ISTEX
L'excellence documentaire pour tous

ANR-10-IDEX-0004-02

L'intégration des ARK dans ISTEX en quelques questions

Nicolas Thouvenin
thouveni

CNRS/Inist

QU'EST-CE QUE ISTEX ?

ISTEX

L'excellence documentaire pour tous

"Construire le socle de la bibliothèque scientifique numérique nationale."

ANR-10-IDEX-0004-02

WWW.ISTEX.FR

www.licencesnationales.fr

"ISTEX plus de ~~18,5~~ 21 **millions de documents** provenant de **19 éditeurs** plus de **7 500 titres** et **13 000 ebooks** entre 1406 et 2015."

DES MILLIONS DE DOCUMENTS

- à réceptionner
- à analyser
- à enrichir
- à reformater
- à normaliser
- à ré-océriser
- à mettre en ligne
- …

⇒ à identifier

COMMENT IDENTIFIER UN "OBJET DOCUMENTAIRE" ISTEX ?

Avec un système automatique !

1. 7AF88EC8E547846C736021BF26ED71EFABB36991
2. B14CC18B27A2A5A13D87E5DE4052EE9E340C02CE
3. 4471A7AC4A966A679545A477299512E5C0C38230
4. 7E195F8F9F9A75BBB8AFDDC56A5A6B7F30E6F01B
5. A0DEFD7E09821E0A5F50D31FA70B86B7F960B493
6. DF0E8799A961756E13EF36BDB170C2A30E2C3D44
7. DB7FB95E6ABD3065F754578ACF4BE00257286346
8. C0CE73FE3B5B1D1EBDC9A5B3CFC079391F36CCF5
9. 5E780B4F23B9A4E198EC8742064ED77015B5D872
10. 9E66D0A322F35469F4BC12AC03C9961FFF2800DE

OU BIEN AVEC UN SYSTÈME NORMALISÉ

- DOI,
- HANDLE
- URI
- ARK
- …

oui, mais avec lequel ...

POURQUOI CHOISIR LE SYSTÈME ARK?

Analyse de nos contraintes

- Présence d'un DOI sur 96% des documents ISTEX
  - ~~DOI~~
- Une architecture informatique existante (non java)
  - ~~HDL~~
- Un système d'identifiant ad-hoc existant faiblement normalisé
  - ~~URI~~

Analyse de nos besoins

- Pouvoir s'intégrer facilement dans l'existant
  - ~~HDL~~
- Éviter une dépendance avec un système externe
  - ~~HDL, DOI~~
- Identifier avec le même système des données différentes (documents, référentiels, etc.)
  - ~~ADHOC~~

Analyse des avantages du système ARK

- système décentralisé
- sans contrainte technique
- dissocier d'un type de données

QUELLE AUTORITÉ NOMMANTE CHOISIR ?

Choix du NAAN

- ISTEX est un **projet** et non une **institution**
- L'Inist-CNRS est une **unité** du CNRS (UPS 076)
- Le CNRS est l'un des **acteurs** du projet avec l'ABES, l'université de Lorraine, Couperin et le ministère

Qui a autorité sur les identifiants ISTEX ?

Pour ISTEX

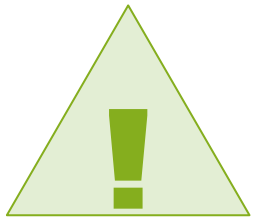
l'Inist-CNRS est chargé :

- de la mise en oeuvre de la plateforme
- de réceptionner les documents ISTEX
- ...

⇒ **il est l'autorité nommante**

```
naa:
who:    Institut de l'information scientifique et technique (=)
Institute for scientific and technical information (=) INIST
what:   67375
when:   2016.03.30
where:  http://www.inist.fr
how:    NP | (:unkn) unknown | 2016 |
```

Pour le CNRS



l'Inist-CNRS est chargé :

- de la mise en oeuvre d'autres plateformes
- de traiter des documents non ISTEX
- de traiter des données différentes (terminologie, etc.)
- ...

