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

INVESTISSEMENTS D'AVENIR

Equipex ORTOLANG

Etienne Petitjean - ATILF - CNRS / Université de Lorraine

Historique

- Projet monté en 2011
- En réponse à l'appel d'offre Equipex dans le cadre des PIA2
- Un équipement d'excellence de mutualisation de ressources et d'outils sur la langue et son traitement informatique
- Budget total de 2,6M € sur 7 ans
- Dès le départ vision nationale et internationale
- Volonté marquée d'axer le projet sur les données ouvertes

un consortium réunissant des compétences complémentaires

Rappel des objectifs

- **Mutualisation de ressources et d'outils pour des travaux de recherche** :
  - la notion de ressources et d'outils est aujourd'hui incontournable spécifiquement en linguistique et en TAL
  - sans une véritable mutualisation chaque équipe de recherche se verrait dans l'obligation de tout réinventer
  - or la constitution et la normalisation de ressources et d'outils de qualité est très coûteuse
- **Valorisation des donnés de la recherche (corpus, lexiques, dictionnaires et outils de traitement)**
  - un des aspects essentiels aujourd'hui est leur informatisation et leur disponibilité sur la toile sous une forme facilement accessible et exploitable par l'ensemble de la communauté scientifique (FAIR)

Quelques caractéristiques

- Une ouverture pluridisciplinaire
  - SHS & Informatique
- Intégrant l'existant des centres nationaux de ressources
  - CNRTL et SLDR
- Gérant des ressources pour l'ensemble de la communauté scientifique:
  - ORTOLANG est une infrastructure de mutualisation : les ressources et les outils restent propriété des laboratoires.
- Droits d'accès définis par les propriétaires des corpus mais recommandation d'ORTOLANG :
  - **« Aussi ouvert que possible, aussi fermé que nécessaire »**

Les équipements matériels

- Plateforme:
  - 6 serveurs Dell avec 384 Go de RAM par machine
  - Baie de stockage Dell d'une capacité de 190To
  - Virtualisation avec VMWare
  - Sauvegarde:
    - Robotique LTO
    - Quantum DXI 4800
    - Externalisation (vaulting)
  - Hébergement à l'INIST (Réseau, Sécurité, Salles serveurs, Exploitation, Continuité de service et Haute disponibilité)
  - Mise à jour de toute l'infrastructure en 2019

- Matériels spécifiques:
  - Articulographe (Loria)
  - Système EEG (LPL)

La plateforme logicielle

- Héberger des objets numériques liés à la langue et à son traitement (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
  - Contrôler l'accès aux objets (ex: restriction à l'ESR)
  - Identifier de manière unique les objets (Identifiants pérennes)
- Assurer la fiabilité du stockage
  - Intégrité des données
  - Garder un historique des états des objets (gestion de versions)
  - Assurer l'archivage à long terme des ressources linguistiques
- Valoriser, diffuser et partager les ressources
  - Faciliter le travail avant publication
  - Présentation de statistiques d'usage aux utilisateurs
  - Faciliter les recherches multi-critères

Visibilité d'ORTOLANG

- Insertion dans le dispositif national
- Insertion dans le dispositif international
- Indices d'adhésion de la communauté au projet

