ISTEX : une nouvelle corde à son ARK

Pascale Viot, Nicolas Thouvenin

OUTILS SERVICES

Les documents issus des bouquets acquis dans le cadre des négociations ISTEX et disponibles sur la plateforme possèdent un identifiant technique interne de 40 caractères, l'ID ISTEX. Ils ont été enrichis depuis peu par un nouvel identifiant normé, pérenne, gratuit et distribué : l'ARK.

ISTEX : *une nouvelle corde à son ARK*

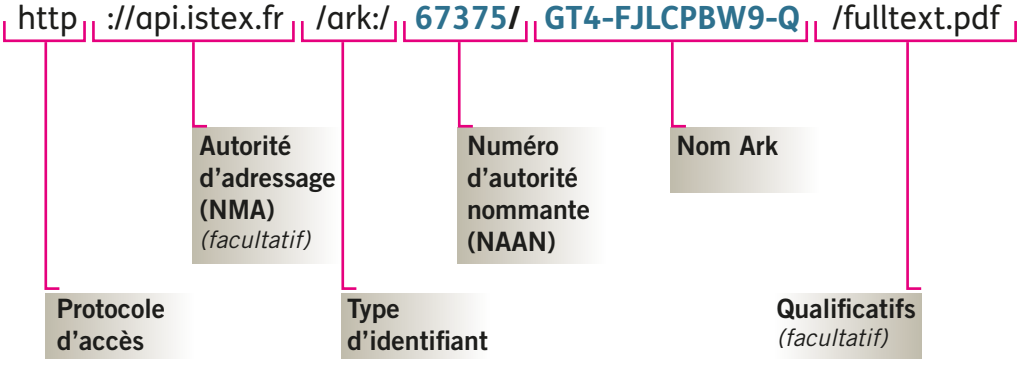
La réflexion autour de la normalisation des identifiants ISTEX a démarré fin 2015 lors d'une immersion au sein de la BnF. L'objectif était de mieux appréhender la démarche d'attribution et de pérennisation des identifiants des ressources numériques de la BnF. En 2016, la norme ARK (Archival Research Key) était présentée à l'Institut de l'information scientifique et technique (Inist) par Sebastien Peyrard en vue de son implémentation dans les données ISTEX. En effet, bien que le fonds documentaire ISTEX soit composé d'objets possédant majoritairement un DOI pointant vers le document éditeur, l'usage des ARK offre la possibilité d'identifier les documents du fonds, ce qui évite la confusion entre l'identifiant de l'objet d'archive et celui du document original. En outre, par rapport au système *Handle,* l'approche décentralisée, gratuite et sans contrainte technique proposée par le système ARK constitue un avantage déterminant pour s'intégrer au mieux dans une plateforme technique préexistante.

QU'EST-CE QU'UN ARK ?

Il s'agit d'un système d'identifiants basé sur la norme URI, initié, mis en place et maintenu par la CDL[1] et intensivement utilisé par la BnF. On accède au document grâce à une URL ARK composée de deux parties : une première, constituée du protocole d'accès associé à l'adresse du site, qui peut donc être amenée à changer au cours du temps ; une seconde, constituée de l'ARK proprement dit, partie fixe et pérenne composée d'un numéro d'autorité attribué par la CDL, d'un nom ARK et d'un qualificatif de service facultatif.

ORGANISATION DES ARK DANS UN ÉTABLISSEMENT

Un ARK est constitué d'un numéro, le NAAN (Name Assigning Authority Number), attribué par la CDL permettant d'identifier l'institution habilitée (67 375 pour Inist-CNRS). Celle-ci s'engage à garantir l'unicité et la pérennité des identifiants qu'elle produit. Pour y parvenir, la CDL conseille d'introduire la notion de préfixe – ou *subpublisher,* constitué d'une suite de 3 caractères alphanumériques, incrémenté automatiquement sur la base d'un algorithme.
Le *subpublisher* caractérise toutes les ressources d'un même jeu de données, d'un même projet, d'un même service. Il permet de définir un périmètre fonctionnel, dans lequel il est possible de créer et d'assurer une unicité des identifiants produits. L'Inist-CNRS a donc développé un registre central[2] pour mémoriser les *subpublishers* et garantir leur unicité et la cohérence du système d'identification. Chaque *subpublisher* est caractérisé par quatre éléments :



[1] California Digital Library

[2] https://github.com/Inist-CNRS/ezark

[3] http://search.cpan.org/~jak/Noid/noid#NOID_CHECK_DIGIT_ALGORITHM

• un nom (de projet, d'application, de service…),
• un sujet (le nom d'un dépôt, d'une étude, d'un jeu de données),
• une description libre,
• l'URL du service proposant l'accès aux ressources.

Outre le *subpublisher,* le nom ARK est suivi d'un identifiant opaque non séquentiel de 8 caractères alphanumériques et d'un caractère de contrôle permettant d'assurer la validité de l'ARK. L'Inist-CNRS a repris le même algorithme que celui proposé par la CDL, le NCDA *checksum algorithm*[3]. Chaque élément (*subpublisher,* identifiant, caractère contrôle) est séparé par un tiret.