Le NAAN Inist ne peut pas être associé uniquement à ISTEX

Comment gérer une sous autorité nommante ?

ISTEX
L'excellence documentaire pour tous

REGISTRE

Un **registre unique** à l'Inist
partagé par tous les projets et services
pour **enregistrer & conserver** les préfixes.

> un fichier Excel partagé !

ou via une application similaire...

Registre officiel des subpublishers ARK pour l'Inist-CNRS

Cette interface centralise les subpublishers ARK (prefix) du NAAN de l'Inist-CNRS (67375). Ce registre permet de générer, recenser et garantir l'unicité de ces subpublishers. Chaque équipe ou projet de l'INIST peut ainsi venir déclarer un subpublisher pour ensuite venir associer des ARK aux données gérées par une plateforme donnée (ex : plateforme ISTEX ou plateforme Loterre).

NAAN

67 375

Nombre de subpublishers

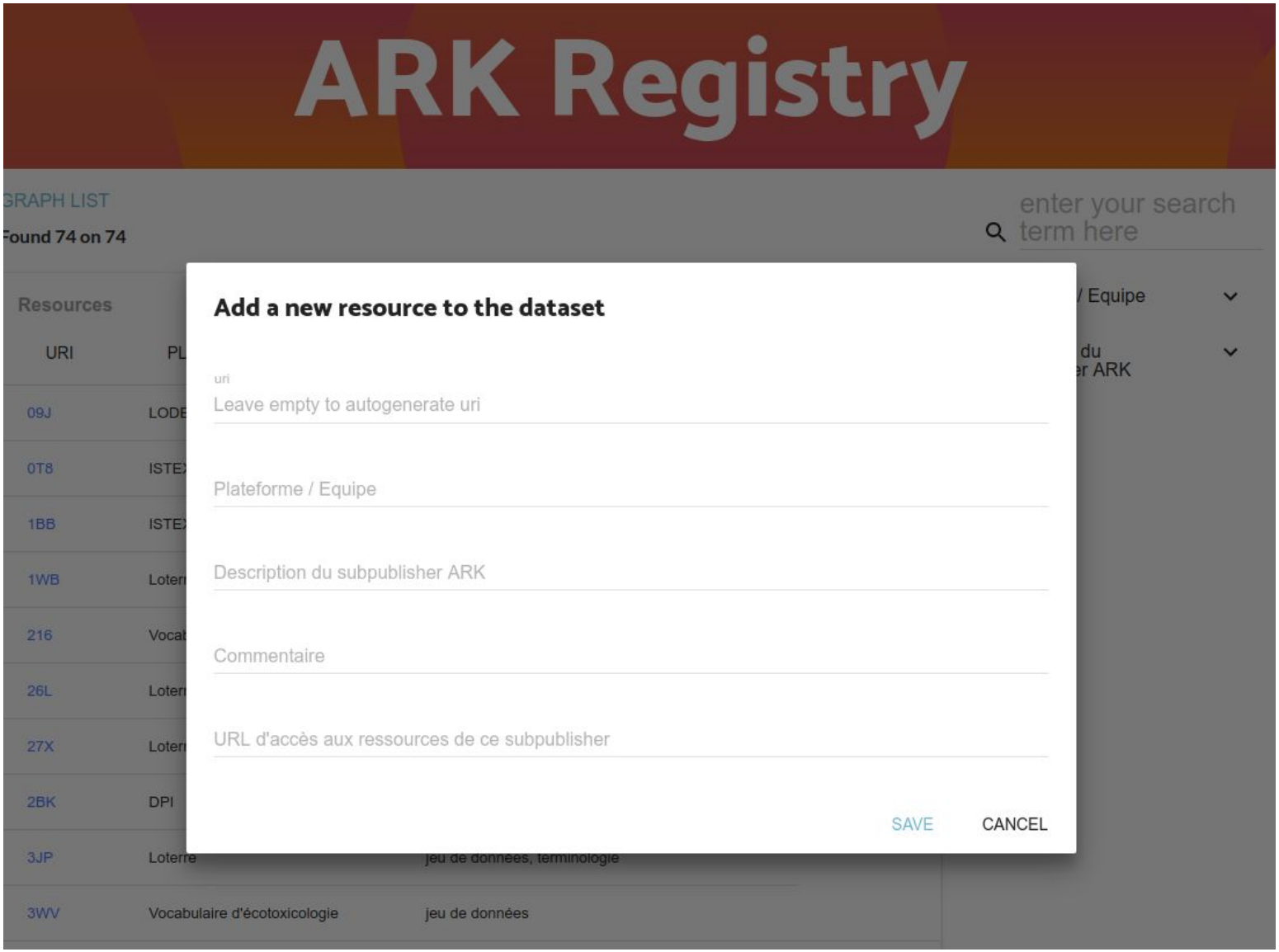
74

Liste des subpublishers déclarés

ISTEX property-dataset	ISTEX Corpus IOP	ISTEX Corpus Springer
Loterre jeu de données, terminologie	Loterre jeu de données, terminologie	DPI projets
Loterre jeu de données, terminologie	Loterre jeu de données, terminologie	ISTEX Corpus Emerald

`ezark` - https://github.com/Inist-CNRS/ezark

http://inist-registry.ark.inist.fr/


ARK Registry
GRAPH LIST
Found 74 on 74
enter your search term here
Resources
URI
Add a new resource to the dataset
uri
Leave empty to autogenerate uri
Plateforme / Equipe
Description du subpublisher ARK
Commentaire
URL d'accès aux ressources de ce subpublisher
SAVE
CANCEL
/ Equipe
du
r ARK
09J
0T8
1BB
1WB
216
26L
27X
2BK
DPI
3JP
3WV
Vocabulaire d'écotoxicologie
jeu de données
jeu de données, terminologie

UN PREFIXE DE 3 caractères :

exemple

HCB

49 préfixes pour ISTEX

ISTEX
L'excellence documentaire pour tous

ANR-10-IDEX-0004-02

Comment générer des identifiants ?

Une nomenclature

