Istex a pour ambition de renforcer la recherche et l'enseignement supérieur français sur la scène mondiale. Elle vise la réalisation d'un socle documentaire numérique pérenne pour l'ensemble de la communauté scientifique en offrant des services et des usages complémentaires et interopérables avec ceux déjà mis en place dans les établissements et organismes concernés.

Les services Istex
Istex au-delà de l'acquisition

Istex[1] a engagé, depuis 2013, une politique nationale d'acquisition massive de documentation afin d'offrir un accès en ligne aux collections rétrospectives de la littérature scientifique dans toutes les disciplines. Mais le projet a également pour objectif l'agrégation de ces ressources en une unique collection au sein d'une plateforme nationale, ainsi que la création de services apportant une plus-value basée sur le traitement des données en texte intégral.

Ces services qu'Istex se propose de mettre en place sont de deux natures : des services dits « de base » tout d'abord, avec un accès vers les textes intégraux d'articles ou vers tout objet documentaire numérique individualisé (thèse, chapitre de livre, etc.) et ce, via une interrogation de métadonnées et une indexation plein texte ; des services à valeur ajoutée, ensuite, qui s'appuient sur l'analyse de sous-corpus individualisés auxquels sont appliqués des méthodes de *text mining* basées sur le traitement des données en texte intégral.

Ces services ont pour but de permettre non seulement la production de synthèses documentaires à des fins de recherche, mais aussi des représentations de données basées sur des technologies de cartographie de la connaissance ou encore des utilisations à des fins de recherche en ingénierie de la langue (lexicographie, terminologie, 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 d'une ressource dans le cadre d'Istex.

UNE INTÉGRATION DANS LES SYSTÈMES DES ÉTABLISSEMENTS

Pour faciliter l'usage des ressources par l'ensemble des usagers, Istex propose une intégration forte dans les systèmes de diffusions existants pour éviter à l'utilisateur de devoir réaliser ses recherches sur un énième portail d'information. S'appuyant sur la plateforme de base et son interface de programmation (API), Istex offrira donc un accès aux corpus à travers les services numériques des établissements - Environnement Numérique de Travail (ENT) et *Discovery Tool* (DT). Il s'agit de proposer des éléments compatibles avec les solutions assez diverses mises en œuvre dans les établissements. Les fonctionnalités visées sont multiples: recherche sur les métadonnées et le plein texte avec affinage par facettes, accès transparent au plein texte en un minimum de clics, interface utilisateur ergonomique et intégrée aux services cibles, installation technique simple et documentée. Pour faciliter l'appropriation par la communauté scientifique, le code source développé est mis à disposition sous licence libre Cecill (licence française élaborée notamment par le CNRS et l'Inria) et partagé dans un dépôt GitHub[2]. Concernant l'intégration dans les ENT, des *widgets* compatibles avec le site d'un ENT ont été définis. Concernant les DT, l'approche est assez différente. L'index d'un *Discovery Tool* contient des métadonnées d'articles (identifiés généralement par des DOI) inclus dans des revues (identifiées généralement par des ISSN) qui sont possédées par des éditeurs. Il contient également la liste des abonnements des différents établissements qui correspondent à des bouquets de revues. Cet index devra donc être complété par les métadonnées de la plateforme Istex puis dédoublonné avec les articles déjà présents dans l'index au moment de la génération des résultats. Pour que les résultats soient intégrés de façon homogène avec leurs autres souscriptions, les établissements devront donc, en accord avec leurs prestataires dans le cas d'outils commerciaux, déclarer le bouquet Istex dans leurs abonnements. Une première étape a consisté à réaliser cette intégration à l'Université de Lorraine tout en associant à la réflexion d'autres partenaires universitaires afin, à terme, de proposer des solutions de même type sur les diverses plateformes des établissements.

DES SERVICES D'EXPLOITATION DU PLEIN TEXTE

Outre la recherche sur les métadonnées[3] descriptives des collections et des articles et l'indexation plein texte, trois services « de base » supplémentaires sont en cours de finalisation.

- **Recherche de termes et de leurs variantes**

Il convient de déterminer dans le texte initial les

[1] Le projet Istex (Initiative d'excellence pour l'information scientifique et technique) s'inscrit dans le programme « Investissements d'avenir » initié par le ministère de l'Enseignement supérieur et de la Recherche. Ce projet (ANR-10-IDEX-0004-02) est porté par quatre partenaires : le CNRS, l'Abes, Couperin et l'Université de Lorraine.

[2] https://github.com/istex

[3] L'équipe du hub de métadonnées (Abes) travaille à la normalisation et à l'enrichissement de ces métadonnées d'éditeurs acquises dans le cadre d'Istex.

séquences textuelles les plus susceptibles d'être de bons candidats termes dans le domaine scientifique correspondant au document analysé. Un groupe de travail réunissant les compétences de l'équipe Traitement Automatique du Langage Naturel (TaIn) du Laboratoire d'Informatique de Nantes Atlantique (Lina) et de l'Inist vise la réalisation de cet objectif afin d'assurer, dans des domaines spécialisés, la détection et le balisage de termes et de leurs variantes en plein texte, ainsi que le maintien d'un référentiel de terminologie scientifique pour l'exploitation des données.

