ORTOLANG : Un équipement d'excellence pour la mutualisation et la valorisation des ressources sur le français et les langues de France

Jean-Marie Pierrel

TLRF – ACTES – Jean-Marie Pierrel : ORTOLANG[1] : un équipement d'excellence pour la mutualisation et la valorisation des ressources sur le français et les langues de France

Vous savez tous, et nous en avons suffisamment parlé, que les ressources linguistiques sont fondamentales pour pouvoir développer un certain nombre de travaux de recherche en traitement automatique des langues, à la fois pour faire émerger des modèles (approche stochastique ou symbolique) ainsi que pour les valider. Il est également indispensable d'avoir des ressources partagées et pérennes pour pouvoir comparer les résultats obtenus par différents systèmes de traitement.

Enfin, il est important de pouvoir valoriser nos langues et cet important travail de production de ressources. Pendant de trop nombreuses années, un nombre très important de ressources dont la constitution a pourtant coûté très cher, ont été perdues. Dans les années 90-2000, beaucoup de travaux de recherche et de thèses ont commencé par construire une ressource spécifique, un corpus, un lexique ou un outil de traitement. Une fois la thèse ou le projet terminés, le thésard s'en va, le corpus reste sur un ordinateur, et cinq ans après plus personne n'est capable de savoir où il est ni comment cela fonctionne.

Le premier objectif est donc la pérennisation de toutes les ressources produites par nos laboratoires sur les langues, et, au travers de leur mutualisation, permettre de conforter nos recherches. Par ailleurs il est évident qu'une équipe de recherche ne peut pas être performante dans tous les domaines et sans une véritable mutualisation de ressources chaque équipe de recherche se verrait dans l'obligation de tout réinventer. La constitution de ressources est tellement chère qu'il était nécessaire de faire quelque chose pour faciliter cette mutualisation et cette valorisation de ressources linguistiques et c'est précisément l'objectif d'ORTOLANG.

En termes de positionnement institutionnel, ORTOLANG est un équipement d'excellence, validé dans le cadre du programme d'Investissement d'avenir lancé par le gouvernement. Géré par le CNRS, il regroupe un certain nombre d'équipes, l'ATILF avec son centre de ressources sur l'écrit (le CNRTL), l'INIST et le LORIA, le LPL, qui gérait le SLDR centre de ressources sur l'oral, MODYCO et le LLL, et implique quatre universités (Université de Lorraine, Aix Marseille Université, Université Paris Ouest Nanterre, Université d'Orléans), le CNRS et l'INRIA.

Les objectifs d'ORTOLANG sont essentiellement d'outiller et valoriser les travaux qui se font sur le français et les langues de France, avec en particulier une commande de l'État pour proposer un certain nombre de ressources et d'outils de base pour le français. Nous aurons une présentation d'Huma-Num tout à l'heure, qui couvre tous les aspects des sciences humaines et sociales, ORTOLANG est un service spécialisé dans la langue, complémentaire de l'offre générale d'Huma-Num.

Comme pour tous les équipements d'excellence, deux phases ont été définies par le programme d'investissement d'avenir, une phase dite d'investissement, qui s'étale jusqu'à la fin de l'année 2016 et durant laquelle nous devons définir et stabiliser la plate-forme que nous proposons, puis une phase de fonctionnement dont les financements sont assurés jusqu'en 2020. La plate-forme ORTOLANG actuelle s'appuie sur une grappe de trois serveurs biprocesseurs avec 128Go de RAM chacun, offrant ainsi une puissance de calcul et de stockage suffisamment importante. Nous disposons en effet actuellement de 40To de disque utile partagés et d'un espace de sauvegarde allant jusqu'à 312To. Cela nous

[1] *Ortolang* bénéficie d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme "Investissements d'avenir" portant la référence ANR-11-EQPX-0032

permet d'assurer, au fur et à mesure que des ressources sont déposées sur la plate-forme, des sauvegardes journalières et incrémentales. Pour permettre une adaptation maximale aux besoins des chercheurs, nous avons développé une plate-forme logicielle spécifique qui peut être schématisée par le flux de travail présenté dans la figure 1 et qui structure les grandes fonctionnalités que nous proposons dans ORTOLANG.

