ISTEX, vers des services innovants d'accès à la connaissance

Synthèse rédigée par Raymond Bérard, directeur de l'ABES, à partir du dossier de candidature d'ISTEX aux Initiatives d'excellence et des réunions de travail des partenaires du dossier.

ISTEX en quelques mots

ISTEX – Initiative d'excellence de l'information scientifique et technique (IST) – fait partie des initiatives d'excellence (IDEX) financées par les investissements d'avenir qui ont pour ambition de hisser l'enseignement supérieur français au niveau des meilleures universités du monde.
ISTEX vise à la réalisation d'un socle documentaire numérique pérenne, commun à l'ensemble des Initiatives d'excellence, offrant des services et des usages complémentaires et interopérables avec ceux mis en place dans les établissements et organismes concernés.
ISTEX comporte deux volets :
• l'acquisition sous forme de licence nationale d'un corpus inégalé de ressources documentaires ;
• l'agrégation de ces ressources au sein d'une plateforme nationale apportant une plus-value basée sur le traitement des données en texte intégral.

Pourquoi ISTEX ?

Est-il encore nécessaire de souligner que disposer de ressources documentaires riches est essentiel pour une production scientifique de rang mondial ? Plusieurs études ont montré la corrélation entre la disponibilité de ces ressources et la productivité et la qualité de la recherche[1]. Une autre corrélation a été mise en évidence entre l'information scientifique disponible et l'innovation exprimée en termes de dépôt de brevets.

Cependant, notre pays ne dispose pas encore de ce socle documentaire du fait d'acquisitions en retrait par rapport aux grands pays européens. De plus, la structure actuelle de l'accès aux ressources numériques, qui constituent désormais la majeure part de la documentation scientifique, ne permet pas d'en tirer le meilleur profit. Celles-ci sont en effet disponibles séparément sur les plateformes de leurs éditeurs.
La construction d'un outil national et pluridisciplinaire est la mieux à même :
• de garantir la capacité de maintenance d'un système de stockage et d'accès de grande ampleur ;
• de renforcer le poids des acheteurs publics de la recherche regroupés face à une offre oligopolistique ;
• de favoriser toutes les formes de valorisation scientifique de corpus documentaires très étendus, en permettant les croisements interdisciplinaires ainsi qu'avec les données de la recherche. L'ensemble de données ainsi constitué sera disponible en permanence pour une ingénierie scientifique d'un niveau sans commune mesure avec celle des communautés qui travaillent aujourd'hui sur des plateformes juxtaposées ;
• de contribuer à l'intégration d'un espace européen de recherche.

Un contexte favorable

[1] Voir notamment celle du *Research Information Network*, menée auprès de 115 universités du Royaume-Uni, qui fournit des résultats sur la corrélation entre les dépenses documentaires, les téléchargements d'articles de revues scientifiques et la productivité de la recherche exprimée en termes d'articles publiés, de doctorats délivrés et de contrats de recherche.
Research Information Network, « E-journals : their use, value and impact », 2009.
En ligne : http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/e-journals-their-use-value-and-impact.

À l'heure où la politique nationale de l'enseignement supérieur et de la recherche vise la création de pôles d'envergure internationale, il convient de doter ces derniers d'un outil à la hauteur des enjeux actuels et seule une initiative commune peut être en mesure de relever ce défi. La Bibliothèque scientifique numérique (BSN), qui permet d'appréhender collectivement la question de l'information scientifique et technique en mettant en cohérence les opérateurs nationaux de l'IST, constitue un cadre pour la mise en œuvre d'ISTEX.

Les acquisitions

L'acquisition de ressources documentaires sous forme de licence nationale, après validation par un comité de pilotage fédérant l'ensemble des acteurs, présente un double intérêt :

- garantir à chaque chercheur la documentation scientifique qui lui est nécessaire, quel que soit le caractère minoritaire ou interstitiel de son champ de recherche au sein de son établissement ou de la communauté nationale ;
- mobiliser l'ensemble des communautés pour valoriser au mieux un système transformant appelé, à partir d'un noyau de base, à élargir ses services au plus près des besoins de la recherche. L'appropriation interdisciplinaire de l'outil est en effet un facteur de production de valeur ajoutée.

