Produire des données de recherche enrichies et interopérables avec PACTOLS et Opentheso

Formation MASA, 22 septembre 2021





Blandine NOUVEL, Centre Camille Jullian (Aix Marseille Univ, CNRS) Miled ROUSSET,



Maison de l'Orient et de la Méditerranée (Univ Lyon 2, CNRS)



Objectifs de la formation

Comprendre les enjeux de l'enrichissement des données de recherche à l'aide d'un vocabulaire de référence interopérable.

Exemple PACTOLS:

- l'intégrer à sa démarche scientifique,
- l'appliquer à ses outils grâce aux fonctionnalités du gestionnaire de thésaurus Opentheso.

Plan

- 1. Contexte : Métadonnées, ouverture et partage des données de la recherche
- 2. PACTOLS, thésaurus pour l'archéologie, normalisé et interopérable
- Bien choisir les mots-clés pour qualifier ses données, principes de base et méthode de l'indexation
- 4. PACTOLS dans la chaîne de traitement des données : les 4 niveaux d'intégration
 - a. consulter et copier/coller le concept dans son propre réservoir
 - b. associer son propre vocabulaire et PACTOLS, le rôle des URI
 - c. lier le thésaurus à son système : exploiter l'API et son widget
 - d. valoriser l'interopérabilité : réciprocité des liens et requête de retour
- 5. Conclusion : participer à l'évolution du thésaurus



Je suis..., et j'ai des projets,... j'ai besoin de, j'ai envie de...

1. Les métadonnées dans l'ouverture et le

partage des données

Ouvrir et partager les données

Un contexte international, un engagement national

- 2018 : lancement du "Plan National pour la Science Ouverte"
- 2021 : 2e PNSO, axé sur les données de la recherche et les bonnes p
 - accès ouvert aux publications
 - structurer, partager, ouvrir les données de la recherche
 - ouvrir les codes sources des productions scientifiques
 - faire évoluer les pratiques et l'évaluation



Pour les données :

S'engager dans une démarche FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable)

Respecter le principe « Aussi ouvert que possible, mais aussi fermé que nécessaire »

Métadonnées : définition et rôle

Métadonnées / métadonnées descriptives : "informations structurées sur tout ce qui peut être nommé comme des pages Web pages, livres, articles scientifiques, images, chansons, produits, processus, personnes (et leurs activités), données de la recherche, concepts, des services". (trad. de https://www.dublincore.org/resources/metadata-basics/)



Métadonnées : structure

La description suit un schéma structuré, conforme (ou pas) à des standards internationaux ou disciplinaires.

Par exemple, le Dublin Core :

- Standard le plus répandu, obligatoire dans certains protocoles d'échange de données.
- Composé de 15 éléments de base
 + des éléments supplémentaires et des éléments "de raffinement" concernant certains éléments de base (date, titre, relation, format, droits, couverture) et des éléments supplémentaires (audience, identifiant)

L'utilisation de métadonnées standardisées en général est un enjeu pour le respect des principes FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable)

Nom de l'élément en	Nom de l'élément en	Décrit
anglais	français	Decrit
Title	Titre	
Subject	Sujet	
Description	Description	
Source	Source	contenu
Language	Langue	
Relation	Relation	
Coverage	Couverture	
Creator	Créateur	
Publisher	Éditeur	propriété
Contributor	Contributeur	intellectuelle
Rights	Droits	
Date	Date	
Туре	Туре	instance
Format	Format	particulière
Identifier	Identifiant	

Ouvrir et partager mes les données que je produis

utiliser des outils ouverts

- . production : logiciels non propriétaires et interopérables
- . stockage, archivage : entrepôts et archives institutionnels

soigner la qualité des données

- . veiller à l'homogénéité des jeux de données
- . organiser la validation et le contrôle des données
- . choisir une licence ouverte : CC-BY ou Etalab

consolider les métadonnées

- . utiliser des formats de description reconnus : DC,...
- . exploiter les référentiels ; IdRef, PACTOLS

s'appuyer sur les infrastructures









2. PACTOLS thésaurus et réservoir de métadonnées

thématiques pour l'archéologie

Le thésaurus, vocabulaire contrôlé

- Un ensemble de mots-clés normalisés liés par des relations sémantiques complexes adapté à un domaine de la connaissance
- Les mots-clés sont validés
- Un outil terminologique et documentaire destiné à faciliter la recherche informatique dans des réservoirs de ressources
 - il apporte des informations sur le contenu de la ressource/donnée
 - il évite les ambiguïtés du langage courant
- Vient en appui aux métadonnées de type SUJETS

