Depuis 2013, la politique nationale d'acquisition de documentation confiée à Istex a mis à la disposition de l'ESR un ample gisement documentaire. Les outils pour y accéder et l'exploiter s'intègrent directement dans les systèmes numériques des établissements. Visite dans les arcanes de l'API Istex, sur laquelle reposent ces interfaces et fonctions adaptées aux utilisateurs.

L'API Istex : *le sésame pour accéder aux ressources acquises*

ActuaLitté/Flickr (CC BY 2.0)

Espace commun de démonstration HAL-Istex-Persée lors du Salon du livre 2015.

À la question « Comment accède-t-on aux ressources Istex ? », la réponse la plus courante est : « par la plateforme Istex », ou encore « par l'API Istex », cette dernière étant disponible depuis de nombreux mois à l'adresse **api.istex.fr**. Mais cette réponse est incomplète, voire inadaptée pour qui souhaite simplement consulter ou imprimer un document scientifique. Dans les faits, le personnel d'un établissement peut accéder aux ressources Istex de multiples manières – par son portail d'établissement, ou au moyen d'une extension pour Firefox, par exemple –, sans jamais interroger directement l'API elle-même. Dès lors, on peut reformuler la question autrement : comment ceci est-il rendu possible ? En quoi une API permet et facilite la diffusion de ressources ? Et quels sont les divers modes d'accès ?

UNE BRIQUE SIMPLE, STANDARDISÉE ET OUVERTE

En français, API *(Application Programming Interface)* signifie « interface de programmation applicative » : un système ou un ensemble de composants permettant d'interagir avec un service. Dans notre cas, le service est la plateforme Istex, qui propose un ensemble de fonctionnalités pour accéder aux ressources acquises auprès des éditeurs. Se voulant universel, ce service est disponible sur Internet à travers le protocole standard HTTP.

Comme de nombreux services « orientés ressources », l'API Istex se base sur une architecture REST[1] et utilise un format simple et léger : JSON[2].

Ce service web est donc une « brique » simple, standardisée et ouverte[3], qui va servir de base à tout site web ou à tout logiciel qui souhaite donner à l'utilisateur final un accès aux ressources Istex.

L'API Istex gère un grand nombre de fonctionnalités courantes : contrôle d'accès, recherche de documents avec possibilité de filtres, tris et facettes, choix du format de la réponse, récupération des documents dans leur forme d'origine, normalisée ou enrichie[4], export d'un ensemble de résultats, etc. Toutes ces fonctionnalités peuvent donc être proposées dans les interfaces de l'utilisateur final.

[1] REST est une architecture logicielle souvent utilisée dans les API web pour sa simplicité et son adéquation avec le protocole HTTP.

[2] JSON (JavaScript Object Notation) est un format léger, permettant de stocker des objets typés (booléens, textes, tableaux…).

[3] Sa documentation complète est disponible sur **api.istex.fr/documentation**

[4] Références bibliographiques, catégorisations, etc.

ISTEX DANS LES ÉTABLISSEMENTS

À ce jour, plusieurs établissements pilotes proposent déjà l'accès aux ressources Istex, notamment les universités de Lorraine, de Rennes et de Saint-Étienne. Les personnels de ces universités peuvent utiliser la partie documentaire de leurs portails pour rechercher parmi les ressources Istex et accéder aux documents en texte intégral. Ceci a été rendu possible par une intégration de l'API, complètement transparente pour l'utilisateur.

Ces intégrations ont été réalisées à l'aide de *plugins* génériques, notamment pour les logiciels Drupal et uPortal. Ces briques logicielles étant disponibles librement, elles peuvent être réutilisées par d'autres établissements qui disposeraient des mêmes types de portail.

Quant aux établissements dont les sites ne s'appuient sur aucun logiciel ou espace numérique de travail particulier, ils peuvent intégrer l'accès aux ressources Istex grâce à des *widgets* web génériques et paramétrables – une opération qu'un webmestre réalise facilement en insérant quelques lignes de code informatique.

À plus long terme, les établissements ayant souscrit à un outil de découverte (EBSCO, OCLC…) pourront sélectionner les bouquets Istex et afficher un lien vers le plein texte hébergé sur la plateforme Istex.

LE PLEIN TEXTE À PORTÉE DE MAIN

Deux fonctionnalités de l'API Istex combinées entre elles facilitent l'accès aux documents sans bouleverser les habitudes des usagers.