• Recherche sur les entités nommées

Cela nécessite de détecter, de normaliser et de baliser de telles entités nommées dans le plein-texte. Une équipe réunissant les compétences du Laboratoire d'Informatique (Li) de Tours et de l'Inist est en charge de cet aspect. Par « entités nommées » on entend, les dates, les noms de lieux (villes, régions, pays), les noms d'individus ou de groupes d'individus (nom d'équipe, de laboratoire ou d'institution). On pourrait y ajouter les adresses web de ressources ou de données, les noms de projets liés à une ressource ou cités dans une publication. Dans un domaine spécialisé, cela pourrait être beaucoup plus fin : noms des astres en astronomie, des molécules en chimie, des plantes en botanique, etc.

• Accès aux principaux champs des références bibliographiques

Un balisage automatique préalable de ces informations dans les références bibliographiques des articles est en cours de réalisation à l'Inist. De tels accès permettront d'offrir de nouvelles facettes de sélection de ressources et faciliteront l'hypernavigation d'un article à un autre. Pour les sous-domaines, il sera possible de construire des cartographies scientifiques et de répondre à des questions du type : qui travaille avec qui ? Quels sont les réseaux de citations existants ? Quels sont les vecteurs de publication les plus marquants ? Quelles sont leurs évolutions au cours du temps ? Etc. Les résultats de ces trois types de traitement viennent ainsi enrichir les métadonnées initiales des articles et vont permettre d'adjoindre de nouvelles facettes de sélection dans la vaste collection Istex.

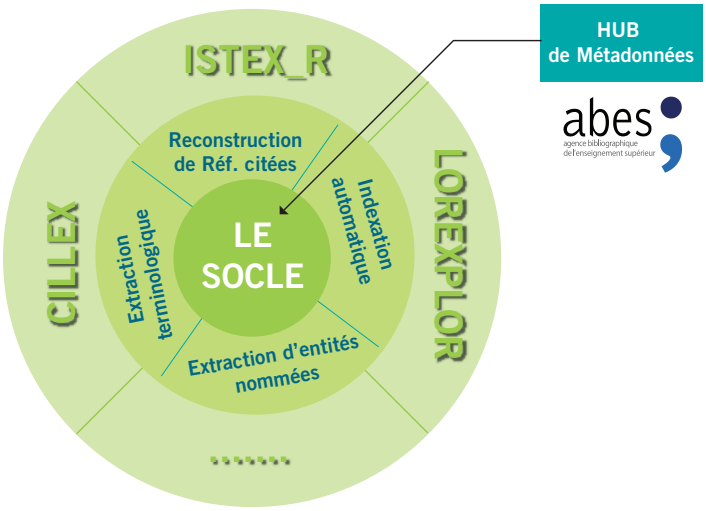
DE VÉRITABLES SERVICES À VALEUR AJOUTÉE

Enfin, reposant sur ces services « de base », trois types de services à valeur ajoutée sont développés dans le cadre de trois projets (Cillex, Istex-R et Lorexplor).

• Un moteur de réponse offrant des outils de classification automatique

Compte tenu du volume de données d'Istex, les résultats de requêtes pourront correspondre à des collections très importantes de documents qui devront être classifiées.

CARTOGRAPHIE DES PROJETS ISTEX



Le projet Cillex, mené par le laboratoire CLLE (Cognition, Langues, Langage, Ergonomie) de Toulouse, souhaite répondre à ce besoin. Son but est de développer des outils de métrologie, fondés sur des structures de réseaux « petits mondes » omniprésentes dans les bases documentaires, pour mieux identifier l'information pertinente et caractéristique d'un document dans une collection.

• La caractérisation de l'évolution des recherches et des connaissances dans le temps grâce à la construction de cartes diachroniques

Le projet Istex-R, porté à Nancy par le Loria (Laboratoire Lorrain de Recherche en Informatique et ses Applications), l'Atilf (Analyse et Traitement Informatique de la Langue Française) et l'Inist, vise à créer des outils d'accès aux contenus textuels pour construire et capitaliser des connaissances sur un domaine scientifique. Il a pour objectif de compléter la plateforme de base par une analyse du contenu afin de caractériser l'évolution des recherches et des connaissances dans le temps.

• Une bibliothèque *open source* de composants XML d'exploitation du corpus ISTEX

Le projet Lorexplor (Exploration des enjeux et besoins de l'université de Lorraine sur l'exploration des connaissances), se propose de constituer une bibliothèque *open source* de composants XML pour construire, en interaction avec des wikis sémantiques, des systèmes de recherche, des serveurs d'exploration de corpus et des chaînes de curation de données sur des sous-corpus. L'objectif est de faciliter la construction de plateformes intermédiaires (régionales, thématiques ou institutionnelles) analysant de 100 000 à 1 000 000 documents préalablement sélectionnés sur la plateforme Istex et cela, pour des besoins spécifiques d'analyse (requête particulière, bibliographie, méta-analyse, etc.)

Jean-Marie Pierrel
Université de Lorraine
jean-marie.pierrel@atilf.fr