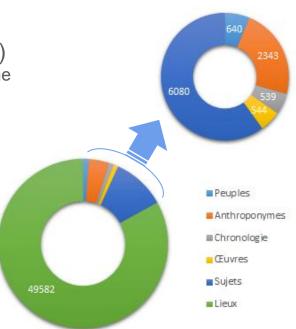
PACTOLS, thésaurus pour l'archéologie

Contient (pour la version publique)

- 60000 concepts multilingues(fr, en, it, es, de, nl, ar)
 - o sur l'archéologie, de la Préhistoire à la période contemporaine
 - o et les sciences de l'Antiquité
- 6 collections
 - + 1 liste gérée hors du thésaurus

Produit et administré

- avec le gestionnaire Opentheso
- par un GT + experts scientifiques



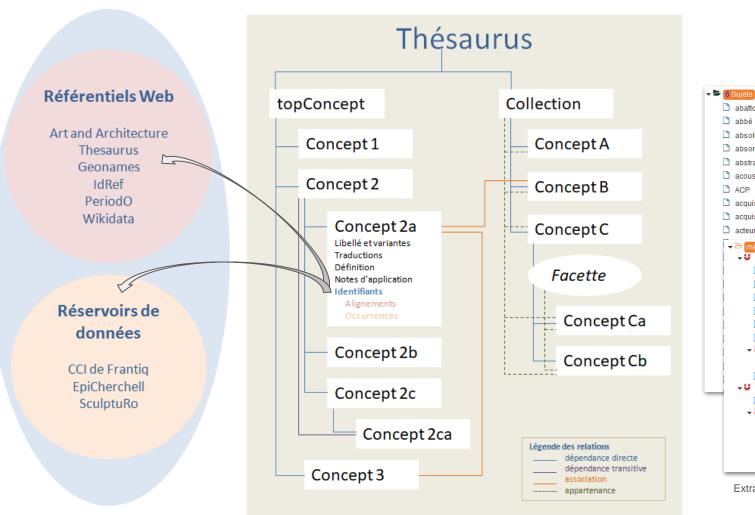
Répartition des concepts dans les collections PACTOLS, janvier 2019

PACTOLS, vocabulaire normalisé

ISO 25964 sur la structure et l'interopérabilité des thésaurus

SKOS, Simple Knowledge Organisation System pour les éléments relatifs aux concepts et à leurs relations

- des mots-clés ou concepts = éléments de description
 - o libellés : prefLabel, altLabel, hiddenLabel
 - identifiant : idPactols + Idark
- des groupes de concepts = collections et facettes
- des relations sémantiques internes = structure hiérarchiques, associations
- des relations sémantiques externes = alignement avec d'autres réservoirs en ligne



absolu absorption atomique abstraction divinisée acoustique ACP acquisition de ressource naturelle acquisition des données acteur → U par spécialité psychanalyse paléopathologie ophtalmologie neurologie médecine vétérinaire médecine légale → Chirurgie Chirurgie dentaire gynécologie → U par tradition médecine arabe → Description médecine antique médecine assyro-babylonienne médecine égyptienne médecine grecque médecine romaine

abattoir sacré nabbé

Extrait de Pactols v2

PACTOLS, référentiel terminologique pour l'archéologie



thématique/mot-clé : clés méthodologiques de l'indexation

3. Bien choisir la métadonnée

L'indexation documentaire, méthode QCoCoJuLi

QUOI ? adopter le bon point de vue : on indexe le contenu de la ressource (dc:subject) pas le ressource, celle-ci étant décrite avec les autres métadonnées du format (dc:title, dc:creator, dc:date, etc.)

COMMENT ? "lire, comprendre" la ressource, être capable d'en faire un résumé/analyse. De quoi ça parle, que représente-elle, quel est le sujet ?

- Utiliser les éléments d'information : titre/légende, TdM, introduction/conclusion,...
- Se méfier des mots du titre, choisis pour inciter à la lecture mais pas toujours explicites voire trompeurs sur le contenu

COMBIEN ? 1 à 10 mots-clés, souvent plus en archéologie (thématique, chronologie, géographie,...)

JUSQU'OÙ ? opter pour un niveau de précision adapté :

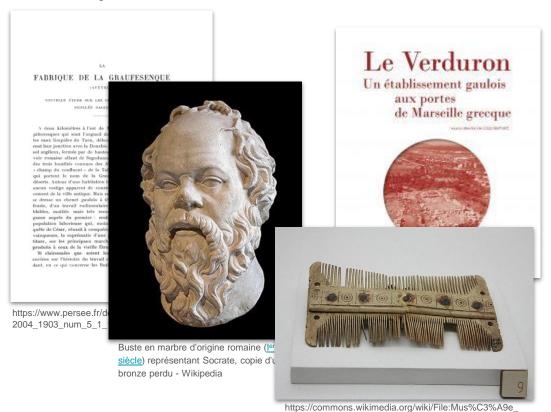
- industrie lithique, burin ou burin d'angle ?