Les acquisitions porteront sur des corpus de livres électroniques, de grands corpus patrimoniaux numérisés, des archives de bases de données, des collections rétrospectives de périodiques. Un comité technique, composé des responsables des négociations de Couperin et du CNRS ainsi que des responsables des services de documentation des autres établissements publics scientifiques et techniques et d'autres représentants des différentes communautés scientifiques, a déjà travaillé à un premier recensement des ressources utiles pour la recherche dans le cadre de la coordination des politiques d'acquisition de ressources documentaires électroniques de BSN. Ce premier travail constitue un socle mais doit être revu de façon plus systématique et structurée avec un dispositif formalisé de remontées des besoins des communautés disciplinaires et des procédures d'évaluation des ressources au plan des contenus, des modalités techniques d'accès et d'usage et de l'estimation de leur coût. La liste des ressources sélectionnées sera rendue publique.

Les achats d'amorçage, effectués en 2011 par l'ABES (archives Springer, corpus de dictionnaires de Classiques Garnier numérique, base EEBO – *Early English Books Online*), ont permis de rôder les procédures et la méthodologie et d'établir une licence type détaillant nos exigences vis-à-vis des éditeurs (en matière de périmètre, de droit de réutilisation des données et métadonnées, etc.). Ces premières acquisitions ont démontré la pertinence de l'achat centralisé qui engendre des économies substantielles par rapport à des achats individuels.

La plateforme

Elle proposera, sur une interface unique, les données compilées de produits acquis nationalement auprès de multiples éditeurs, mettant ainsi facilement à disposition de la communauté de l'enseignement supérieur et de la recherche un même corpus de documentation scientifique et permettant de nouveaux usages qui seraient sinon très difficiles à proposer sur le plan technique. Il est envisagé dans un second temps d'enrichir ce contenu par le moissonnage d'entrepôts tiers (revues en *Open Access*, archives institutionnelles…).
Le corpus résultant permettra par son importance la fouille de texte et pourra faire surgir de nouvelles pistes pour la recherche. Dans le domaine des sciences humaines et sociales, il devient potentiellement une source pour l'histoire, la sociologie et la philosophie des sciences et un objet de données de recherche pour la linguistique.

Les services attendus

Les ressources acquises sont de types multiples : dictionnaires, corpus, livres électroniques, articles scientifiques. Il y a donc constitution d'une masse de connaissance, structurée selon des pratiques éditoriales variables : revues scientifiques éditées il y a plusieurs dizaines d'années, livres anciens, dictionnaires structurés, corpus de textes assemblés.
L'intérêt de disposer de tels corpus en texte intégral permet d'offrir des services allant largement au-delà de la simple consultation d'items unitaires sélectionnés après interrogation d'un moteur de recherche indexant des métadonnées de type Google. En effet, la mise en ligne de ces informations en texte intégral structuré permet de développer des fonctionnalités modernes d'extraction de connaissances basées sur les technologies de la fouille de texte (*Text Mining*, *Data Mining*). Ces opérations consistent à extraire des éléments d'information d'un texte ou d'un ensemble de textes sur

la base de critères d'expression, de similitude ou de proximité de termes figurant dans le texte intégral mais pas nécessairement présents dans les métadonnées produites antérieurement.
Ces traitements d'extraction de connaissances représentent des voies extrêmement prometteuses d'accès aux textes mais permettent également de produire les référentiels terminologiques, indispensables aux mécanismes d'indexation automatique.
En ce qui concerne les articles scientifiques, l'inflation de leur nombre constatée ces dernières années conduit à penser qu'ils ne seront plus les seuls vecteurs de la transmission de la connaissance (même s'ils restent la voie normale de l'évaluation des activités de recherche) et que les méthodes de *Data Mining* seront en mesure d'offrir aux chercheurs des produits d'information nouveaux : cartes de la connaissance, *Overlays Journals*, synthèses documentaires, etc.
À ces ressources acquises auprès des éditeurs s'adjoindront celles que constituent les données de la recherche en provenance des laboratoires. L'enjeu, essentiel, est ici d'associer celles-ci aux éléments de publication qui en sont issus, offrant ainsi un *continuum* informationnel.

