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

Une infrastructure de mise à disposition et de partage de ressources linguistiques

Jean-Marie.Pierrel@atilf.fr

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

Positionnement institutionnel

- Un équipement d'excellence au service de l'ensemble de la communauté s'appuyant sur un réseau de compétences
  - 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 complémentaires de développement de ressources et d'outils de base pour le français

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

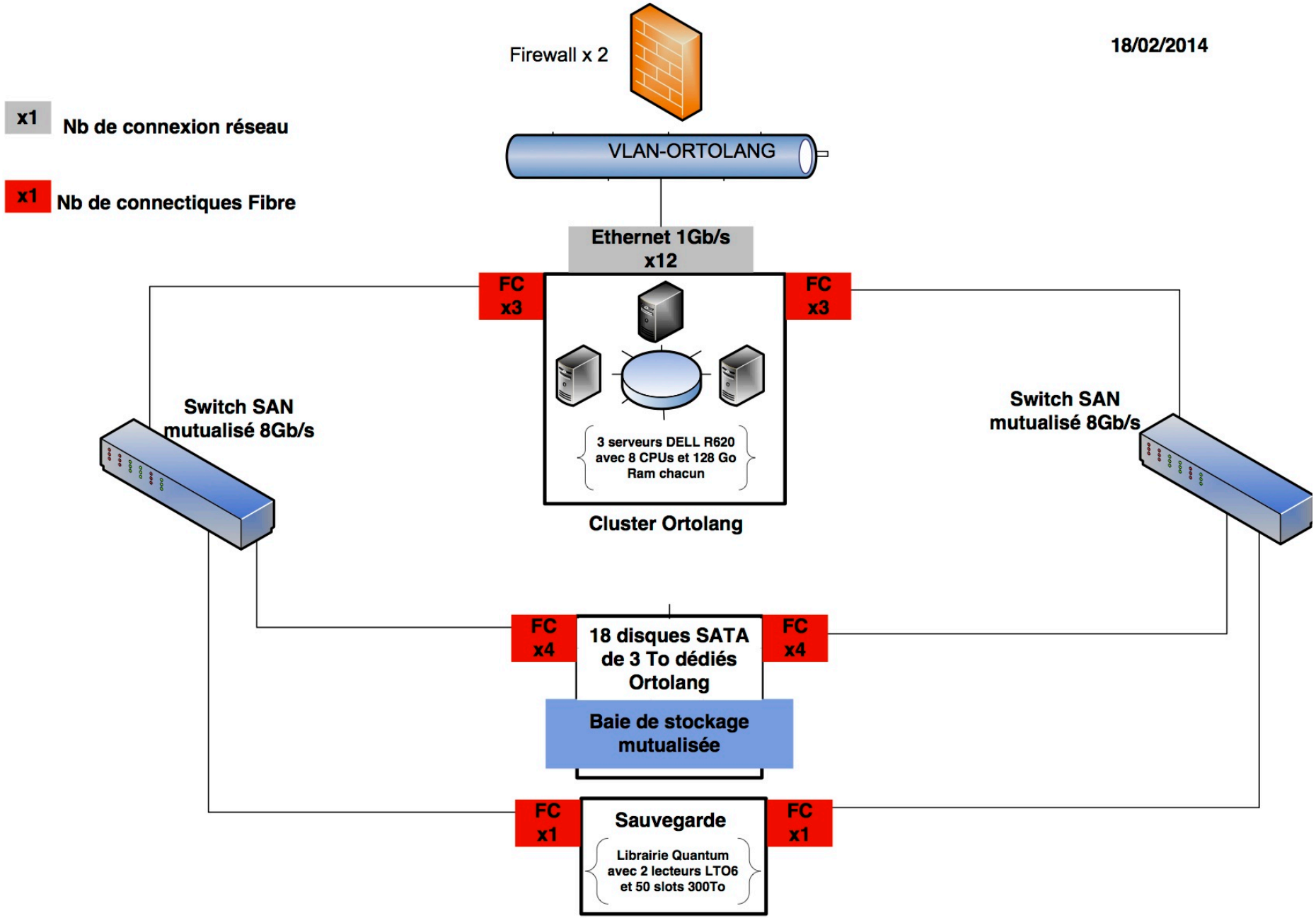
Pourquoi un tel équipement ?

