

Infrastructure de Recherche CORPUS

Réunion du 06 Mai 2011

Ce PPT contient à la fois les éléments présentés par Laurent Dousset au sujet de l'IR Corpus lors de la réunion avec un certain nombre de linguistes ou de personnels rattachés à des unités de linguistique, qui a eu lieu au siège de CNRS le 06 Mai 2011, ainsi qu'un certain nombre d'éléments qui ont été évoqués lors de la discussion. Ce PPT vaut de Compte Rendu de la réunion.

Il contient des propositions qui doivent être évaluées et discutées. Ce PPT doit donc servir de base de discussion. Il ne vise pas l'exhaustivité des questions et problèmes qui doivent être adressés, et l'auteur est conscient du fait que d'autres interlocuteurs, qui auraient dû être présents à la réunion mais qui pour diverses raisons n'ont pas pu venir ou n'ont pas été convoqués, doivent contribuer à l'évolution des propositions.

CORPUS – Infrastructure de Recherche

<http://www.corpus-ir.fr/>

4 Infrastructures SHS décidées dans la feuille de route de Décembre 2008

- Adonis : unification des données (moteur de recherche généralisé) + services divers
- Progedo : gestion et réservoirs des données sociales « quantitatives »
- BSN : bibliothèque scientifique numérique
- **Corpus : gestion et production de réservoirs de données « qualitatives » (archives scientifiques). L'IR Corpus est créé cette année, 2011.**

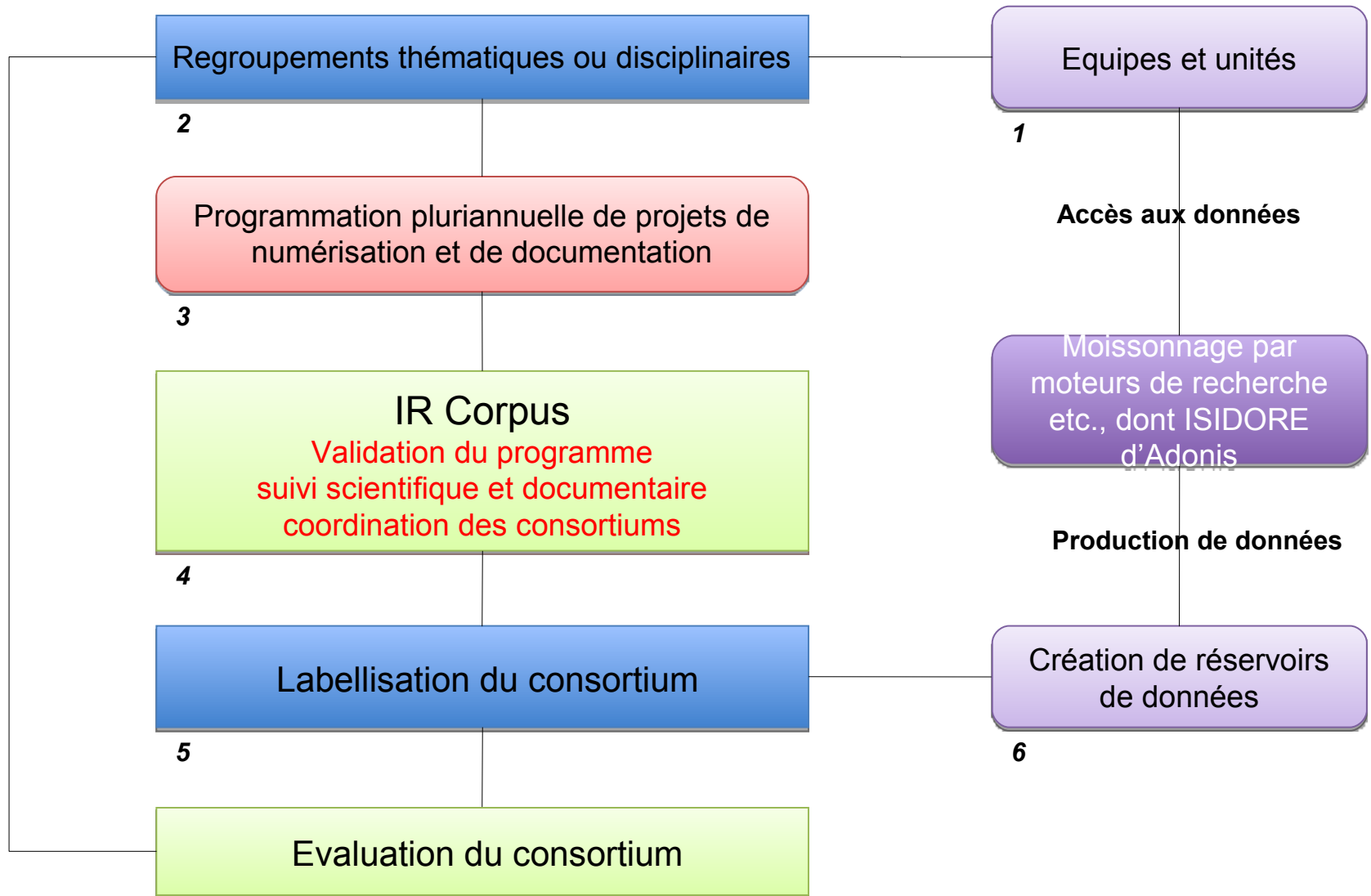
IR CORPUS créée à partir de plusieurs constats/besoins:

