casuHAL
club utilisateur

Journées casuHAL 2022

À la découverte de l'API de recherche HAL

9 et 10 juin 2022

Brigitte Bidegaray-Fesquet (CNRS, Univ. Grenoble Alpes)

À la découverte de l'API de recherche HAL

- Introduction
- Structure d'une interrogation
- Les types de champ
- Exemples
  - pour la gestion d'un portail
  - pour la gestion d'une collection
  - pour la « gestion » d'un auteur
- Les API des référentiels

Introduction

Qu’est-ce qu’une API ? Les API de HAL

Qu'est-ce qu'une API ?

- **API = Application Programming Interface**

  ou « interface de programmation d'application » :

  c'est une interface logicielle qui permet l'échange de données entre logiciels.

- **Comme nous ne sommes pas des logiciels…**

  Nous allons ici apprendre à **utiliser le point d'entrée de l'API** de recherche de HAL, c'est-à-dire à **écrire des requêtes** dans la barre de recherche d'un navigateur et à **formater les sorties** pour que nous, humains, puissions les analyser.

  Pour des résultats de recherche plus complexes, on peut exploiter les résultats en utilisant divers logiciels… mais c'est un autre atelier.

Les API de HAL

- **Il y a plusieurs API dans HAL.**

- **Nous allons nous concentrer sur l'API de recherche.**

- **Les autres API :**
  - API SWORD de dépôt sur HAL
  - Les API de recherche dans les référentiels

anrproject	doctype	metadata
author	domain	metadatalist
authorstructure	instance	structure
europeanproject	journal	

Structure d'une interrogation

Champs de recherche, champs de réponse, filtres, facettes, tri…

Structure générale d'une interrogation

- **Le point d'entrée de l'API**

http://api.archives-ouvertes.fr/search/

- **On précise ensuite**
  - le champ sur lequel on veut chercher : q
  - des filtres sur les réponses : fq
  - des facettes (détaillé plus loin)
  - les champs retournés dans la réponse : fl
  - l'ordre dans lequel on trie les réponses : sort
  - le rang des résultats retournés : start et rows
  - le format de sortie : wt (JSON par défaut)

Format d'une requête

- **Une requête est de la forme** (q comme *query* (requête) en anglais)

  ?q=<champ>:<valeur>

- Pour chercher tous les documents dont le titre contient la chaîne de caractères « Bloch » :

  http://api.archives-ouvertes.fr/search/?q=title_t:"Bloch"

- **On peut utiliser les opérateurs binaires AND (et) et OR (ou).** L'opérateur par défaut est AND.
- documents dont le titre contient les chaînes de caractères « Bloch » et « Maxwell » :

  http://api.archives-ouvertes.fr/search/?q=title_t:("Bloch" "Maxwell")

- documents dont le titre contient les chaînes de caractères « Bloch » ou « Maxwell » :

  http://api.archives-ouvertes.fr/search/?q=title_t:("Bloch" OR "Maxwell")

- **ainsi que l'opérateur unaire NOT (non).**
- documents dont le titre contient la chaînes de caractères « Bloch » mais pas « Maxwell » :

  http://api.archives-ouvertes.fr/search/?q=title_t:("Bloch" NOT "Maxwell")

Filtres

- **On filtre avec** (fq comme *filter query*)

  ?fq=<champ>:<valeur>

- documents dont le titre contient les chaînes de caractères « Bloch » et « Maxwell » :

  http://api.archives-ouvertes.fr/search/?q=title_t:"Bloch"&fq=title_t:"Maxwell"

- **L'avantage du filtre est bien sûr que l'on peut filtrer par rapport à un autre champ que la requête initiale**

- documents dont le titre contient les chaînes de caractères « Bloch », déposés entre 2016 et 2020 :

  http://api.archives-ouvertes.fr/search/?q=title_t:"Bloch"&fq=submittedDateY_i:[2016 TO 2020]

- **On peut cumuler plusieurs filtres** et donc avoir plusieurs fois &fq=…

- thèses et HDR dont le titre contient les chaînes de caractères « Bloch », déposées entre 2016 et 2020 :

  http://api.archives-ouvertes.fr/search/?q=title_t:"Bloch"&fq=submittedDateY_i:[2016 TO 2020]&fq=docType_s:(THESE OR HDR)

Facettes (1/2)

- **Pour générer des facettes, il faut ajouter le paramètre facet=true à une requête**
- Ensuite, on précise
  - le champ qui sert pour la facette

    facet.field=<champ>:<valeur>

  - le type de tri

    facet.sort=index (tri lexicographique)

    facet.sort=count (tri par nombre d'occurrence)
- répartition par domaine dans la collection LJK :

  https://api.archives-ouvertes.fr/search/LJK/?q=*:*&rows=0&facet=true&facet.field=level0_domain_s&facet.sort=count

Facettes (2/2)

- **On peut aussi préciser**
  - Le préfixe par lequel commencent les facettes

facet.prefix=<préfixe>

  - répartition par mots-clés commençant par V dans la collection LJK :

https://api.archives-ouvertes.fr/search/LJK/?q=*:*&rows=0&facet=true&facet.field=keyword_s&facet.prefix=V

  - On peut se servir de champs non multi-valués pour servir de pivot

facet.pivot=<pivot>

  - répartition par type de documents des types de dépôts dans la collection LJK :

https://api.archives-ouvertes.fr/search/LJK/?q=*:*&rows=0&indent=true&facet=true&facet.pivot=docType_s,submitType_s

Affichage (1/2)