La suite de mots-clés doit permettre de connaître le contenu sans en avoir pris connaissance.

Contrôler son indexation avec une LECTURE INVERSÉE: "d'après les mots-clés, le document parle de..."



S'adapte à toute ressource: texte, objet, image, son



Saint-Raymond - Toilette 06.jpg

Fouilles de l'épave de la Madrague de Giens, 1972-1982 THE AMERICAN PROPERTY AND ADDRESS OF THE Mary on a training When the same married in particular of Francisco and native contained deposits on the containing about the other sections. L. office periods from the west of the work to be to be the extrait des notes manuscrites de Ph. Leveau sur la

fouille de la tombe 1A nécropole occidentale de

Cherchell



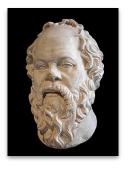
Indexer ces documents

FABRIQUE DE LA GRAUFESENQUE

NOUVELLE ÉTUDE SUR LES ORIGINES DE LA POTERIE SIGILLÉE GALLO-BONAINE

A deux kilomètres à l'est de Millau, dans un de ces sites pittoresques qui sont l'orgueil des montagnes du Bouergue, les eaux limpides du Tarn, débouchant du Gausse noir, opèrent leur jonction avec la Dourbie. C'est là, dans une vallée au sol argileux, fermée par de hautes murailles rocheuses, que la voie romaine alliant de Segodunum à Lodève rencontrait une des trois localités connues des Ruifnes, le Condotomagus ou « champ du confluent » de la Table de Peutinger. Ces lieux, qui portent le nom de la Graußesenque, sont sujourd'hui déserts. Autour d'une habitation isolée, perdue dans les terres, aucon vestige apparent de construction ne marque l'empla-cement de la ville antique. Mais sur un de ses murs de clôture se dresse un chenet gaulois à tête de bélier, bloc de pierre fruste, d'un travail rudimentaire. Deux autres chenets sem-blables, mutilés mais très reconnaissables, gisent dans le habbles, mutilie mais très reconnissables, giornt dans le gama nagrès des premiers : antès d'aves des foyres d'une gama nagrès des premiers par les des presents de la con-quirite de Criare, rémais à computire eller même, aux dépress des vaisqueures, les appendents d'une grante industrie, et à misqueures, les principaux marchés de Franjare, ses propres de l'appendent de l'appendent de l'appendent des l'appendent de Si chiarendes que soient les informations des auteurs anciens sur l'histoire du tuvuil ser Gaube, nons avons ceptur-dant, ne ce qui concerne les fluidiers, quelques témologiques

https://www.persee.fr/doc/rea_ 0035-2004 1903 num 5 1 1298



Buste en marbre d'origine romaine (ler siècle) représentant Socrate, copie d'un bronze perdu -Wikipedia



https://ark.frantig.fr/ark:/2 6678/cci725036



Fouilles de l'épave de la Madraque de Giens. 1972-1982



extrait des notes manuscrites de Ph Leveau sur la fouille de la tombe 1A nécropole occidentale de Cherchell



https://commons.wikimedia.org /wiki/File:Mus%C3%A9e Saint -Raymond_-_Toilette_06.jpg

4. Intégrer PACTOLS dans la gestion de ses données

Présentation du gestionnaire Opentheso Les 4 niveaux d'utilisation : consultation, mapping, ajout dynamique, retour de requête

Opentheso, gestionnaire de thésaurus

Un outil original

- Logiciel de gestion de thésaurus
 - multilingue
 - conforme à la norme ISO-25964
 - collaboratif
 - full web
- Développé à la MOM par la plateforme Web Sémantique et Thesauri
- Diffusé en open-source sur Github
- Intégré à la grille de services de la TGIR Huma-Num
- Partenaires :