- 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)
- Or la constitution et la normalisation de corpus de qualité est très couteuse et sans une véritable mutualisation chaque équipe de recherche se verrait dans l'obligation de tout réinventer

La plateforme ORTOLANG

- Une plateforme matérielle et logicielle s'appuyant sur
  - 3 serveurs bi-processeur avec 128 Go de RAM chaque.
  - 40 To utiles de disques en Raid 6
  - 2 Librairies LTO6 de 125 To
    - capacité totale de sauvegarde (312,5 To compressés)
- Un upgrade de cette plateforme prévu en 2016 et en 2019
  - Financements assurés dans le cadre du contrat ANR-PIA en cours

18/02/2014


x1
Nb de connexion réseau
x1
Nb de connectiques Fibre
Firewall x 2
VLAN-ORTOLANG
Ethernet 1Gb/s
x12
FC
x3
FC
x3
3 serveurs DELL R620
avec 8 CPUs et 128 Go
Ram chacun
Cluster Ortolang
Switch SAN
mutualisé 8Gb/s
Switch SAN
mutualisé 8Gb/s
FC
x4
18 disques SATA
de 3 To dédiés
Ortolang
FC
x4
Baie de stockage
mutualisée
FC
x1
Sauvegarde
Librairie Quantum
avec 2 lecteurs LTO6
et 50 slots 300To
FC
x1

Les services offerts

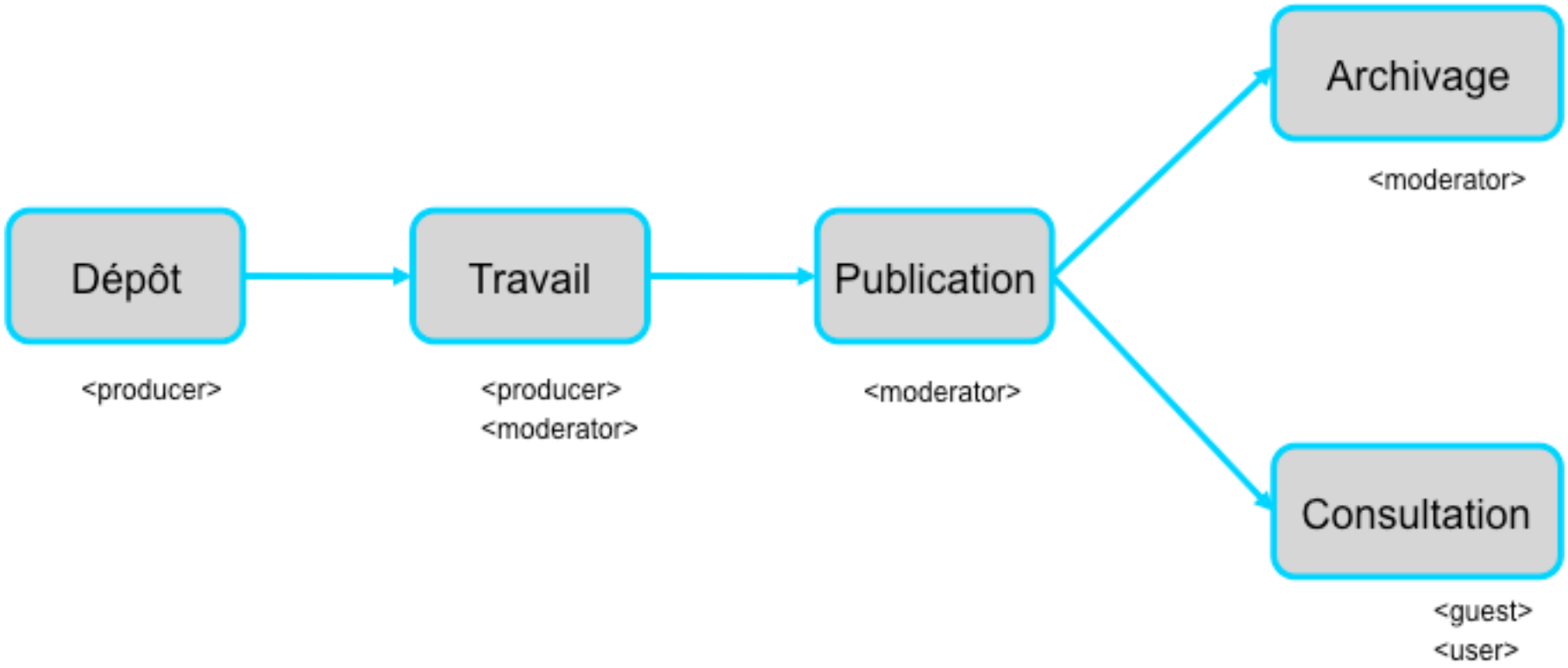
- Héberger des objets numériques (corpus, dictionnaires, lexiques et outils de traitement)
  - Organiser les objets dans des collections
  - Enrichir les objets avec des métadonnées
  - Proposer un catalogue des objets disponibles
- Assurer la fiabilité du stockage
  - Contrôler l'accès aux objets
  - Identifier de manière unique les objets
  - Garder un historique des états des objets
  - Assurer l'archivage à long terme des ressources via la solution proposée par HumaNum et le CINES
- Valoriser, diffuser et partager les ressources
  - Faciliter le travail avant publication
  - Collecter des statistiques
  - Assurer des notifications d'usage aux utilisateurs