ark:/67375/**XXX**-**YYYYYYYYYY**-**Z**

- **Un préfixe**
- **Un identifiant sur 10 caractères**
- **Un caractère de contrôle - NCDA**
  NOID CHECK DIGIT ALGORITHM

UN CODE LIBRE

node-inist-ark

build passing | bitHound 99

NodeJS package used to handle "normalized" ARK for the INIST organization. This library can be used to generate a lot of random and valid ARKs dedicated to a specific NAAN and subpublisher, or to parse an existing ARK as a nice JSON object, or to validate the content of a given ARK (ex: checking this ARK as not been misspelled thanks to its checksum).

INIST's ARK anatomy is:

```
    ark:/67375/39D-S2GXG1TW-8
    \__/ \___/ \__/\______/\_/
     |     |    |     |     |
ARK Label  |    |     |     Check sum (1 char)
           |    |    Identifier (8 chars)
           |   Sub-publisher (3 chars, it has to be generated in the centralized INIST ARK registry)
           |
Name Assigning Authority Number (NAAN) (67375 is dedicated for INIST)
```

- INIST NAAN will not change and is this integer: 67375
- Sub-publisher is handled by a [centralized ARK registry for INIST](todo add the link)
- Identifier is a string of 8 uppercase characters from this alphabet 0123456789BCDFGHJKLMNPQRSTVWXZ
- Check sum is 1 character calculated from the ARK identifier following the NCDA checksum algorithm. It is used to help detecting mispelled ARK.

Install

```
npm i inist-ark
```

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

ANR-10-IDEX-0004-02

QUELS QUALIFICATIFS UTILISÉS ?

UN OBJET DOCUMENTAIRE DANS ISTEX

peut être composé de :

- Métadonnées
  - XML, MODS
- Texte intégral
  - PDF, TEI, TXT, OCR, ZIP, TIFF
- Annexes
  - PDF, TXT, DOC, JPEG, QT, MPEG, MP4, PPT, XLS, XLSX, AVI, XML, RTF, GIF, WMV
- Couvertures
  - PDF, GIF, JPEG, TIFF, HTML
- Enrichissements
  - Entités nommées, Références bibliographiques, etc.