• **Identification** : depuis peu, l'accès à l'API Istex peut se faire au moyen de l'authentification délivrée par le fournisseur d'identités de son établissement (par exemple, Janus, pour le CNRS). En effet, l'API s'appuyant sur la fédération d'identités Éducation-Recherche de Renater, l'ensemble de la communauté scientifique bénéficie dès à présent de cette facilité.

• **Recherche plein texte :** une autre fonctionnalité intéressante de l'API est son résolveur de liens, compatible avec la norme OpenURL. À partir de métadonnées simples (titre, auteur…) ou d'identifiants standards (DOI, PMID…), elle permet de savoir si un document est présent dans la base de documents Istex et, si oui, de faire un rebond vers le plein texte.

Comment ces fonctionnalités sont-elles utilisées par le grand public ? En premier lieu, par les extensions pour les navigateurs web Chrome et Firefox. Disponibles sur GitHub et sur les boutiques d'applications, ces extensions analysent les pages web visitées en recherchant des identifiants standards DOI, PMID et PII. Une fois la disponibilité dans la plateforme Istex vérifiée, un lien vers le plein texte PDF s'affiche. L'authentification par fédération d'identités est automatiquement demandée en cas de besoin. L'installation de ces extensions se fait en quelques clics, ce qui nous permet d'élargir le spectre d'utilisateurs Istex.

De manière similaire, des liens vers la plateforme Istex sont établis depuis le moteur de recherche Google Scholar. Dès lors que l'utilisateur a sélectionné Istex dans la partie « liens vers les bibliothèques » de ses paramètres, des liens OpenURL vers le résolveur de l'API Istex sont automatiquement affichés pour tous les documents ayant été déclarés à Google[5] comme faisant partie du « bouquet Istex ». Comme pour les extensions, si la résolution aboutit, l'utilisateur est redirigé vers le plein texte PDF, avec les mécanismes de contrôle d'accès présentés précédemment.

TEXT & DATA MINING

Une spécificité d'Istex en tant qu'archive documentaire est la possibilité d'utiliser le texte intégral comme matière première pour des travaux de recherche, en lui appliquant par exemple des techniques de *text & data mining*. Il a fallu pour cela que l'API soit facilement « moissonnable » par les chercheurs.

Plusieurs actions ont été menées en ce sens. La première a été de mettre à la disposition de la communauté un outil de moissonnage automatique. Écrit en langage NodeJS, il est disponible sur GitHub

[5] Cette déclaration se base sur les fichiers KBART fournis par la base Bacon.

TROIS ÉQUIPES POUR UNE PLATEFORME

Le développement de la plateforme est réalisé à l'Inist-CNRS de Vandoeuvre-lès-Nancy, où le travail est mené par trois équipes aux objectifs complémentaires et coordonnés.

• **ISTEX-DATA** : vérification et curation des données livrées par les éditeurs en amont de la mise en ligne. Son expertise des formats lui permet également de mettre au point les transformations XSLT vers les formats pivots MODS et TEI. L'équipe est également en train de bâtir une « chaîne d'OCRisation » (reconnaissance optique de caractères), qui permettra d'ajouter ou d'améliorer le plein texte « brut » des documents, matière indispensable pour les travaux de text & data mining.

• **ISTEX-API** : mise à disposition des documents. Ce travail s'articule autour de trois aspects : une « chaîne d'ingestion » qui prépare, reformate et indexe les données, une API qui expose les documents sur Internet, et des outils qui facilitent l'exploitation de l'API (l'extension pour Firefox, par exemple).

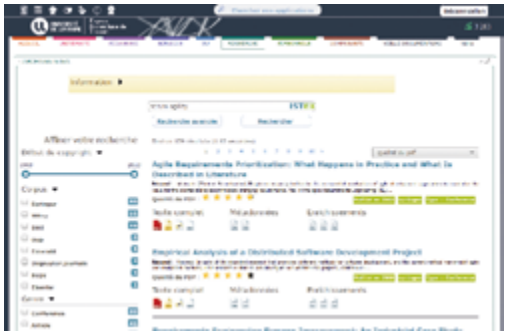
• **ISTEX-RD** : recherche et développement pour l'amélioration des données. L'objectif est de mettre au point des outils permettant d'enrichir les données initiales en repérant ou extrayant de nouvelles informations. Les données produites (références bibliographiques ou entités nommées, par exemple) sont ensuite reversées dans l'API et mises à disposition de la communauté.