- Utilisation partagée des données qualitatives des SHS insuffisante
- Sauvegarde du patrimoine scientifique et humain accumulé par les SHS nécessaire
- Evolution générale des « Digital Humanities »: nécessité d'interopérabilités entre réservoirs

Objectifs généraux de l'IR CORPUS:

- Création de **consortiums** disciplinaires ou thématiques qui s'accordent sur les méthodologies de sauvegarde et de partage des données numériques **autour d'objets numériques identifiés.**
- Création de réservoirs de données numériques des consortiums

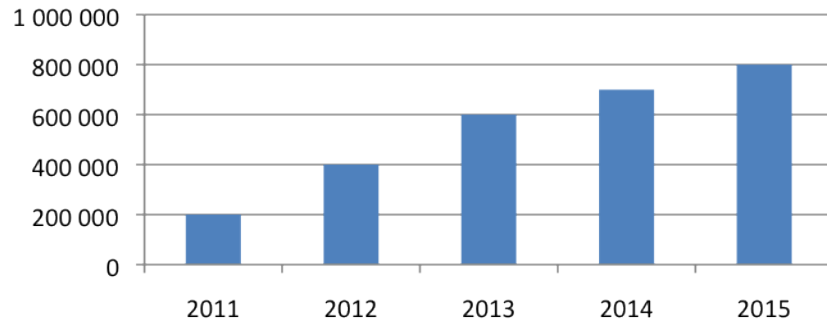
Workflow: création des consortiums au sein de CORPUS



Principes:

- Corpus reçoit sa dotation de la cellule des TGIR/TGE du Ministère (actuellement via CNRS).
- Corpus redistribue sa dotation aux Consortiums (moins une retenue pour son propre fonctionnement)
- Les Consortiums reçoivent une dotation annuelle sur le quadriennal ou quinquennal de leur labellisation
- Les Consortiums gèrent eux-mêmes, en accord avec Corpus, leurs dotations

Evolution programmée (espérée) de la dotation de Corpus:



Financement des Consortiums:

- Dotation récurrente le temps de la labellisation (4 ou 5 ans)
- Dotations utilisables uniquement pour numérisations, documentation et organisation du Consortium (services et financements de développement d'outils et d'hébergement à voir avec Adonis)
- Moyenne de 50 à 70 kEuros par an et par Consortium

Qu'est-ce qui est attendu des Consortiums?

Principes de l'interopérabilité:

- Concertation sur les standards et formats de l'archivage et du traitement numérique (et leur évolution)
- Production d'archives et de données
- Valorisation / visibilité / accessibilité des fonds (à ne pas confondre avec « open access »)
- Ouverture vers / intégration de la communauté
- Devenir une référence sur les « manières de faire » et standards au sujet des types de corpus/objets traités (par exemple par le biais de la publication de Guides, de l'organisation de formations etc.)