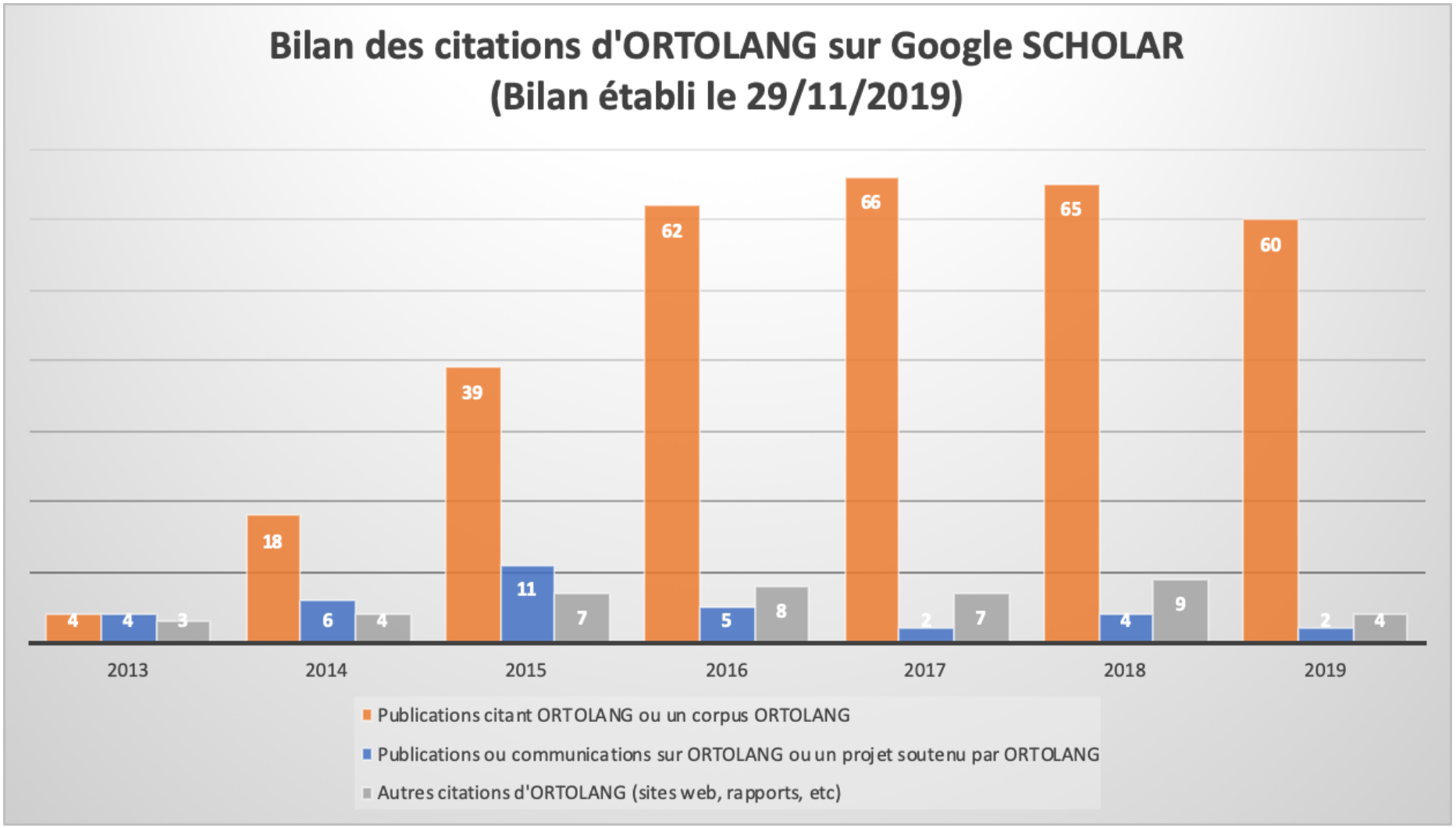
Insertion dans le dispositif national

- Liaison avec la TGIR Huma-Num
  - Service spécialisé pour la langue complémentaire de l'offre générale d'Huma-Num
  - Moissonnage des métadonnées et visibilité via ISIDORE
  - Appels communs à projets avec les consortiums Ecrit, IRCOM et CORLI « Corpus, Langues et Interactions »
- Partenariat avec la DGLFLF
  - participation au colloque TLRF 2015 (Technologies pour les Langues Régionales de France)
  - soutien de la DGLFLF : convention de 51 k€ signée pour 2015/2016
- Partenariat avec
  - Des projets ANR : ORFEO, OTIM, TermITH, Restaure, etc.
  - Des projets PIA : Labex EFL (Paris), BLRI (Aix-Marseille), etc.

