,C., Sagot, B., Antoniadis, G., Wigham,C. R. , Hriba, L.,Longhi, J. & Seddah, D. (2014) « The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres ». Special issue on « Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics ». *JLCL (Journal of Language Technology and Computational Linguistics)*. pp1-31. http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf

CMC genres

SMS	Tweets	Email	Text chat	Multimodal
- cmr-smslareunion	- cmr-polititweets	- cmr-simuligne	- cmr-getalp_org	- cmr-copeas
- cmr-smsalpes	- cmr-intermittent		- cmr-favi	- cmr-tridem06
- cmr-88milsms		**Discussion forum**	- cmr-favi (POS tagged)	**Multimodal + 3D**
	Weblog	- cmr-simuligne	- cmr-simuligne	- cmr-archi21

Accès au site www.ortolang.fr

- Ortolang est totalement fonctionnel pour la partie dépôt et accès aux corpus
- Ortolang comprend des corpus et des outils issus des données des précédents centres (CNRTL, SLDR) et des partenaires (Modyco, LLL)
- Ortolang comprend de nombreux corpus ajoutés depuis l'ouverture et également grâce aux appels corpus communs à Ortolang, Corpus Ecrits, IRCOM et CORLI.

Pour en savoir plus
et pour utiliser Ortolang

www.ortolang.fr

contact@ortolang.fr