ETUDES DES URLs dynamiques pré-existantes

TYPOLOGIE	FORMAT	dans l'url Istex	
record	MODS	metadata	mods
record	XML	metadata	xml
record	JSON		
annexes	PDF	annexes	pdf
annexes	JPEG	annexes	jpeg
annexes	TXT	annexes	txt
annexes	XLS	annexes	xls
annexes	XLSX	annexes	xlsx
annexes	GIF	annexes	gif
annexes	ZIP	annexes	zip
annexes	PPT	annexes	ppt
annexes	QT	annexes	qt
annexes	DOC	annexes	doc
annexes	MP4	annexes	mp4
annexes	MPEG	annexes	mpeg
annexes	AVI	annexes	avi
annexes	WMV	annexes	wmv
annexes	RTF	annexes	rtf
covers	TIFF	covers	tiff
covers	HTML	covers	html
covers	GIF	covers	gif
covers	PPT	covers	ppt
fulltext	PDF	fulltext	pdf
bundle	ZIP	fulltext	zip
fulltext	TEI	fulltext	tei
fulltext	TXT	fulltext	txt
fulltext	OCR	fulltext	ocr
entities	TEI	enrichments	unitex
refbibs	TEI	enrichments	refbibs
keywords-nb	TEI	enrichments	nb
keywords-multicat	TEI	enrichments	multicat
keywords-teeft	TEI	enrichments	teeft
keywords-abes	TEI	enrichments	abesSubjects
authors	TEI	enrichments	abesAuthors

TYPOLOGIE	FORMAT	dans l'url Istex	
fulltext-original	PDF	fulltext	original
record-original	XML	metadata	original
covers-original	TIFF	cover	original
annexes-original	ZIP	annexes	original
annexes-original	PDF	annexes	original
annexes-original	JPEG	annexes	original
annexes-original	TXT	annexes	original
annexes-original	XLS	annexes	original
annexes-original	XLSX	annexes	original
annexes-original	GIF	annexes	original
annexes-original	ZIP	annexes	original
annexes-original	PPT	annexes	original
annexes-original	QT	annexes	original
annexes-original	DOC	annexes	original
annexes-original	MP4	annexes	original
annexes-original	MPEG	annexes	original
annexes-original	AVI	annexes	original
annexes-original	WMV	annexes	original
annexes-original	RTF	annexes	original
covers-original	TIFF	covers	original
covers-original	HTML	covers	original
covers-original	GIF	covers	original
covers-original	PPT	covers	original

Beaucoup de combinaisons possibles

Une nomenclature simplifiée

`ark:/67375/XXX-YYYYYYYYYY-Z/`**`TYPOLOGIE.FORMAT`**

- **3 typologies :**
  fulltext, metadata, bundle
- **7 formats :**
  zip, pdf, xml, txt, json, tei, mods

ANR-10-IDEX-0004-02

QUELS FORMATS UTILISER PAR DÉFAUT ?

Problèmes des qualificatifs pour chaque document

- non obligatoire
- pas toujours disponible

⇒ Ils varient en fonction

- des données,
- des outils,
- de la plateforme

L'API ISTEX RENVOIE POUR

un ARK :

- soit le document demandé dans le format sélectionné
  - ark:/67375/VQC-ZJHCRXRV-B/fulltext.pdf
- soit la liste des formats disponibles pour la typologie sélectionnée (en JSON)
  - ark:/67375/VQC-ZJHCRXRV-B/fulltext
- soit la liste des typologies et formats disponibles pour le document sélectionné (en JSON)
  - ark:/67375/VQC-ZJHCRXRV-B