Contexte général

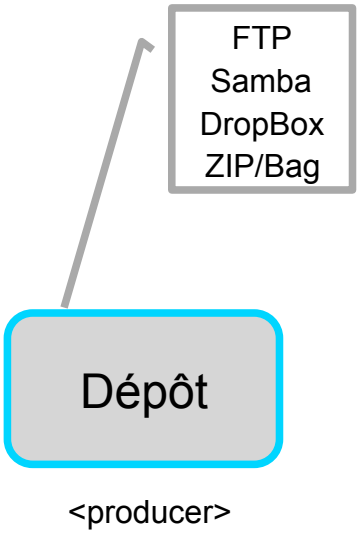
- L'hébergement, le stockage et l'archivage des objets ne sont en fait qu'une partie du processus.
- En amont, la phase de dépôt et de mise en forme est capitale.
  - Les utilisateurs émettent le souhait de déposer une nouvelle ressource,
  - Ils proposent alors un ensemble de données (fichiers) et de métadonnées, fruit d'un travail de recherche pas toujours prêt à être publié en l'état car il ne répond pas forcément aux contraintes de publication ou d'archivage.
  - Les administrateurs de la plateforme doivent alors assurer un travail de mise en forme en collaboration étroite avec le producteur des données
  - Une fois les objets prêts à être diffusés, il faut garantir la pérennité des données (elles ne changeront plus).
  - 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.