ATTRIBUTION AUTOMATIQUE

Bien qu'il existe un logiciel *open source* (NOID, Nice Opaque Identifiers) qui génère des identifiants et donc potentiellement des ARK en utilisant un paramétrage spécifique, un outil a été développé en interne. Ce développement *ad hoc*[4] se justifie pour les raisons suivantes : le choix d'un nom ARK en trois parties, l'existence d'une plateforme technique d'injection, les besoins de stockage et de sauvegarde de tous les ARK générés, soit plusieurs millions.
À partir du registre central, l'attribution des ARK est réalisée automatiquement : pour les documents, au cœur même de la plateforme ISTEX ; pour les référentiels documentaires, dans l'application Lodex[5].
Avant d'attribuer des ARK aux 19 millions de documents, une première phase a consisté à travailler sur les référentiels documentaires liés aux documents ISTEX. À partir d'une méthode de publication des référentiels, l'attribution des ARK a été expérimentée sur les différentes catégories de classification des documents ISTEX. La méthode a ensuite été validée et généralisée sur plusieurs référentiels consultables et citables au travers du site https://data.istex.fr via leur ARK, comme, par exemple :

Extrait du référentiel des types de documents :
https://content-type.data.istex.fr/ark:/67375/XTP-94FB0L8V-T
Extrait du référentiel des catégories Inist-CNRS de documents : https://inist-category.data.istex.fr/ark:/67375/RZL-8WV8N6BQ-7

PROTOTYPE ET MISE EN PRODUCTION

Une fois la méthode éprouvée sur les référentiels, au printemps 2017, une équipe composée de documentalistes et d'informaticiens a développé un prototype d'attribution et d'accès aux documents ISTEX via le protocole HTTPS, associé à l'adresse de la plateforme ISTEX, suivi de la partie pérenne de l'ARK. Il a été décidé d'utiliser plusieurs *subpublishers* : chaque bouquet éditeur ayant été enregistré dans le registre central, les documents ISTEX ne possèdent donc pas tous le même préfixe. Quant au qualificatif, il identifie le document selon sa typologie (fulltext.pdf, fulltext.tei…). Compte tenu de la masse de documents, la mise en production des ARK est réalisée progressivement au fur et à mesure des mises à jour de la plateforme.

CITABILITÉ FACILITÉE, LISIBILITÉ AMÉLIORÉE

L'attribution d'ARK aux objets documentaires offre de nombreux avantages. C'est tout d'abord une norme, utilisée par de nombreuses institutions publiques, qui assigne des identifiants pérennes de façon gratuite et avec une liberté de pratique pour l'autorité nommante. La citabilité est facilitée par une chaîne de caractères plus courte que l'ID ISTEX actuelle. La lisibilité est améliorée par une hiérarchisation bien identifiée dans le Nom ARK. La pérennité, quant à elle, est assurée en interne, contrairement à certains identifiants tel le DOI.
Ce nouveau type d'accès vient en complément de l'accès par ID ISTEX, mais ne le remplacera pas. Ainsi, pour l'utilisateur ayant déjà cité un document ISTEX, l'accès peut se faire sous les deux formes, comme, par exemple :
https://api.istex.fr/document/087661D669BF44CA05AA6CE08ADD6399F6A439C4/fulltext/pdf et
https://api.istex.fr/ark:/67375/GT4-FJLCPBW9-Q/fulltext.pdf
Actuellement, tous les corpus ISTEX sont enregistrés dans le registre de subpublisher (code de 3 caractères), un code de 8 caractères étant en cours d'attribution pour les documents issus de chacun des corpus. Il est donc possible de citer un document avec cette nouvelle URL sachant que la partie la plus courte de l'URL, sans les qualificatifs, permet de connaître l'ensemble des typologies et formats possibles pour un même document :
https://api.istex.fr/ark:/67375/GT4-FJLCPBW9-Q
Demain, grâce à la structure hiérarchisée et l'utilisation des qualificatifs, il sera possible non seulement de citer une notice en mods ou en xml, un fulltext en txt ou en pdf, mais également une page, voire une illustration du document.

Pascale Viot
Équipe Plateforme ISTEX, Inist-CNRS
pascale.viot@inist.fr

Nicolas Thouvenin
Responsable du service R&D, Inist-CNRS
nicolas.thouvenin@inist.fr

[4] https://github.com/Inist-CNRS/node-inist-ark

[5] https://github.com/Inist-CNRS/lodex/

POUR EN SAVOIR PLUS

Sur la norme ARK, on peut consulter les présentations d'Emmanuelle Bermès, « Des identifiants pérennes pour les ressources numériques : l'expérience de la BnF », International Preservation News, IFLA-PAC, 2006 ; (40) : 16-26, et de Sébastien Peyrard, « The ARK Identifier Scheme: General Characteristics and Implementation at the National Library of France », Workshop on Persistent Identifiers, Köln Universität,: Projet DASISH; 2014. Sur la méthode de publication des référentiels, on peut lire, de Cécilia Fabry et al, Sept étapes pour publier des données ouvertes et liées , I2D : information, données et documents : pratiques & recherche, 2017, pp.12-14. Et aussi : **https://api.istex.fr/documentation/ark, http://blog.istex.fr/des-ark-dans-istex, http://lodex.inist.fr/tag/ark, et www.bnf.fr/fr/professionnels/issn_isbn_autres_numeros/a.ark.html**