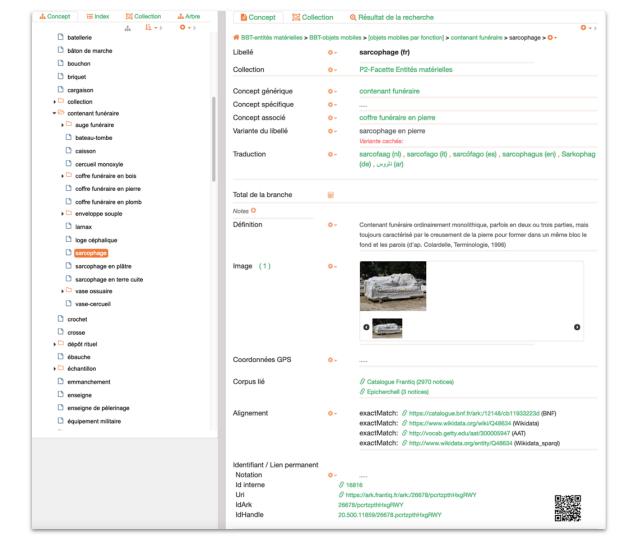




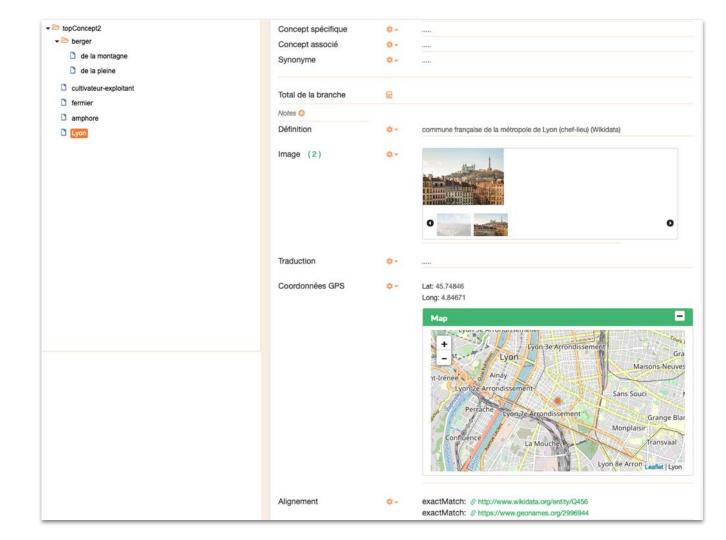




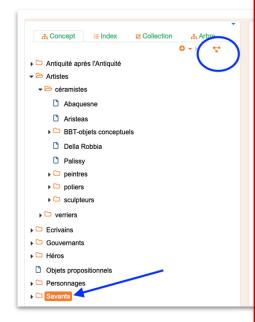
Fiche concept

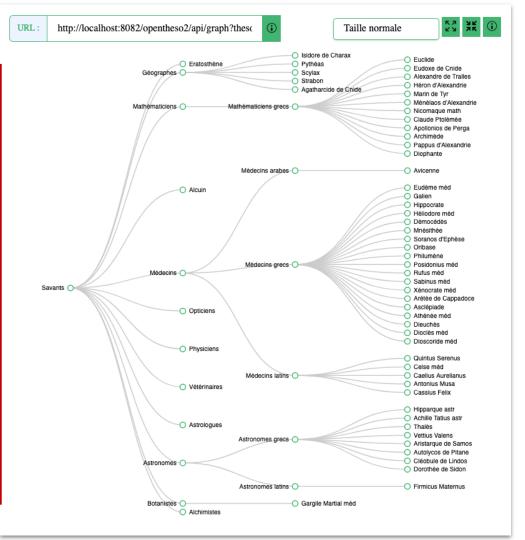


Gestion d'images et de cartes GPS



Les vues





Export des thésaurus (basé sur le frameWork RDF4J)

Export en RDF (SKOS, Json, JsonLd et Turtle)



Export en CSV en respectant un format défini



Exploitation des thésaurus et contrôle de l'indexation

Opentheso dispose d'une ARI complète de type webservices **REST** : **Opentheso 3**koha nd-ArcheoGRID nuxeo 💿 telemeta **WORDPRESS**

du manuel à l'interopérabilité complète

Les 4 niveaux d'utilisation :

Du plus basic au plus sophistiqué,

1. Consulter le thésaurus

/1

Utiliser la version en ligne : https://pactols.frantiq.fr

NB. La version en développement, libre à la consultation uniquement sur https://pactols2.frantiq.fr

Méthode:

- 1.consulter PACTOLS,
- 2.sélectionner le-s concept-s pertinent-s
- 3.le-s copier/coller dans son système



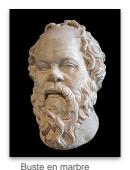
Indexer ces documents avec PACTOLS



PETERS OFFICE HER LISS ORDERED HE LA POTEN-

A was Manistere I for the Willes, show on the soul formation and complete the contragent of Manistere and Registeries and Regi

https://www.persee.fr/doc/ rea_0035-2004_1903_num_5_1_12 98



d'origine romaine (<u>ler siècle</u>) représentant Socrate, copie d'un bronze perdu - Wikipedia



https://ark.frantiq.fr/ark:/2 6678/cci725036



Fouilles de l'épave de la Madrague de Giens, 1972-1982



extrait des notes manuscrites de Ph. Leveau sur la fouille de la tombe 1A nécropole occidentale de Cherchell



https://commons.wikimedia.org /wiki/File:Mus%C3%A9e_Saint -Raymond_-_Toilette_06.jpg