En résumé les services proposés seront de deux natures :
• accès vers le texte intégral d'un article ou de tout objet documentaire numérique individualisé (thèse, livre, chapitre de livre, etc.) *via* une interrogation de métadonnées. Ce service de base peut s'effectuer dans un premier temps alors que seules les métadonnées ont été fournies et chargées localement. Il sera intégré aux outils de signalement existants (SUDOC) ;
• services à valeur ajoutée basés sur le traitement des données en texte intégral comme :
- interrogation en texte intégral sur les objets numériques indexés dans leur totalité,
- production de synthèses documentaires par analyse de sous-corpus, individualisés pour l'occasion, auxquels sont appliqués des méthodes de *Text Mining*,
- services de représentation et visualisation de données basés sur les technologies de cartographie de la connaissance,
- production de corpus terminologiques,
- utilisation à des fins de recherche en ingénierie de la langue : lexicographie, morphosyntaxe, traduction, etc.
Ces services à valeur ajoutée ne peuvent être opérés qu'aux conditions suivantes : posséder les données localement, dans des formats manipulables et structurés (XML natif ou XML/PDF), et disposer des droits d'extraction et de traitement. Ce sont deux conditions majeures de l'achat de la ressource.

Un projet en deux phases

L'accès aux ressources documentaires acquises se fera, dans un premier temps, par le biais des plateformes des éditeurs (garantie dans la licence type pour cinq ans). Dans une deuxième phase (dans un délai de deux ans après le démarrage du projet) sera créée la plateforme d'accès aux ressources électroniques achetées au niveau national.

L'articulation entre le niveau national et celui des sites et des communautés

La plateforme garantira une articulation entre le niveau national et celui des sites et des communautés. Outre sa dimension nationale – une base documentaire commune issue des licences nationales – la plateforme ISTEX permettra à toutes les structures qui le souhaitent, et notamment à chaque IDEX, de développer des services propres en configurant à partir de ce socle sa propre bibliothèque numérique en fonction de ses objectifs en matière de documentation et d'IST. De même ces ressources, bien que centralisées, seront utilisées et partagées par des communautés scientifiques en réseau pour leurs besoins propres.

Les partenaires d'ISTEX

La mise en œuvre d'ISTEX est fédérative. Elle repose sur la complémentarité entre opérateurs au sein d'un dispositif national d'ensemble, cadré par l'infrastructure Bibliothèque scientifique numérique. La répartition, au sein d'ISTEX, des rôles des partenaires du projet, travaillant ensemble de longue date, est assise sur leurs domaines respectifs de compétences :
• le consortium Couperin (avec les organismes hors périmètre Couperin) pour le recueil des besoins des communautés de recherche, l'évaluation des ressources, l'établissement des listes ;
• l'ABES pour la négociation et l'acquisition des ressources, leur signalement, la gestion des accès et des droits (en partenariat avec le CNRS-INIST et en relation avec les opérateurs locaux) ;

• le CNRS, en s'appuyant sur l'INIST, pour la conception et l'hébergement de la plateforme d'agrégation du texte intégral : hébergement et exploitation des données, développement des services à valeur ajoutée, analyse de l'utilisation et des usages ;
• l'université de Lorraine pour la recherche et les services.
L'archivage pérenne des données sera assuré par le Centre informatique national de l'enseignement supérieur (CINES).

Où en est-on ?[2]

Le dossier de candidature a été déposé par le CNRS et l'université de Lorraine à l'été 2010. Après l'évaluation du jury international, l'avis du comité de pilotage puis du Commissariat général à l'investissement, une décision du Premier ministre, en date du 14 décembre 2011, a autorisé l'Agence nationale de la recherche (ANR) à contractualiser sur le projet ISTEX à hauteur de 60 M €.
Cette décision est assortie d'un certain nombre de recommandations notamment : détailler davantage les services à valeur ajoutée et les performances supplémentaires de la plateforme nationale, se référer aux projets existants au niveau international, porter attention à la compatibilité des données provenant d'éditeurs multiples et assurer l'articulation avec le dispositif national pour les acquisitions numériques (BSN) et leur signalement.

Les partenaires du projet travaillent actuellement sur le projet de convention avec l'ANR, en lien étroit avec la Mission de l'Information scientifique et technique et du Réseau documentaire (MISTRD) du ministère de l'Enseignement supérieur et de la Recherche : actualisation du projet ; réponses aux recommandations ; gouvernance, organisation et pilotage d'ISTEX ; définition des jalons et des indicateurs.

Cette infrastructure nationale pluridisciplinaire de rang mondial, indépendante des plateformes des éditeurs, va bientôt voir le jour. Elle constituera une garantie de compétitivité de nos équipes de recherche dans leur accès à l'information scientifique. La maîtrise conjointe des contenus et des technologies de traitement de l'information et de la connaissance, dans un contexte mutualisé, assurera une infrastructure opérationnelle crédible au développement des pôles d'enseignement et de recherche français.

Raymond Bérard
Directeur de l'ABES

[2] À la date du 9 février 2012.