Insertion dans le dispositif international

- DARIAH : Digital Research Infrastructure for the Arts and Humanities
  - Participation active de ses laboratoires supports en 2014
  - Contribution spécifique Ortolang en 2015-2016
  - Participation à la proposition H2020 Infra-Dev-3 portée par DARIAH, le centre DANS (NL) et l'université de Göttingen (DE) et le CNRS
- CLARIN : Common Language Resources and Technology Infrastructure
  - mi 2015 : décision de la France de participer comme observateur à CLARIN
  - début 2016 : instanciation de cette décision au travers d'Huma-Num
  - En 2018 : inclusion dans le VLO (Virtual Langage Observatory)
  - Ortolang souhaite se positionner comme un des nœuds français de CLARIN.

Visibilité (Google Scholar)



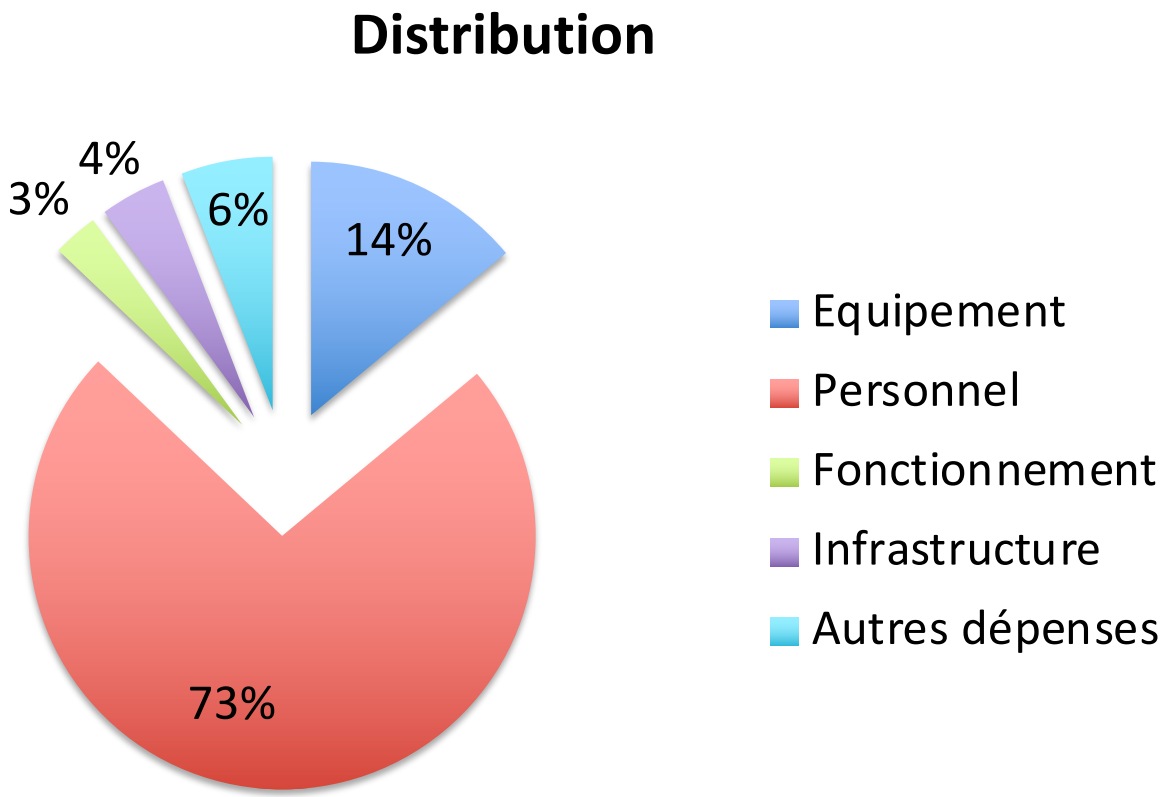
Ortolang

Bilan

- Bilan financier
- Evaluation ANR
- Résultats

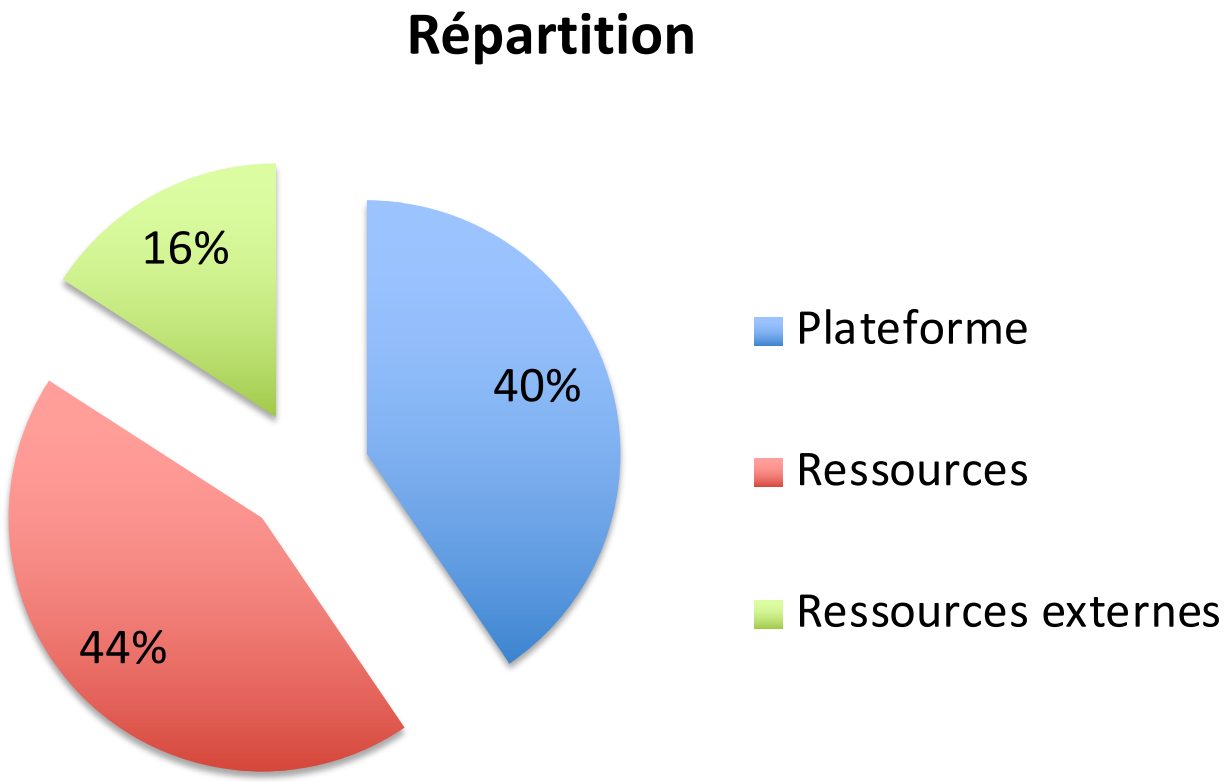
Budget tranche 1

Budget total : 2 200 000 €



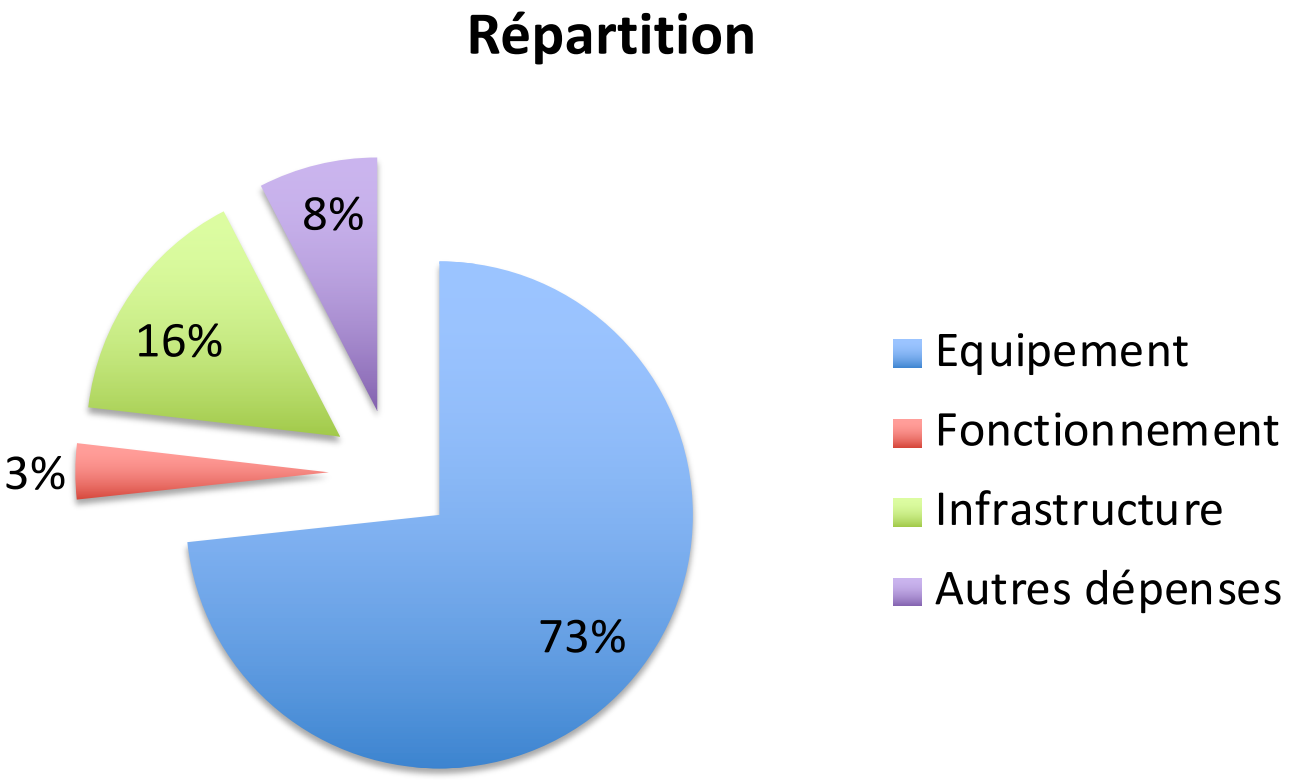
Personnel

Salaires: 1 615 000 €



Budget tranche 2

Budget total: 400 000 €