Flux du travail

On distingue 5 types de travail sur la plateforme



Le dépôt

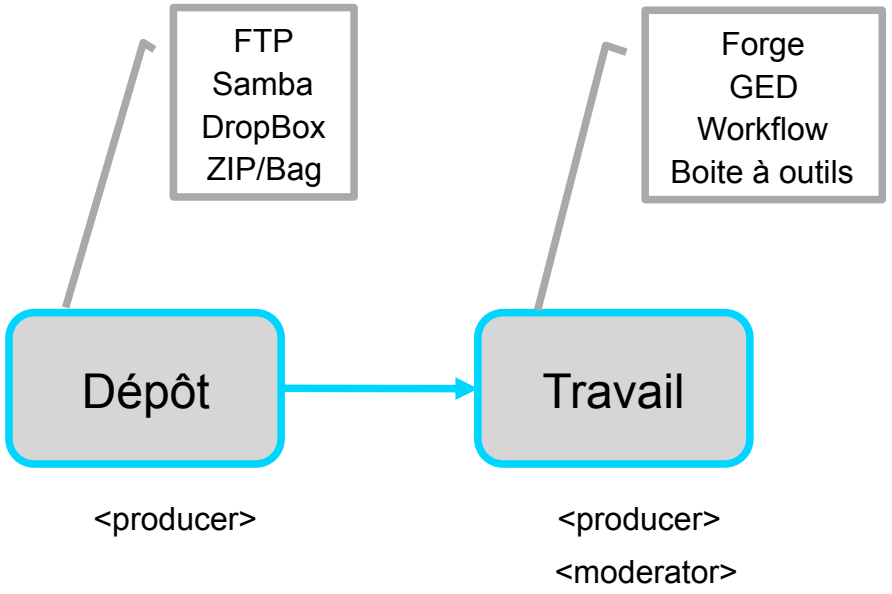


- Le producteur dispose d'un moyen simple de dépôt de ses données même si ces données sont encore impropres à la publication.
- Différentes méthodes sont proposées pour le dépôt ou téléchargement :
  - via FTP
  - sur un partage réseau
  - via une interface web
  - par téléchargement de fichiers compressés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

Un accompagnement dès cette phase de dépôt

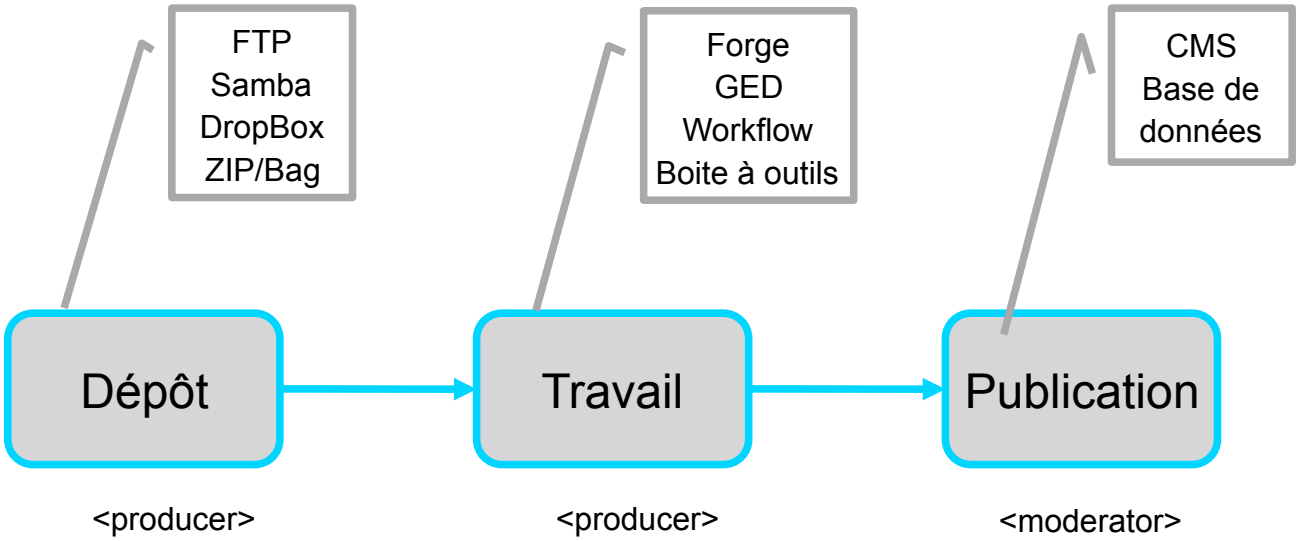
- Trois Centres de compétences thématiques
  - Oral : (SLDR & Modyco)
  - Multi-modal (SLDR & Modyco)
  - Ecrit (ATILF/CNRTL)
- Interactions directes et personnelles avec les membres de ces centres compétences
  - pour le SLDR : resp. Arnaud Gingold successeur de Bernard Bel
  - pour MoDyCo : resp. Christophe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