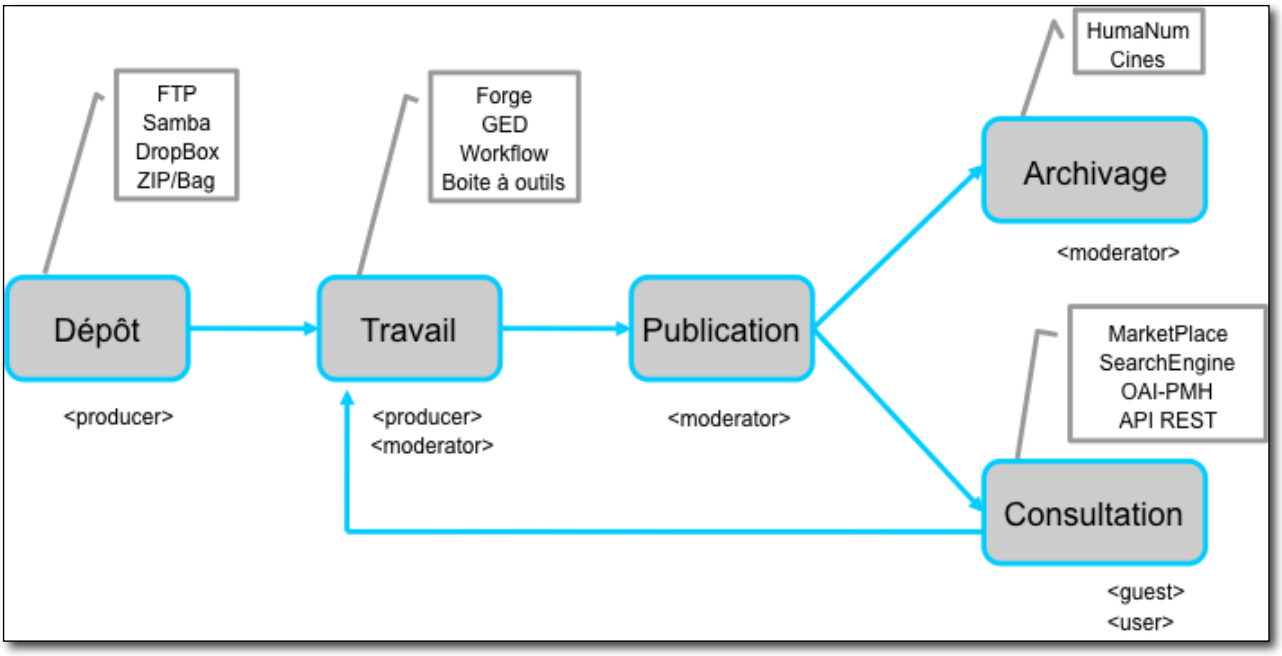

Figure 1 : Flux de travail au sein d'ORTOLANG

Chacun sait qu'héberger, stocker, archiver des ressources est une chose, mais que la plupart du temps quand des chercheurs ou un groupe de passionnés de la langue construisent une ressource, elle est difficilement publiable en l'état, car elle ne respecte pas forcément un certain nombre de standards indispensables pour qu'elle puisse être réutilisée par d'autres.

Le schéma classique, auquel on se trouve donc confronté, est que les utilisateurs souhaitent pouvoir déposer une ressource, mais ces derniers n'étant pas informaticiens la plupart du temps, il faut leur proposer une solution simple pour qu'ils puissent le faire. Cependant lors du dépôt, les ressources ne sont pas forcément prêtes à être publiées. Il est donc nécessaire d'accompagner les utilisateurs pour la standardisation et la documentation de leurs ressources en termes de métadonnées facilement interopérables avec d'autres systèmes. Il faut également mettre en place un processus de validation technique de la ressource avant de prendre la décision de la diffuser. À ce moment, la pérennité doit être assurée, c'est-à-dire que dès qu'une ressource est publiée, il faut pouvoir la retrouver sans aucun problème plusieurs années plus tard. Une fois publiée, elle pourra éventuellement être archivée ou exploitée par d'autres pour enrichir de nouvelles ressources. C'est ce qui nous a conduits à proposer un flux de travail qui distingue cinq étapes.