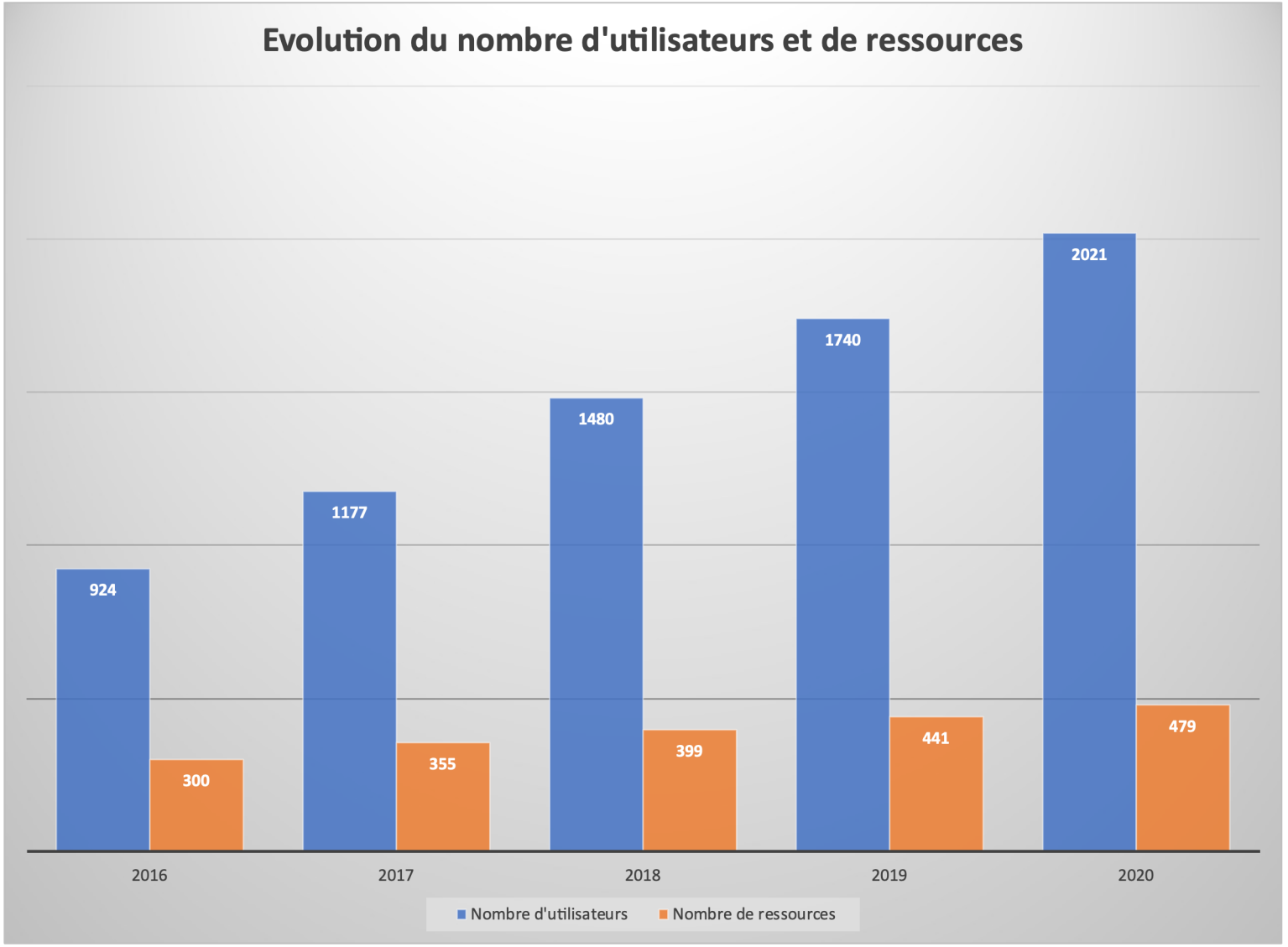
Evaluation ANR

- Evaluation mi-2017 à l'issue de la tranche 1 de l'Equipex
- Jury international constitué par l'ANR
- Rapport scientifique en anglais
- Résultats de l'évaluation:
  - Excellent projet
  - Un apport indiscutable pour la communauté scientifique
  - Validation des choix scientifiques

Résultats

- Une plateforme totalement opérationnelle
  - Respectant les objectifs initiaux
- Un projet et un budget bien géré
- Positionnement national et international
- Statistiques:
  - Plus de 2200 utilisateurs enregistrés
  - Plus de 500 ressources disponibles
  - Pour environ 11To de données

Evolution de l’usage



Pérennisation

- Moyens nécessaires pour la pérennisation d'ORTOLANG:
  - Equipements matériels et jouvence (~50K€ par an)
  - Personnels
    - Maintien des efforts des laboratoires partenaires pour assurer la mission d'accompagnement des utilisateurs
    - Besoin d'un ingénieur de développement (~40K€ par an)
      - Assurer l'évolution logicielle de la plateforme
      - Proposer de nouveaux services (ex: fouille de textes)