Espace de travail sécurisé



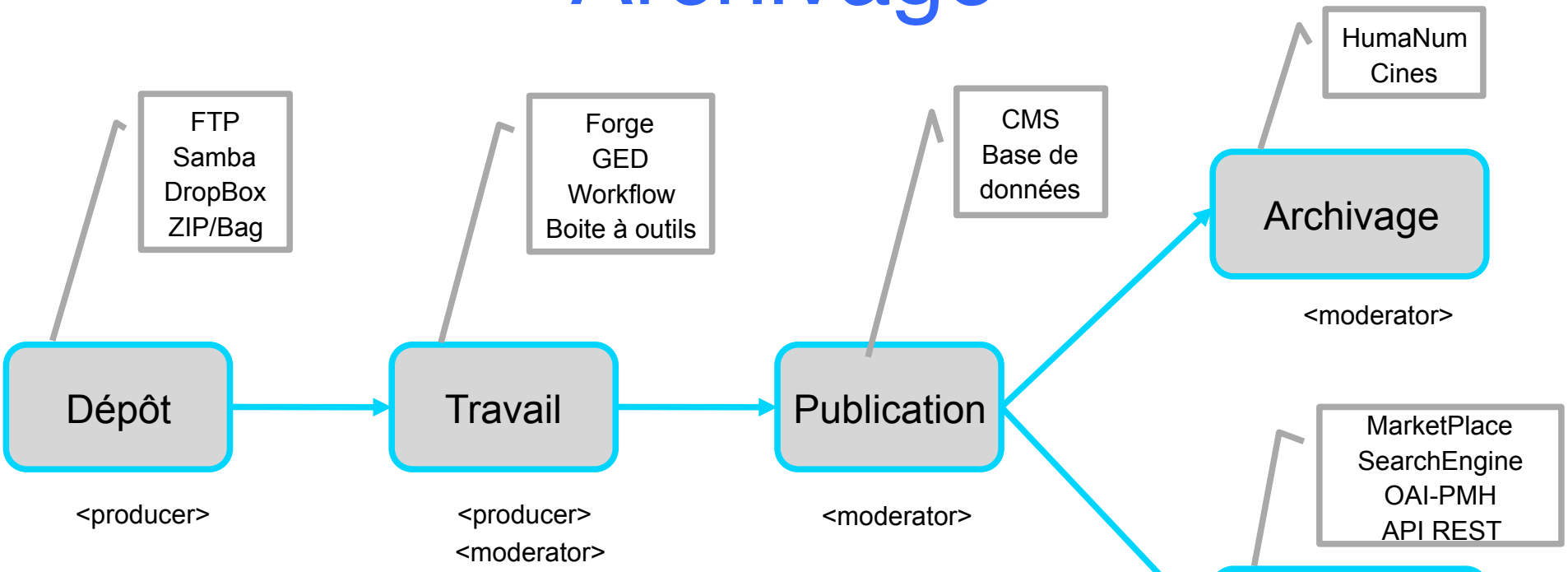
- Via un espace de travail en ligne, un partage de fichiers et une API, le producteur peut travailler sur ses données.
- Cet espace de travail est sécurisé (sauvegarde journalière)
- Il bénéficie en outre d'outils en ligne spécifiques qui vont lui permettre d'enrichir son travail (alignement, annotation, …)
- 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 plateforme