1. Consulter le thésaurus

/2

Utiliser la version en ligne : https://pactols.frantiq.fr

NB. La version en développement, libre à la consultation uniquement sur https://pactols2.frantiq.fr

Méthode:

- 1.consulter PACTOLS,
- 2.sélectionner le-s concept-s pertinent-s
- 3.le⋅s copier/coller dans son système

Avantages



- Facilité d'accès
- Facilité et rapidité de mise en oeuvre
- Pas de connaissance informatique particulière

Inconvénients



- Risques d'erreurs dans le report des mots-clés
- Répétition des copier/coller
- Pas de suivi dans la mise à jour du thésaurus
- Aucune interopérabilité

2. Exploiter les idArk des concepts pour gérer l'interactivité /1

Utiliser des mots-clés d'un référentiel en exploitant leur identifiant pérenne.

Rapprocher les items de sa liste et les concepts de PACTOLS :

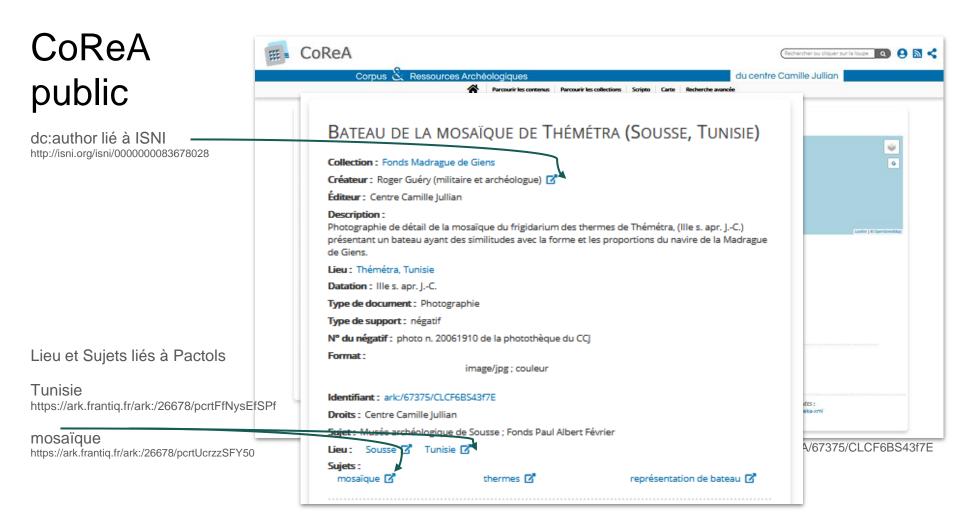
- ajouter une URI/id Ark PACTOLS à chacun de ses items. Chaque item est cliquable et renvoie à PACTOLS
- conserver (ou pas) son propre libellé
- intégrer la liste (ou pas) dans son système de gestion de données

Un exemple avec une base de données Omeka Classic

CoReA, Corpus & Ressources en Archéologie

- des collections thématiques
- des liens vers les référentiels ISNI et PACTOLS
- lien vers le thésaurus et navigation interne



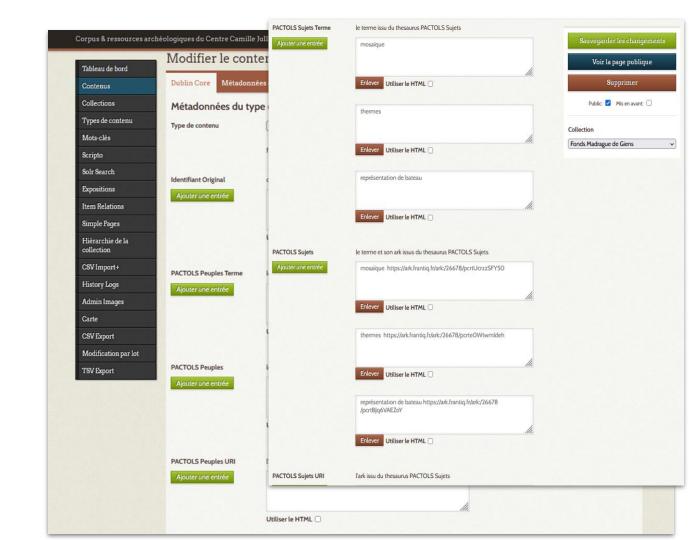


CoReA en mode pro

Définition du Type de contenu = PACTOLS

3 champs / mot-clé dans les Métadonnées du type de contenu

- Terme : nomme le concept dans la ressource = 'mosaigue'
- Concept (libellé + idARK) : lie le Terme au concept dans Pactols = 'mosaïque'
- URI : permet de basculer dans le thésaurus



2. Exploiter les idArk des concepts pour gérer l'interactivité /2

Utiliser des mots-clés d'un référentiel en exploitant leur identifiant pérenne.