Principes de gestion:

- Transparence et concertation dans l'utilisation des ressources
- Comptabilité annuelle des apports en ressources financières, humaines et logistiques dans le domaine de la production de Corpus en général (au delà de la seule contribution de l'IR Corpus).

Principes de justification:

- Conformité avec les procédures d'évaluation et de labellisation
- Organisation des activités sous forme de projet
- Prises de décisions sous forme de concertations

Questions fréquentes

Question: Corpus semble ajouter une couche supplémentaire dans l'organisation des laboratoires, ou un couche supplémentaire dans les institutions...

Réponse: Les organisations existantes (MSH, fédérations,...) peuvent très bien fonctionner comme des partenaires et porteurs de consortiums, et même éventuellement constituer un consortium eux-mêmes. Corpus doit être compris comme étant organisé sous forme de « projets » (de concertation dans, et de visibilité de, l'activité numérique). Les consortiums définissent les procédures, formats et standards au sujet d'objets numériques particuliers.

Question: Les activités de notre laboratoire se situent dans plusieurs types de données, standards....

Réponse: Il n'y a aucun problème à ce qu'un laboratoire fasse partie de plusieurs consortiums. Il y aura même des consortiums dont les interfaces devront être particulièrement transparents. Et il faudra créer des ponts (groupes de travaux etc.) entre les consortiums. Ce qui doit préoccuper les consortiums sont les procédures autour de certains types d'objet numériques qui souvent dépassent les limites disciplinaires: le texte, la photo, les enregistrements sonores....

Question: Les dotations par consortium ne sont pas très importantes, faut-il alors diminuer le nombre de laboratoires partenaires ou démultiplier les consortiums pour pouvoir financer davantage de projets ou d'activités?

Réponse: Plusieurs points:

- Etre un consortium c'est être labellisé. Cette labellisation sera reconnue, et devra avoir des effets leviers dans la recherche d'autres financements (ANR, CE etc.). La dotation doit permettre d'organiser le consortium et d'entamer des projets. Un projet ou équipe qui est déjà financé via l'IR-Corpus devrait pouvoir augmenter ses chances de réussite dans des appels à projets.
- Les consortiums doivent être inclusifs et avoir un effet de concertation collective. Exclure des laboratoires du consortium entraine le risque de ne pas être suffisamment représentatif pour le traitement d'un type d'objet particulier, et donc de ne pas être labellisé.
- La dotation moyenne de 50 à 70kE par an et par consortium est certes limitée, mais s'inscrit dans le temps. Sur une labellisation sur 4 ans, cela signifie entre 200 et 280kE au total. Il est donc important, lors de la création du consortium, d'organiser les projets et travaux selon leurs priorités et de définir le type d'action envisagé (formations, numérisations, réunions, publications, documentations, mise en route de projets corpus....). Le type d'action le plus efficace et nécessaire n'est pas forcément identique d'un consortium à l'autre.

Questions fréquentes (relation IR Corpus - TGE Adonis)

Question: Est-ce que Corpus héberge les sites, bases de données, plateformes sur des serveurs?

Réponse: Non. Corpus ne prend pas en charge ce genre de services. Par contre, l'IR Corpus travaille en étroite relation avec le TGE Adonis, qui lui gère la Grille Adonis sur laquelle des plateformes peuvent être hébergées.

Question: Est-ce que Corpus finance le développement d'outils?

Réponse: Non. Seules l'organisation du consortium, la numérisation et la documentation peuvent être financées par le biais de la dotation de Corpus. Par contre, vous pouvez inclure dans les projets la nécessité de développement d'outils qui sera transmise à Adonis pour évaluation et éventuellement financement.

Question: Est-ce que Corpus moissonnera les bases de données et plateformes pour créer une interface d'accès unifiée?

Réponse: Non. Par contre, Corpus veillera à ce que les plateformes soient moissonnables et moissonnées par Isidore (Adonis) et veillera à ce que les plateformes soient interopérables avec d'autres prestataires/institutions (via OAI-PMH, via les outils développés par Clarin etc.)