Publication



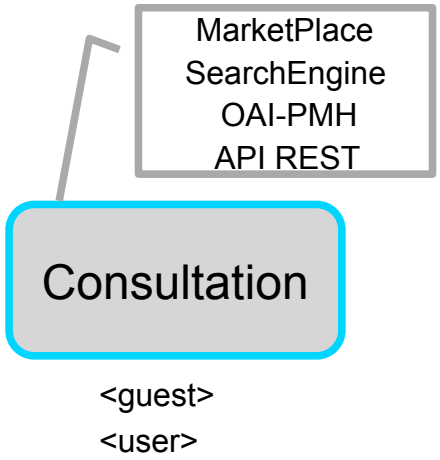
- Une fois que les données sont prêtes, le producteur peut soumettre son travail pour publication.
- Il peut alors suivre l'état de ses demandes et, en collaboration avec les administrateurs, aboutir à une version stable de sa ressource.
- Plusieurs formes de publications sont possibles (tous, restreint,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

Archivage



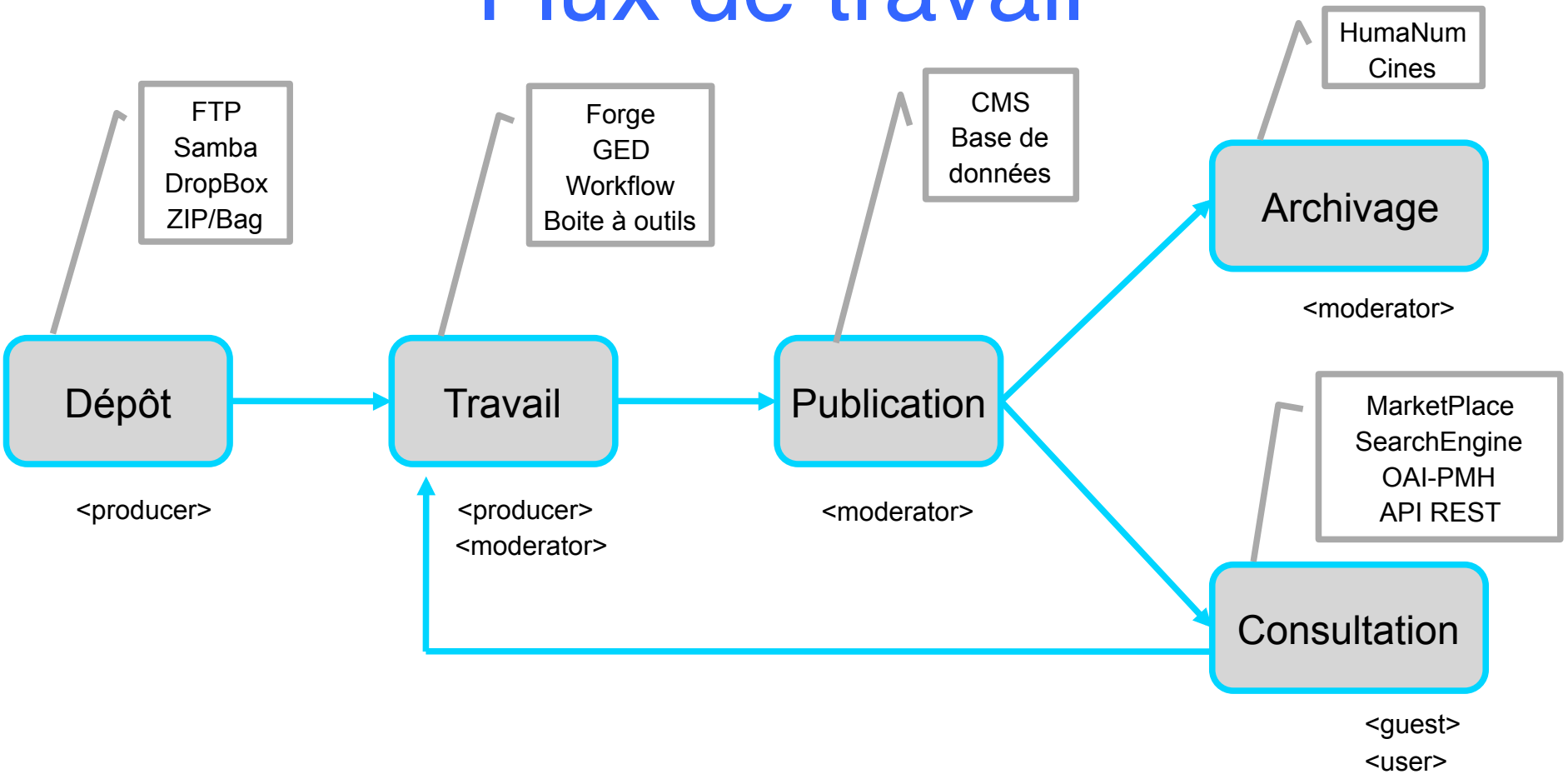
- Les données publiées peuvent être soumises pour archivage et/ou pour consultation simple.
- L'enrichissement automatique des données pendant les phases antérieures a permis de disposer de données 'propres' à être consultées.
- De plus le format d'archivage a été vérifié.