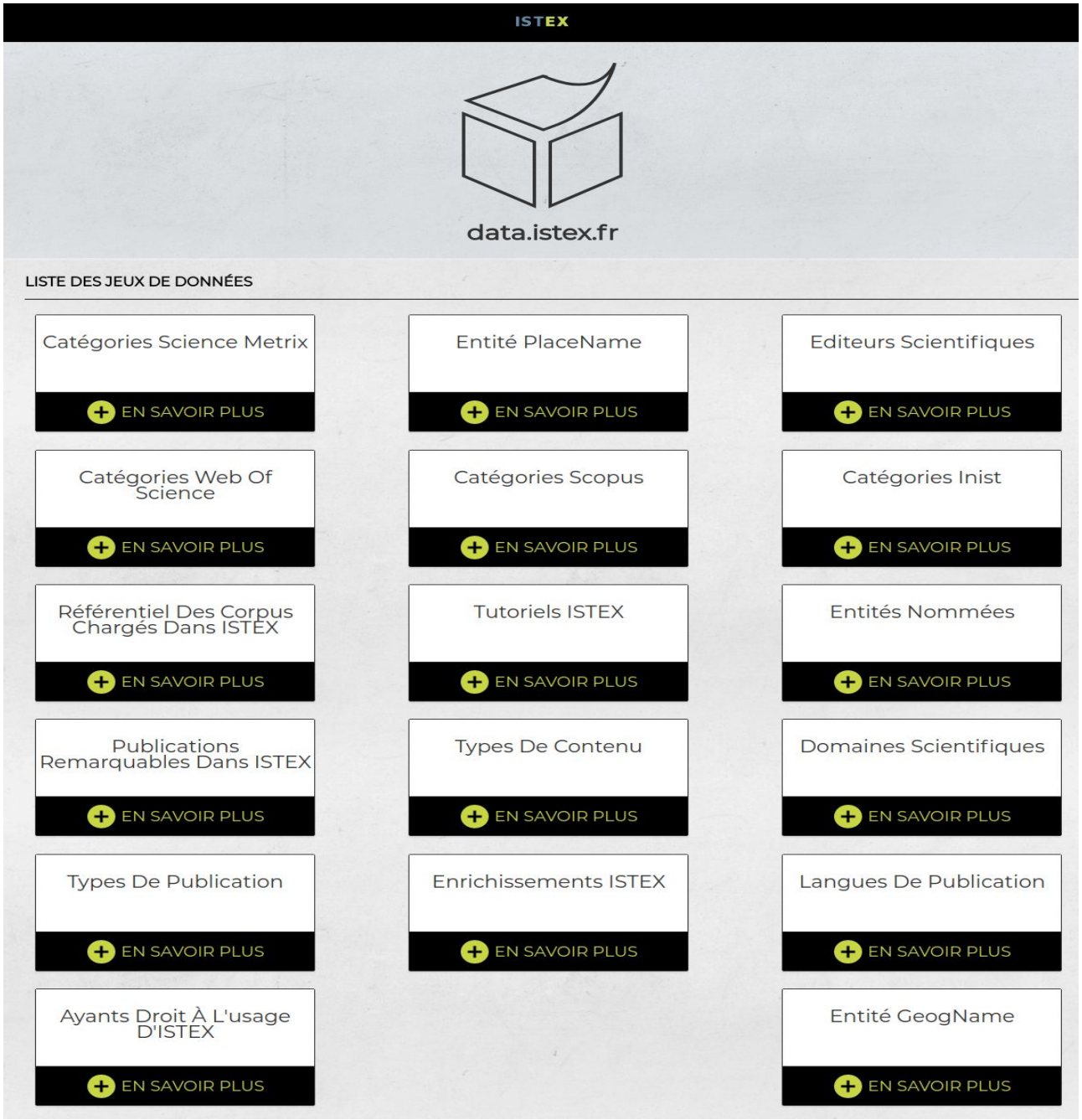
QUELLES "RESSOURCES" IDENTIFIER AVEC DES ARK ?

