ISTEX : de la plateforme de référence à l'infrastructure de recherche

En mettant à disposition près de 26 millions de documents, Istex est aujourd'hui le plus vaste réservoir d'archives scientifiques au service de la recherche française, proposant un usage documentaire pour la consultation de documents, et un usage plus avancé de fouille de textes pour l'exploitation et le traitement de lots de documents.

Né d'une volonté nationale, le projet Istex (Initiative d'excellence de l'information scientifique et technique) s'inscrivait dans le programme « Investissements d'avenir », initié alors par le ministère de l'Enseignement supérieur et de la Recherche. L'idée était d'acquérir massivement des collections rétrospectives de la littérature scientifique dans toutes les disciplines et de se doter d'un outil innovant d'exploitation des données. En s'inspirant du modèle de la Fondation allemande pour la recherche (DFG) qui avait amorcé une démarche d'indépendance envers les éditeurs, ce projet visait un accès pérenne aux publications via une plateforme hébergée sur le territoire national, afin de gagner une certaine autonomie vis-à-vis des éditeurs scientifiques, souverains jusque-là en matière d'accès aux publications.

Quatre acteurs principaux, reliés par un accord de consortium, ont mis en œuvre, chacun avec un rôle spécifique, ce projet doté à sa création le 19 avril 2012 d'un budget de 60 millions d'euros. Le CNRS était porteur du projet, l'Inist avait pour mission de développer l'infrastructure matérielle et logicielle, le Consortium universitaire de publications numériques (Couperin) avait comme mission principale le recueil des besoins et les négociations avec les éditeurs, tandis que l'Agence bibliographique de l'enseignement supérieur (Abes) prenait en charge les acquisitions et le signalement des collections dans les outils documentaires nationaux. Quant à la Conférence des présidents d'université (aujourd'hui France Universités), représentée par l'université de Lorraine, elle avait pour rôle de faire le lien avec les communautés de recherche, en pilotant notamment les projets de services à valeur ajoutée et les chantiers d'usage.

UN PROJET EN DEUX ÉTAPES

La première étape a consisté en une politique volontariste et massive d'achats centralisés d'archives scientifiques sous forme de licences nationales. Celles-ci ont été déterminées en fonction des besoins recensés dans les différentes communautés notamment *via* une enquête de grande ampleur à laquelle quelque 7 000 professionnels de la recherche ont répondu. Un comité de pilotage représentatif de l'ensemble des communautés a ensuite validé les choix et hiérarchisé les priorités d'acquisitions en veillant aux équilibres disciplinaires. S'appuyant sur l'expérience de consortia étrangers, l'Abes et le consortium Couperin ont mené les négociations avec les éditeurs, dans le cadre de contrats d'acquisition innovants.

Portée par l'Inist, la seconde étape du projet était la création de la plateforme destinée à héberger l'ensemble des données, construite en méthode Agile en lien avec les partenaires et utilisateurs.

Un autre choix a été fait, et pas des moindres : celui de ne pas créer d'interface mais plutôt de s'intégrer dans les systèmes existants. Cela a commencé par des widgets, intégrés dans les portails documentaires des établissements, avant de devenir un bouton Istex visible sur les plateformes utilisées par les chercheurs. Depuis mars 2021, les ressources Istex sont accessibles via l'extension unifiée *Click & Read* installable sur les principaux navigateurs Internet.

BONIFIER LES DONNÉES POUR LA FOUILLE DE DONNÉES

Les données reçues n'étant pas toujours de qualité optimale pour l'exploitation par les utilisateurs finaux, un des plus gros défis a été de les nettoyer pour les homogénéiser et les rendre ainsi aptes à être « ingérées ». Pour cela, des feuilles de style ont été créées afin de structurer les données.

Un *workflow* a été mis en place en étroite collaboration avec l'Abes pour les échanges avec les éditeurs et la restructuration des métadonnées mises à disposition par ceux-ci.

Le premier chargement de données s'est déroulé en 2014 avec 6 millions de documents. Le processus s'est ensuite généralisé : enquête, négociation, livraison et chargement pour proposer aujourd'hui plus de 25,5 millions de documents provenant de 32 sources différentes.

En parallèle, des étapes d'enrichissement des données se sont mises en place pour ajouter de nouvelles métadonnées telles que des entités nommées, des références bibliographiques structurées, une indexation ou encore une catégorisation par domaine scientifique.

Grâce aux services Istex qui ont été développés, il est possible d'explorer, d'analyser des données et de faire de la fouille de textes. L'API Istex permet de faire de la recherche documentaire (facilitée grâce à la revue de sommaires : **https://revue-sommaire.istex.fr**), les résultats étant téléchargeables de façon massive avec Istex-DL. Lodex intervient ensuite pour l'exploration et la visualisation des corpus. Sans oublier Data.Istex qui regroupe des exemples de corpus prêts à l'emploi.

Récemment, Istex s'est affiché parmi les 108 infrastructures retenues dans la feuille de route nationale des Infrastructures de recherche 2021, dans la catégorie projet, éditée par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Les principaux objectifs stratégiques exposés pour ce projet d'infrastructure sont :

- Ouvrir la collection aux ressources nativement publiées en accès ouvert et poursuivre son alimentation grâce à une politique d'acquisition ambitieuse
- Faciliter la constitution de corpus cohérents et enrichis, directement exploitables pour du TDM
- Promouvoir le développement de services avancés avec la communauté des chercheurs en TAL
- Offrir des services d'exploration et d'exploitation de corpus accessibles à tous.

Outre le caractère novateur de la réalisation technique, Istex a ouvert la voie à de nouvelles collaborations entre des acteurs de l'IST mutualisant leurs efforts et compétences au service de la communauté ESR. Il est aussi une ressource pour de la fouille de données grâce à la mise à disposition de textes intégraux documentés par des métadonnées riches et téléchargeables massivement.

Alexandra Petitjean-Monnin
Chargée de communication, Inist-CNRS
alexandra.petitjean@inist.fr

Raluca Pierrot
Responsable du service Documentation électronique, Abes
pierrot@abes.fr

Cécilia Fabry
Responsable communication, Inist-CNRS
cecilia.fabry@inist.fr