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
  - une base de données sémantique
  - un connecteur OAI-PMH
  - une interface REST
- L'enrichissement automatique des données pendant les phases antérieures permet de disposer de données propres et homogènes.

Flux de travail



- Les données publiées peuvent être référencées dans un nouvel espace de travail.

Les fonctionnalités offertes : démonstration : www.ortoang.fr

- Ressources disponibles
  - Corpus : https://www.ortolang.fr/index.html#/market/corpora
  - Outils : https://www.ortolang.fr/index.html#/market/tools
  - Lexiques : https://www.ortolang.fr/index.html#/market/lexicons

- Intégration d'outils pour enrichir des données

Echantillon de corpus disponibles

- Modélisation Contrastive et Computationnelle des Chaînes de Coréférence
- CoMeRe (Communication médiée par les réseaux)
- Littéracie Avancée
- Corpus 14
- PERCEO : Corpus d'apprentissage d'Etiqueteur pour l'Ecrit et pour l'Oral
- Corpus journalistique issu de l'Est Républicain
- TCOF : Traitement de Corpus Oraux en Français
- ESLO
- Corpus d'articles de linguistiques issus de la revue "Sciences Humaines"
- Le Pelerinage de Vie humaine de Guillaume de Digulleville
- DEDE : un corpus annoté pour le traitement des DEscriptions DEfinies
- Colaje
- + les corpus du CNRTL et du SLDR

Quelques outils téléchargeables

- FastKwic
  - FastKwic est un outil d'indexation automatique permettant de produire un concordancier.
- Flemm
  - FLEMMv3.1 est un ensemble de modules Perl5 qui effectue l'analyse flexionnelle de textes en français préalablement étiquetés au moyen de Brill ou TreeTagger.
- Jtrans
  - Jtrans est un logiciel pour l'alignement texte parole.
- Jsafran
  - Jsafran est un logiciel Open source pour l'analyse syntaxiques des dépendances en langue française.

Exemples de lexiques disponibles

- Morphalou

  Morphalou3 est un lexique à large couverture. À chaque lemmes sont associées toutes ses formes fléchies (déclinaisons et conjugaisons du lemme). Par exemple le lemme nom commun « reine » donne deux formes fléchies : « reine » (singulier) et « reines » (pluriel.)

- TlfPhraseo

  lexique ouvert des syntagmes définis du Tlfi, c'est-à-dire de syntagmes considérés comme suffisamment remarquables par les rédacteurs du dictionnaire pour disposer d'une définition qui leur est propre.

- Demonette

  un réseau composé de relations morphologiques dérivationnelles.

- Prolex

  plate-forme technologique comprenant un dictionnaire électronique relationnel multilingue de noms propres (Prolexbase), des systèmes d'identification des noms propres et de leurs dérivés, des grammaires locales, etc.

Pour en savoir plus et suivre l'avancement du projet

www.ortolang.fr

contact@ortolang.fr