ALLER AU DELÀ DU PDF

Les objets documentaires ISTEX, mais aussi

- les référentiels documentaires
  - les langues, les types de document, etc.
- les données générées automatiquement
  - entités nommées, catégories, etc.

presque tout ce que l'on diffuse au travers du portail
https://data.istex.fr


ISTEX
data.istex.fr
LISTE DES JEUX DE DONNÉES
Catégories Science Metrix
EN SAVOIR PLUS
Entité PlaceName
EN SAVOIR PLUS
Editeurs Scientifiques
EN SAVOIR PLUS
Catégories Web Of Science
EN SAVOIR PLUS
Catégories Scopus
EN SAVOIR PLUS
Catégories Inist
EN SAVOIR PLUS
Référentiel Des Corpus Chargés Dans ISTEX
EN SAVOIR PLUS
Tutoriels ISTEX
EN SAVOIR PLUS
Entités Nommées
EN SAVOIR PLUS
Publications Remarquables Dans ISTEX
EN SAVOIR PLUS
Types De Contenu
EN SAVOIR PLUS
Domaines Scientifiques
EN SAVOIR PLUS
Types De Publication
EN SAVOIR PLUS
Enrichissements ISTEX
EN SAVOIR PLUS
Langues De Publication
EN SAVOIR PLUS
Ayants Droit À L'usage D'ISTEX
EN SAVOIR PLUS
Entité GeogName
EN SAVOIR PLUS

CONSÉQUENCE

https://data.istex.fr/sparql/

```
1  SELECT *
2  FROM <https://inist-category.data.istex.fr/notice/graph>
3  WHERE {
4    ?subject istex:subjectInist ?complement
5  }
6  LIMIT 100
```

Table | Raw Response | Pivot Table | Google Chart | Geo

Showing 1 to 50 of 100 entries Search: Show 50 entries

	subject	complement
1	https://api.istex.fr/ark:/67375/0T8-021XKRKC-D	http://inist-category.data.istex.fr/ark:/67375/RZL-11ZFRTC3-3
2	https://api.istex.fr/ark:/67375/0T8-021XKRKC-D	http://inist-category.data.istex.fr/ark:/67375/RZL-983FDQQ1-K
3	https://api.istex.fr/ark:/67375/0T8-021XKRKC-D	http://inist-category.data.istex.fr/ark:/67375/RZL-JNRXQT1K-R
4	https://api.istex.fr/ark:/67375/0T8-07CX3D42-7	http://inist-category.data.istex.fr/ark:/67375/RZL-11ZFRTC3-3
5	https://api.istex.fr/ark:/67375/0T8-07CX3D42-7	http://inist-category.data.istex.fr/ark:/67375/RZL-983FDQQ1-K
6	https://api.istex.fr/ark:/67375/0T8-07CX3D42-7	http://inist-category.data.istex.fr/ark:/67375/RZL-JNRXQT1K-R
7	https://api.istex.fr/ark:/67375/0T8-08H1DZ8Q-Q	http://inist-category.data.istex.fr/ark:/67375/RZL-983FDQQ1-K
8	https://api.istex.fr/ark:/67375/0T8-08H1DZ8Q-Q	http://inist-category.data.istex.fr/ark:/67375/RZL-0XQZDS3B-4
9	https://api.istex.fr/ark:/67375/0T8-08H1DZ8Q-Q	http://inist-category.data.istex.fr/ark:/67375/RZL-Q5MMC105-M
10	https://api.istex.fr/ark:/67375/0T8-0DR8N9C5-C	http://inist-category.data.istex.fr/ark:/67375/RZL-11ZFRTC3-3

Usage des ARK dans les triplets RDF du triplestore ISTEX

QUE RESTE-T-IL À FAIRE ?

BEAUCOUP !

- accès direct à certaines parties d'un objet ISTEX
- diffusion / utilisation dans les applications pré-existantes
- service de résolution ARK

- formation
- "respect" des identifiants
- politique de conservation
- (...)

ISTEX
L'excellence documentaire pour tous

INVESTISSEMENTS D'AVENIR

ANR-10-IDEX-0004-02

MERCI