Rapprocher les items de sa liste des concepts de PACTOLS :

- ajouter une URI/id Ark PACTOLS à chacun de ses items. Chaque item est cliquable et renvoie à PACTOLS
- conserver (ou pas) son propre libellé
- intégrer la liste (ou pas) dans son système de gestion de données

Pour CoReA:

Avantages

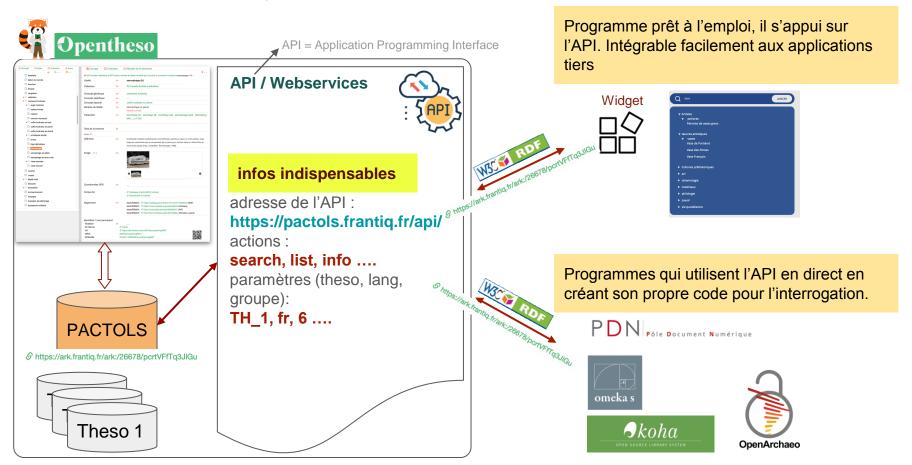
- Vocabulaire lié associé
- Adaptation possible des libellés/concept d'autorité
- suit les mises à jour du thésaurus
- Pas de connaissance informatique particulière

Inconvénients



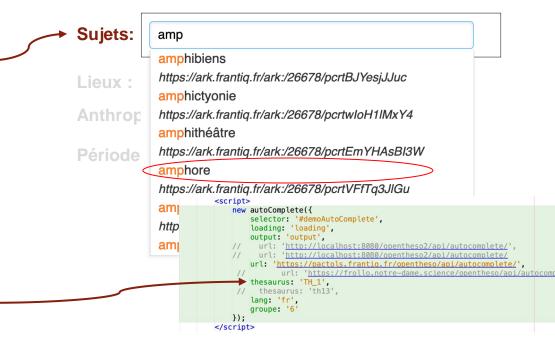
- Multiplication des saisies
- Ajout des id Ark à chaque mot-clé
- Consultation externe du thésaurus
- Interopérabilité limitée

3. API et ajout dynamique



3. Exploiter la liaison dynamique via le widget JavaScript

- Pour le besoin d'indexation dans les systèmes type (CMS, portail de diffusion, base de données...), on peut lier chaque champ de saisie de type métadonnées avec Opentheso via le widget
- la saisie devient alors semiautomatique, l'assistance se fait par un système d'autocomplétion offert par le widget
- il est facilement paramétrable pour s'adapter au bon thésaurus et à la bonne collection/groupe



Test et téléchargement du widget sur : https://github.com/miledrousset/Opentheso2/tree/master/src/main/webapp/scriptAutoComplete

3. Exploiter la liaison dynamique via le widget JavaScript

Avantages

- pas d'erreur de saisie
- assistance dans la recherche des termes
- garantie l'interopérabilité
- mise en place sans connaissance informatique

Inconvénients

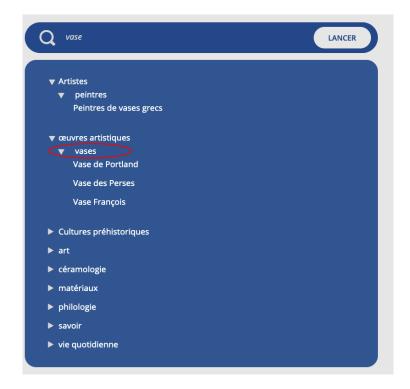
• l'évolution de l'API implique la mise à jour du Widget

Démo Widget + OmekaS

A noter : un deuxième Widget est en cours de finalisation

Il est bien plus complet et offre la possibilité de :

- naviguer dans le thésaurus sans passer par Opentheso mais par son API
- rechercher un terme
- afficher la hiérarchie complète vers le terme trouvé
- afficher à la demande les informations complètes pour chaque concept (définition, formes alternatives, etc.)