La première étape est la plus simple, il s'agit du dépôt. Nous avons choisi de proposer des choses simples pour l'utilisateur. Si la ressource n'est pas trop importante (ressource textuelle, lexique morphosyntaxique par exemple), il suffit de la glisser dans l'espace de travail sur la plate-forme, et le téléchargement se fait automatiquement. Cela est impossible s'il s'agit d'une ressource vidéo, car le débit du réseau ne permet pas ce genre de choses. Nous proposons donc aussi des systèmes de réseaux partagés asynchrones qui permettent de déposer facilement une ressource. Le dépôt peut se faire dans

différents formats, y compris dans des formats compressés qui seront décompressés automatiquement par la plate-forme.

Dès qu'elle est déposée, la ressource est prise en charge par ORTOLANG, ce qui signifie qu'elle va être sauvegardée systématiquement. Il s'agit ici d'une sauvegarde sécurisée : un système informatique assure une sauvegarde dans deux lieux différents, sur des machines différentes, pour être sûr de ne pas avoir de problème. Un nombre important de ressources ont en effet été perdues dans le passé à cause d'un crash du disque du chercheur. Dès que le dépôt est effectué, les déposants sont de plus accompagnés par les centres de compétences d'ORTOLANG : le LPL d'Aix-en-Provence et MoDyCo pour les ressources orales et multimodales, l'ATILF pour les ressources sur l'écrit. Cela implique des interactions directes entre les déposants et les membres des centres de compétences qui les accompagnent pour aboutir à une ressource publiable. Pour permettre à l'utilisateur de suivre l'évolution de sa ressource et être informé du stade où elle se trouve, des informations régulières lui sont alors transmises par les équipes d'ORTOLANG. Ainsi le déposant se voit ouvrir sur la plate-forme un espace de travail sécurisé où il pourra à la fois normaliser, standardiser sa ressource, enrichir ses métadonnées grâce à des systèmes interactifs que nous avons mis au point, et définir les accès qu'il souhaite pour ses données ainsi que la licence attachée à sa ressource : à ce stade, la ressource n'est visible que par les déposants ou par des ayants droit qu'il convient de définir. Ce n'est que lorsque ce travail est terminé que l'on passe à la phase suivante dite de publication. Là, l'utilisateur peut faire des choix sur l'ouverture et la visibilité de sa ressource. Ce que nous souhaitons, c'est que les ressources soient le plus possible entièrement libres. Néanmoins, un certain nombre de ressources ne peuvent être ouvertes qu'à la communauté de l'enseignement supérieur et de la recherche, voire limitées à une liste précise de personnes. Ainsi, si vous avez un corpus vidéo de suivi longitudinal de la constitution de lexiques et de l'apprentissage de la langue chez un enfant autiste en famille, il n'est pas possible de partager librement ces vidéos ne serait-ce qu'à l'ensemble des chercheurs de l'enseignement supérieur et de la recherche. Le déposant a alors la possibilité de n'ouvrir la publication qu'à un nombre limité de personnes clairement définies.

Une fois la ressource publiée, elle devient accessible via le site d'ORTOLANG. À ce moment-là, la ressource peut être archivée et consultée.

- Pour l'archivage, si le déposant le souhaite, une commission commune à ORTOLANG et à la TGIR Huma-Num décide si la ressource doit être archivée à long terme (sur trente ans ou plus). Le coût de l'archivage étant non-négligeable, si ce n'est qu'une version numérisé d'un texte imprimé la meilleure archive est sans doute le papier. Ce n'est peut-être pas nécessaire d'en faire une archive informatique qui coûtera plus cher que sa renumérisation. Par contre dans certains cas, comme pour la ressource ESLO, une enquête sociologique sur le parler français à Orléans dans les années 60, il est indispensable de l'archiver, car il ne sera jamais plus possible de la reconstituer.
- La consultation, quant à elle, peut se faire selon plusieurs modes : en téléchargement simple, avec une visionneuse spécifique présente sur la plate-forme ou sous forme de projet intégré, c'est-à-dire avec tout un site web spécifique pour pouvoir exploiter ou parcourir une ressource particulière, avec des possibilités de navigation, de recherche dans l'ensemble des ressources proposées.