Perspectives

- Mise en place d'un nouvel accord de consortium pour pérenniser le projet en s'appuyant sur les laboratoires partenaires
  - Contribution sous forme de temps de personnel consacré à la plateforme
  - Contribution matérielle

- Prise en compte d'ORTOLANG dans le prochain CPER
  - Positionnement acté dans le pré-projet SHS COVD

Perspectives (2)

- Intégration dans Huma-Num sous forme d'un service spécialisé pour les données langagières opérée par le consortium ORTOLANG

- Intégration dans la future proposition de consortium sur les données langagières de la recherche (ex CORLI)

- Intégration dans le projet européen CLARIN

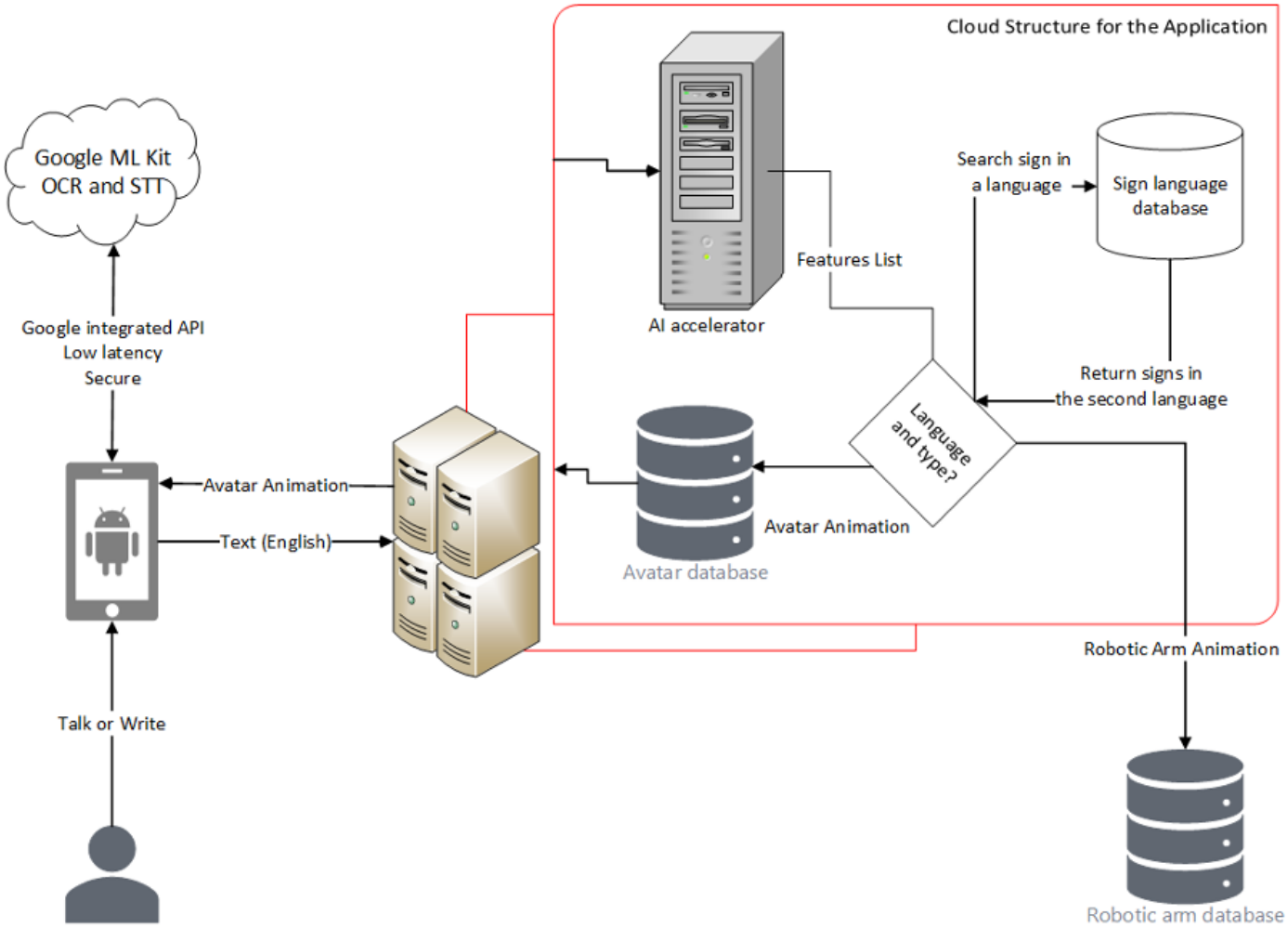
Depuis 2020

- Participation au dépôt d'un projet européen H2020
- Demande de certification CoreTrustSeal
- Préparation à la demande de Centre-B CLARIN
- Réponse à un appel à projet du Fonds National pour Science Ouverte
- Evolutions logicielles de la plateforme

SignFriends

- Projet européen H2020 - ICT-57-2020
- Budget de 4M€
- Partenaires multiples:
  - Laboratoires de recherche (France, Italie, Slovénie)
  - Institutions européennes (EBU)
  - Télévisions (RAI)
- Objectifs:
  - Créer une plateforme de traduction automatique du langage parlé vers le langage signé
- Déposé fin juin 2020

SignFriends (2)



Ortolang

CoreTrustSeal

- Certification internationale pour les entrepôts de données numériques
- « Self assessment »
- Formulaire de 16 questions:
  - Questions institutionnelles
  - Positionnement par rapport à une communauté
  - Qualité & sécurité des données
  - Qualité & sécurité des métadonnées
  - Documentation des procédures
  - Licences
  - Etc.

CoreTrustSeal (2)

- Plusieurs semaines de travail:
  - Rédaction en anglais (ATILF, INIST, Modyco)
  - Réflexion sur notre fonctionnement et nos procédures