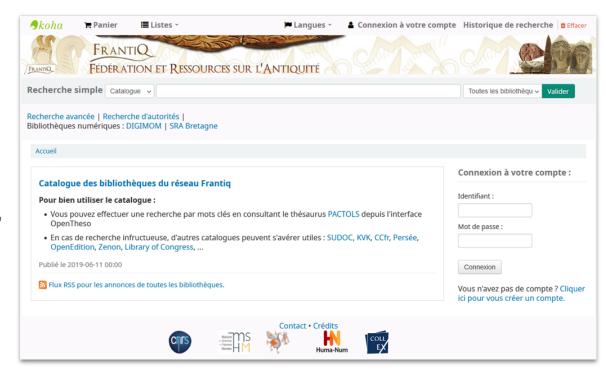


1er exemple avec une base documentaire Koha

Catalogue Collectif Indexé (CCI) de Frantiq

- 650000 notices de documents sur l'archéologie et les sciences de l'Antiquité
- des liens vers les référentiels auteurs (IdRef), revues (Mir@bel) et PACTOLS
- lien dynamique d'ajout de mots-clés PACTOLS

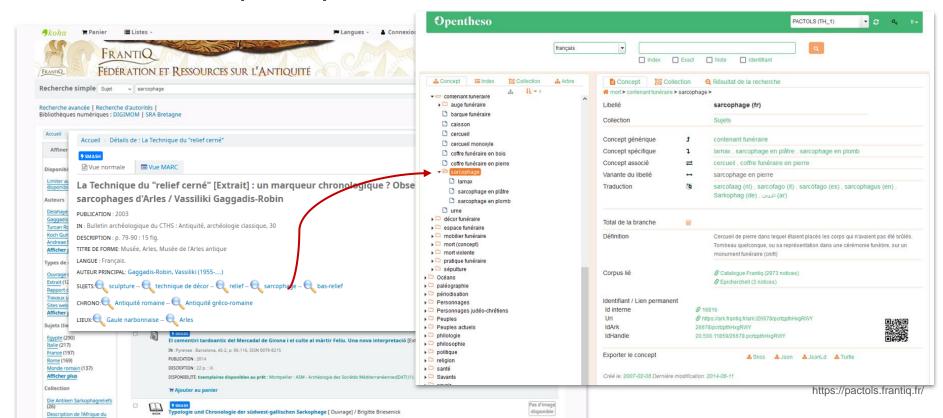
https://catalogue.frantiq.fr/



CCI, interface publique - lien notice/thésaurus

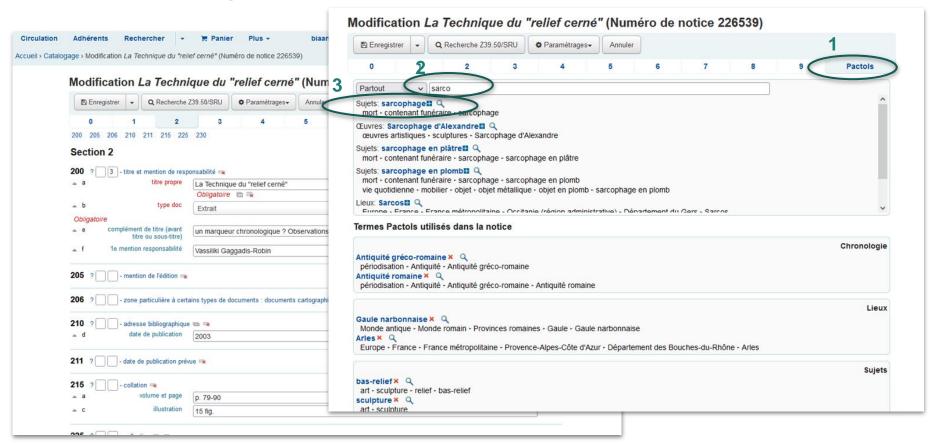
IN: Jahrbuch des römisch-germanischen Zentralmuseums. Mainz

Sarkophag-Studien (11)



https://catalogue.frantiq.fr/

CCI en mode pro



2e exemple avec une base MySQL et un site Wordpress

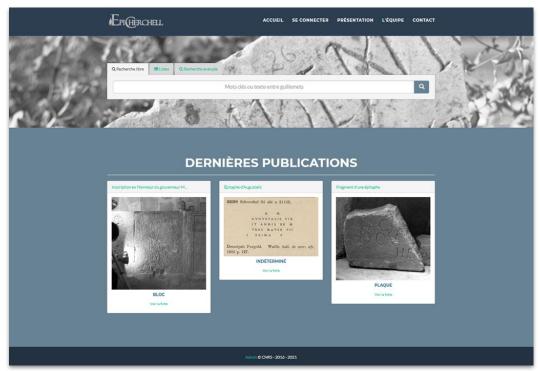
EpiCherchell

Projet d'édition des Inscriptions antiques, grecques et latines, de Césarée de Maurétanie (Cherchel, Algérie)

- potentiel de 1300 textes
- des catégories d'inscriptions très diverses : dédicaces, hommages, épitaphes, etc. gravées sur des supports variés et parfois accompagnées de reliefs.