Questions fréquentes (relation IR Corpus - ANR)

Question: Quelle est la relation entre IR Corpus et les ANR « Corpus »

Réponse: Il n'y a aucune relation institutionnelle ou formelle. Mais l'IR Corpus, le TGE Adonis et l'ANR discutent sur les procédures et le suivi des projets ANR-Corpus. L'IR Corpus et les consortiums doivent, autant que possible, fonctionner comme de garants de la qualité et conformité numérique des partenaires des ANR Corpus.

Consortiums linguistiques: Discussions (1)

Est-il judicieux d'avoir un seul consortium linguistique? Est-il judicieux de parler de consortiums disciplinaires, sachant que les objets traités sont nécessairement pluridisciplinaires? Quel est le type d'actions concret qu'un ou des consortiums linguistiques peuvent envisager?

Nombre de consortiums:

Il n'existe pas de taille critique relative aux consortiums. Ils doivent être cohérents dans le choix des objets qu'ils traitent et les participants doivent être représentatifs/reconnus dans le domaine que le consortium se donne comme définition. Les consortiums ne sont pas figés dans le marbre mais peuvent se redéfinir et se regrouper autrement même pendant le quadriennal de leur labellisation. Bien sûr ces redéfinitions, lorsqu'elle remettent en question la cohérence de l'IR Corpus, doivent être validées par les instances de l'infrastructure.

Un seul consortium linguistique semble difficile à gérer et mal représenter la diversité des objets numériques. Une certain nombre de proposition sont faites:

- Consortium sur l'oral / sur les usages / sur la parole
- Consortium sur le texte / l'écrit
- Consortium multimodal (qui s'intéresserait aux liens entre objets de nature différente) / multilingue (? Le multilingue implique-t-il véritablement des objets numériques distincts?)

- Il est également suggéré de créer, au sein des consortiums, des groupes de travail spécifiques sur des objets ou liens entre objets particuliers. Le « multimodal » pourrait d'ailleurs être un groupes de travail conjoint entre les deux consortiums « oral » et « texte ».

- Est évoqué également la nécessité de s'interroger sur les objets et les liens avec les sciences cognitives, les neurosciences, la psycholinguistique etc.

Consortiums linguistiques: Discussions (2)

Les types d'actions concrètes évoquées sont:

- Formations et réunions d'informations/concertations
- Réunions dont l'objectif est d'élaborer une stratégie de structuration future des consortiums
- Mise en route de projets Corpus
- La question des droits, en particulier dans le domaine de la pérennisation des données, devra être abordée par le(s) Consortium(s), et ceci en relation avec les tutelles, l'Alliance etc.
- Groupes de travail sur des questions particulières et pour établir des ponts entre consortiums

Porteurs du consortium:

Au moins pendant la phase préparatoire, il a été suggéré que les Fédérations et que la DGLFLF jouent un rôle important dans le relais de l'information au sujet des Consortiums en cours de constitution.

Il sera nécessaire de déterminer rapidement un ou des porteurs potentiels.

Les étapes prochaines

Etape 1 : Circuler ce PPT parmi les chercheurs / unités / équipes intéressés.

Etape 2 : Créer une liste de discussion.

Etape 3 : Réfléchir et discuter via cette liste sur le nombre et la nature/objet/objectifs des consortiums, leurs organisations etc.

Etape 4 : Commencer à proposer / déterminer un/des porteur(s) du/des consortium(s) (nécessairement une unité ou équipe qui possède sa propre gestion financière). Le porteur doit être *primus inter pares*, et sera l'interlocuteur principal avec l'IR Corpus en matière administrative et financière.

Etape 5 : Organiser une nouvelle réunion début juin.

Etape 6 : Passer en phase de construction du/des consortium(s) et procédure de pré-labellisation.

Etape 7 : Evaluation par le comité de pilotage de l'IR Corpus du pré-projet avec objectif de financement dès juillet 2011 si le pré-projet (ou projet) est validé.