- **On choisit le format de sortie avec**

?wt=<format>

- où on a une réponse de Apache Solr pour <format> égal à json, xml ou csv
- ou une réponse de l'API pour <format> égal à xml-tei, bibtex, endnote, rss ou atom
- **Dans les formats json, xml ou csv, les champs retournés par défaut sont docid, label_s et uri_s**

http://api.archives-ouvertes.fr/search/?q=title_t:"Bloch"

- **On peut choisir d'autres champs pour la sortie avec**

?fl=<champ>

- Mise en évidence des auteurs publiant sur Maxwell-Bloch :

https://api.archives-ouvertes.fr/search/?q=title_t:"Maxwell-Bloch"&fl=authLastName_s

Affichage (2/2)

- **Les résultats peuvent être triés** (*cf. infra* pour le choix du champ)
- Le tri se fait avec le paramètre

  ?sort=<champ> <sens>

- où <sens> vaut asc ou desc (pour un tri par ordre croissant ou décroissant).
- Tri par année de publication décroissante :

  https://api.archives-ouvertes.fr/search/?q=title_t:"Maxwell-Bloch"&sort=publicationDateY_i desc

- **Dans le cas où il y a beaucoup de résultats, on peut découper les réponses par lots**

  rows=<nombre de résultats>

  start=<rang du premier résultat>

- Publications avec le mot « Ukraine » :

  https://api.archives-ouvertes.fr/search/?q=title_t:Ukraine&sort=publicationDateY_i desc&start=100&rows=50

Les types de champ

Des champs multipliés pour des usages différents

Des champs pour tous les usages

- **Les champs sont dupliqués pour différentes utilisations :** affichage, facettes, recherche ou tri.

On les reconnaît à leur suffixe, parmi lesquels

suffixe	type	affichage	facette	recherche	tri
_bool	booléen	x	x	x	x
_fs	facette	x	x		
_i	nombre	x	x	x	x
_id	identifiant			x	
_sort	alpha (lexicogr.)				x
_s	chaîne de caractères	x	x		x
_sci	chaîne (sans casse)	x	x	x	x
_t	texte			x	
_tdate	date (ISO 8601)	x		x	x

Exemples

pour la gestion d'un portail, d'une collection, d'un auteur

Exemple pour la gestion d'un portail

- Evaluer l'évolution du dépôt en texte intégral des articles dans un portail :

https://api.archives-ouvertes.fr/search/saga/?q=docType_s:ART&rows=0&facet=true&facet.pivot=submittedDateY_i,submitType_s

Cette interrogation comprend :

- la restriction à un portail,
- la requête uniquement sur les articles,
- l'utilisation de facettes,
- et les deux pivots utiles : la date de soumission et le type de soumission.

Si pour l'utilisation par un logiciel c'est la même chose, un humain ne voit pas les résultats de la même manière si on inverse l'ordre des pivots :

https://api.archives-ouvertes.fr/search/saga/?q=docType_s:ART&rows=0&facet=true&facet.pivot=submitType_s,submittedDateY_i

Exemple pour la gestion d'une collection

- Trouver les articles co-publiés avec les Etats-Unis en 2021 :

  https://api.archives-ouvertes.fr/search/LJK/?q=instStructCountry_s:us&fq=publicationDateY_i:2021

Cette interrogation comprend :

- la restriction à une collection (*idem* portail mais les collections sont en majuscules),
- la requête uniquement sur les publications avec le pays 'us' (code ISO 3166),
- et un filtre sur la date de publication.

En utilisant des facettes, on peut lister les pays avec lesquelles il y a des co-publications pour une collection :

https://api.archives-ouvertes.fr/search/LJK/?q=*:*&rows=0&facet=true&facet.pivot=instStructCountry_s

Exemple pour la « gestion » d'un auteur

- **Pour sensibiliser au dépôt en texte intégral :**

https://api.archives-ouvertes.fr/search/?q=authIdHal_s:brigitte-bidegaray-fesquet&fq=docType_s:ART&fl=journalSherpaCondition_s,journalSherpaPrePrint_s,journalSherpaPostPrint_s,submitType_s,uri_s

Cette interrogation comprend :

- la requête uniquement sur un auteur *via* son idHAL textuel,
- un filtre sur les articles dans des revues seulement,
- l'affichage d'un certain nombre données Sherpa/Romeo de la revue
  - détail des conditions
  - autorisation des pre- et post-prints
- du type de soumission (texte intégral, notice)
- et du lien vers la notice.

API de recherche dans les référentiels

L'exemple du référentiel auteur

Exemple pour la « gestion » d'un auteur

- **Essayer d'identifier un auteur, un cas simple :**

https://api.archives-ouvertes.fr/ref/author/?q=Bidegaray-Fesquet&fl=*_s

*_s est un champ dynamique qui retourne toutes sortes d' « identifiants »

- **Un cas plus compliqué qui permet d'y voir plus clair :**

https://api.archives-ouvertes.fr/ref/author/?q=lastName_s:Picard&fq=firstName_s:C*&fl=*_s

Certaines requêtes plus compliquées peuvent nécessiter l'utilisation de plusieurs référentiels ou de l'API de recherche et d'un référentiel et dans ce cas là il est difficile de combiner les recherches sans logiciel adapté.

Vous avez envie d'en savoir plus ?

La documentation de l'API :

https://api.archives-ouvertes.fr/docs/search

La documentation des référentiels :

https://api.archives-ouvertes.fr/docs/ref

Requêtes sur les ressources de HAL (sur le wiki) :

https://wiki.ccsd.cnrs.fr/wikis/hal/index.php/Requêtes_sur_les_ressources_de_HAL