Grâce à ce flux de travail, qui peut paraître un peu complexe, nous sommes certains que lorsqu'une ressource est proposée en consultation les données sont propres et homogènes : tous les tests ont été faits préalablement pour s'assurer de sa robustesse. Une fois qu'elle est publiée et consultable, la ressource peut être réutilisée dans une autre

espace de travail ou projet pour pouvoir l'enrichir, en respectant bien entendu la licence qui lui est propre.

Pour finir, précisons comment ORTOLANG s'insère dans le dispositif national et international

Au niveau national, ORTOLANG est un service spécialisé pour la langue, complémentaire de l'offre générale proposée par la TGIR Huma-Num (très grande infrastructure de recherche). L'archivage de nos ressources se fait via la solution proposée par Huma-Num. Notons par ailleurs que dans le cadre du partenariat que nous avons la TGIR nous avons lancé des appels communs avec les consortiums Ecrit et IRCOM pour la standardisation et la finalisation de corpus ou de ressources existantes dans les laboratoires, mais qui étaient jusqu'ici non diffusées pour la plupart.

Au niveau international, nous participons aux infrastructures européennes de recherche, dont Dariah qui sera présenté tout à l'heure. Faisons donc un petit focus sur l'autre infrastructure de recherche européenne, CLARIN, spécialisée sur les langues. La position de la France actuellement vis-à-vis de CLARIN n'est pas complètement claire, parce que la France n'a pas encore pris la décision de la rejoindre, mais nous sommes prêts à contribuer à la création d'une CLARIN-France dès que la France donnera son feu vert officiellement [2]. Pour mémoire, notre laboratoire, l'ATILF, était contractant du projet européen qui a conduit à cette infrastructure.

Quels sont les objectifs de CLARIN ?

- CLARIN veut proposer des ressources dans diverses langues : nous sommes prêts à gérer les ressources pour le français et les langues de France.
- Pour la diffusion, CLARIN veut qu'il n'y ait pas d'autre restriction que celle découlant de considérations éthiques ou juridiques : nous plaidons pour des ressources entièrement ouvertes aussi.
- CLARIN veut une intégration maximale des données, c'est-à-dire avoir des métadonnées de recherche des contenus qui permettent aux chercheurs, où qu'ils soient, de parcourir les ressources qui sont dans le réseau CLARIN. Nous respectons au niveau d'ORTOLANG les normes de représentation, en particulier les métadonnées compatibles avec celles de CLARIN.
- CLARIN souhaite qu'il y ait une intégration de service, pour toutes les modalités (texte, parole, geste), sous forme en particulier de web services. Nous allons proposer de tels web services d'exploitation au niveau d'ORTOLANG sur les ressources que nous gérons.
- CLARIN souhaite que les ressources accessibles via son réseau soient pérennisées, nous avons fait ce choix aussi. Cela signifie que chaque ressource peut être retrouvée à travers son identifiant pérenne n'importe quand, avec bien entendu un versionnage, puisque les ressources peuvent évoluer.
- CLARIN plaide pour une durabilité dans le temps. Dans notre système français ce point peut apparaître comme problématique : vous savez que même les plus grands laboratoires en France ont une durée de vie limitée, leur renouvellement étant discuté tous les cinq ans. Pour le moment notre existence est assurée jusque 2020, et c'est la réussite du projet qui fait qu'ORTOLANG sera pérennisé.

Aujourd'hui la première version de la plate-forme est accessible à l'adresse www.ortolang.fr, nous y intégrons toutes les ressources que nous avons déjà récupérées, et celles existant au sein des centres de ressources que le CNRS avait lancées en 2006

[2] Depuis la conférence, la position de la France vis-à-vis de CLARIN s'est clarifiée mi-juin 2015 par une décision d'être observateur au sein de cette infrastructure européenne.

(Centre de ressources textuelles, CNRTL, et centre de ressources sur l'oral SLDR à Aix) ainsi au cours de cette année ce sont plusieurs centaines de ressources (Corpus, Lexiques et outils de traitement)n qui seront directement accessibles sur notre plate-forme.