Ces trois équipes travaillent en étroite collaboration, en s'appuyant sur la méthode agile Scrum. Elles ajustent en permanence leurs objectifs en fonction des retours venant des utilisateurs.

et s'utilise en interface « ligne de commande ». Ses nombreuses options permettent de cibler un sous-corpus de la même manière qu'on le ferait en interrogeant l'API – filtrer sur l'ensemble des champs, trier, choisir les formats de sortie – ou encore de limiter le nombre de documents souhaités. Ce script fonctionne sur l'ensemble des systèmes d'exploitation et a déjà été utilisé pour extraire plusieurs millions de documents.

Pour ceux qui ne sont pas familiers avec la ligne de commande et qui souhaitent récupérer des volumétries inférieures à 10 000 documents, l'API Istex propose également une fonctionnalité d'export au format Zip. Tout comme avec l'outil précédent, l'utilisateur peut spécifier finement sa requête et choisir les formats souhaités. Les utilisateurs les plus techniques pourront, quant à eux, réaliser eux-mêmes leurs scripts ou programmes de moissonnage dans leur langage favori. Un partage avec l'ensemble de la communauté est d'ailleurs vivement apprécié.

Claude Niederlender
Responsable du projet Istex-API
Inist-CNRS
claude.niederlender@inist.fr



TUTORIELS ET CODE SOURCE

Une série de tutoriels à destination des webmestres ou des utilisateurs les plus techniques a été réalisée par le service formation de l'Inist. Interactifs, ils permettent d'apprendre à l'aide de cours et d'exercices, de manière ludique, à interroger l'API. Ils sont disponibles sur le site de l'Inist.

Dans une optique de partage et d'émulation de la communauté, le code source d'un certain nombre d'outils spécifiques d'Istex a été mis en ligne sur le site GitHub à l'adresse github.com/istex. On y retrouve notamment les plugins pour Drupal et uPortal, les extensions pour navigateurs, les scripts de moissonnage de l'API, les widgets pour sites web ou le code source d'un mini-site de démonstration. Toute contribution est bienvenue.

À sept mois du clap de fin, Istex approche les 20 millions d'articles

La politique d'acquisition de ressources documentaires à l'échelon national dans le cadre du projet Istex, porté par le CNRS, Couperin, l'Université de Lorraine et l'Abes, a donné lieu à la constitution d'un fonds documentaire unique, mis à disposition à la fois sur les plateformes des éditeurs et *via* l'API Istex.

Le programme d'achats de licences nationales a débuté par des bases de données qui recèlent des documents à valeur patrimoniale. Les documents concernés sont de natures hétérogènes. Ce sont d'abord les grammaires et les dictionnaires de Classiques Garnier Numérique, qui permettent un voyage dans l'histoire de notre langue. Ce sont aussi les deux bases complémentaires, Early English Books Online (EEBO) et Eighteen Century Collection Online (ECCO), qui offrent une vue globale sur les publications de langue anglaise – mais pas seulement – sur plus de trois siècles. Les collections rétrospectives de revues scientifiques : un autre grand axe que le projet Istex poursuit en privilégiant la profondeur de collection et l'adossement aux abonnements actuels des bibliothèques. Par la signature de douze contrats avec des éditeurs de stature internationale comme Springer, Wiley, Cambridge University Press, Institute of Physics ou encore Emerald, tous les établissements français ayant des missions d'enseignement supérieur et de recherche accèdent aux archives de plus de 8 000 revues dans tous les domaines disciplinaires : médecine, chimie, physique, économie, mais aussi sociologie, histoire, droit, linguistique, etc. La couverture chronologique a été étendue le plus possible pour se rapprocher des publications courantes.

Mais Istex, c'est aussi des *e-books* : les achats réalisés à ce jour (10 000) portent sur des titres en sciences, en chimie, en science politique et en droit international.

Les mois qui restent jusqu'à la fin du projet – prévue au 31 août 2017 – seront employés pour parfaire cet ensemble, notamment en augmentant le contenu francophone et en renforçant le corpus en sciences humaines et sociales. Les derniers achats seront réalisés au début de l'été.

Carole Melzac
Responsable du service Achats documentation électronique, Abes
carole.melzac@abes.fr

POUR EN SAVOIR PLUS

Vous pouvez retrouver toutes les informations sur chaque ressource sur le site LicencesNationales.fr