



Opération LR-1 : axe ressources Compte-rendu de la réunion du 1^{er} juin 2012

Laboratoires représentés :

ALPAGE, HTL, LACITO, LATTICE, LLF, LPP (Paris 3), MII, SEDYL

Objectifs de la réunion :

- ▲ Présenter le travail effectué depuis janvier 2012 par Sarra El Ayari, Clément Plancq et Benoît Sagot
- ▲ Discuter de la valorisation des ressources (moissonnage par les catalogues internationaux)
- ▲ Présenter les modalités de demande d'interventions de l'ingénieure recrutée à cet effet

Pour rappel, il s'agit d'une opération transversale, qui concerne toutes les équipes participantes ainsi que tous les axes. Les objectifs que nous nous sommes fixés sont les suivants : interopérabilité des ressources, utilisation des standards, visibilité et accessibilité.

Le travail d'inventaire des ressources

Dans un premier temps, nous avons dressé un inventaire des ressources (données linguistiques et outils) utilisées et/ou produites par ces laboratoires. Cet inventaire, toujours en cours de réalisation, a été effectué en rendant visites aux équipes participantes. Il nous a permis d'élaborer une grille de description des ressources compatible avec les standards de métadonnées. Nous disposons à l'heure actuelle de 72 ressources recensées. Afin de servir les objectifs mentionnés précédemment, nous avons développé une plate-forme en ligne d'enregistrement et de consultation de fiches descriptives de ces ressources : <http://ressources.labex-efl.org> Une fois qu'une ressource est enregistrée, il est alors possible de générer des métadonnées aux différents formats standards que sont le *Dublin Core*¹, le *TEI Header*², *OLAC*³ et *CMDI (CLARIN)*⁴.

Cette plate-forme permet également de rechercher des ressources selon différents critères (module de recherche à facettes) et propose à la fois une boîte à outils de scripts de conversions de formats ainsi qu'une page dédiée à la proposition d'autres ressources externes au LabEx. Ces deux dernières fonctionnalités ont pour but d'être alimentées par les participants eux-mêmes de façon collaborative.

Le processus de soumission et de sélection de appels à intervention de l'ingénieure

Dans un second temps, il est prévu que les participants puissent soumettre des demandes d'interventions de l'ingénieure de l'axe, que ce soit pour du développement logiciel, du support/conseil ou bien de la formation. Pour en bénéficier, il y a certains critères à respecter : le laboratoire demandeur doit avoir réalisé son inventaire au préalable, la demande doit être en lien avec une opération soutenue ou financée par le LabEx et il est souhaitable que le projet concerne plusieurs participants. Pour ce faire, nous avons mis au point un formulaire de demandes à remplir. Les demandes reçues seront validées par un comité *ad hoc* formé par les responsables de l'opération (Sarra El Ayari, Clément Plancq et Benoît Sagot), ainsi que Cédric Gendrot (LPP) et Isabelle Léglise (INALCO) – sous réserve d'acceptation. Les demandes pourront être envoyées au fil de l'eau.

Une page web consacrée aux projets sera accessible à tous et un email sera envoyé aux équipes partenaires après acceptation des projets afin que d'autres intéressés puissent se manifester.

1 Le Dublin Core (DC) est un schéma générique de description de ressources qui comprend 15 descripteurs.

2 Le TEI Header cherche à définir un format d'échange, de création, et de stockage de textes annotés.

3 OLAC est un formalisme de métadonnées qui étend le *Dublin Core*, avec des spécificités linguistiques.

4 CLARIN présente une organisation modulaire pour définir des métadonnées précises et complètes.

Discussions générales au LabEx :

Parmi les points soulevés lors de la réunion, deux ont fait l'objet d'échanges plus fournis : celui de l'abréviation des langues, notamment celle des langues des ressources, et celui du format et de l'abréviation des gloses, qui est une question qui se pose lors de l'étiquetage des corpus. Pour l'heure, nous utilisons ce qu'offre le [WALS](#)⁵, mais cet outil ne permet de noter toutes les langues recensées dans notre plate-forme. Les participants ont été d'accord pour dire que le LabEx était un bon endroit pour discuter ses questions et faire des propositions sur tous ces points. A cette fin, ils ont pensé qu'il serait bon d'organiser une réunion de travail, dont un des objectifs immédiats serait de choisir un référent adéquat pour les langues. L'idée d'organiser ces réunions en partenariat avec l'axe « Typologie » a été évoquée et semblait naturelle.

Nous avons également discuté de la procédure d'enregistrement des ressources répertoriées dans les catalogues de métadonnées français ([ISIDORE](#)), européen ([CLARIN](#)) et international ([OLAC](#)). Cela nécessite un hébergement pérenne des métadonnées. Pour les données orales, il serait intéressant de prendre contact avec le [CRDO](#)⁶. En ce qui concerne l'écrit, un partenariat avec le [TGIR Corpus](#)⁷ et [ORTOLANG](#)⁸ pourrait aboutir à des solutions à moyen terme.

5 Le [WALS](#) (*World Atlas of Language Structures*) est une base de données regroupant des propriétés de langues (phonologiques, grammaticales, lexicales).

6 Il s'agit du *Centre de Ressources pour la Description de l'Oral*, lequel repose principalement sur le [LACITO](#).

7 [TGIR CORPUS](#) (*Coopération des Opérateurs de Recherche Pour un Usage des Sources numériques*) est une *plateforme de coopération* pour l'accès aux principaux ensembles documentaires (images, sons, textes) produits en priorité dans le cadre de recherches dans les SHS en général, et notamment en linguistique, psychologie, histoire, archéologie, philosophie, anthropologie, géographie, lettres et arts.

8 [ORTOLANG](#) (*Open Resources and TOols for LANGuage*) est un équipement d'excellence dont le 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.