  - Traduction de toute la documentation en anglais
  - Création d'un site multilingue de documentation (Gilles Toubiana)
    - https://ortolangdoc.atilf.fr
    - Ce site deviendra la vitrine d'ORTOLANG
- Calendrier:
  - Dépôt de la version 1 en septembre 2020
  - Première évaluation fin janvier 2021
  - Dépôt de la version 2 fin mars 2021
  - On attend la suite…

CLARIN

- Infrastructure de recherche européenne sur les données langagières
- Participation de 24 pays
  - La France est « observateur »
- Réseau de centres « nœuds »
  - B, C, K
- Notre objectif:
  - Devenir le premier Centre B français (22 au total)
  - Renforcer la visibilité internationale d'ORTOLANG

CLARIN (2)

- Certification technique:
  - CoreTrustSeal
  - Intégrer la fédération d'identité EduGain
  - Sécurité des échanges (HTTPS)
  - Moissonnage des métadonnées par le VLO (Virtual Language Observatory)
  - Gérer des identifiants pérennes (PIDs)
  - Implémenter le Federated Content Search (optionnel)

Projet FNSO

- Appel à projet 2021 autour des publications
- Projet CORPUCIT (Corpus, citations et visualisations)
- Objectif:
  - gérer les citations d'extraits de textes ou de corpus en générant un identifiant pérenne pour chaque citation, et de lier finement écrits scientifiques et données de langage (écrits, sons, vidéo, images) présentées dans leur contexte, facilitant la réflexion scientifique et la réutilisation des données.
- Dépôt fin mars 2021
- Budget: 200 000 euros

Evolutions logicielles

- Tous les développements pour être certifié CLARIN
- Mise à jour des composants logiciels
  - Keycloak, Wildfly, Postgres, Etc.
- Mise à jour du déploiement des composants
  - Docker
- Ajout de nouvelles fonctionnalités
  - Recherche par facettes
  - Démarrage du travail avec le CINES autour de l'archivage pérenne