1 fiche = étude et transcription de l'inscription + images

1 base de données MySQL interfacée à un site WordPress



http://ccj-epicherchel.huma-num.fr

http://ccj-epicherchel.huma-num.fr/interface/index.php

Les mots-clés sont liés à l'identifiant ark du concept équivalent PACTOLS

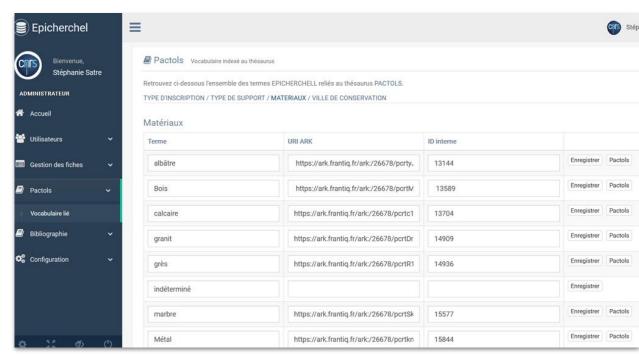
4 thématiques :

- type d'inscription,
- type de support,
- matériaux
- ville de découverte,

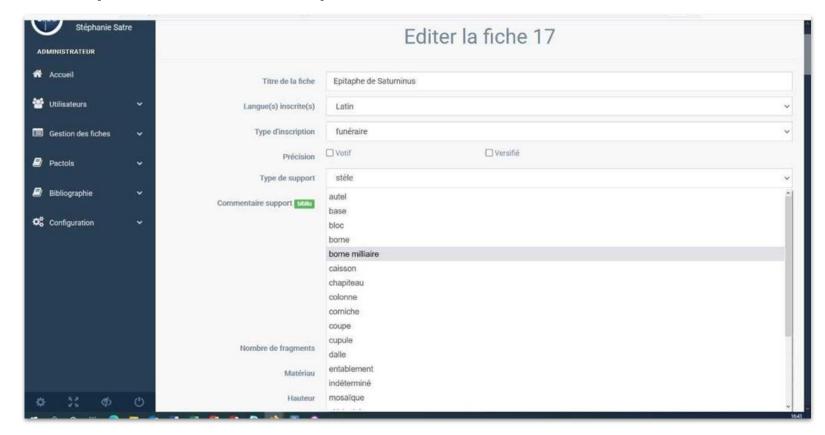
4 listes de mots-clés

1 mot-clé/champ

Ex. pour les Matériaux

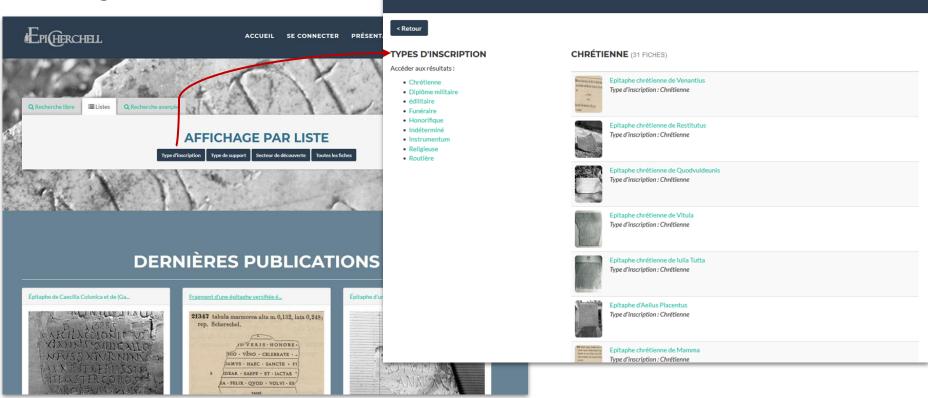


Les champs s'alimentent par sélection dans la liste

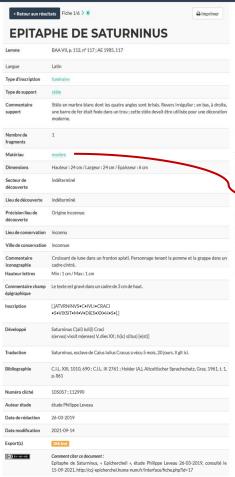


Navigation interne







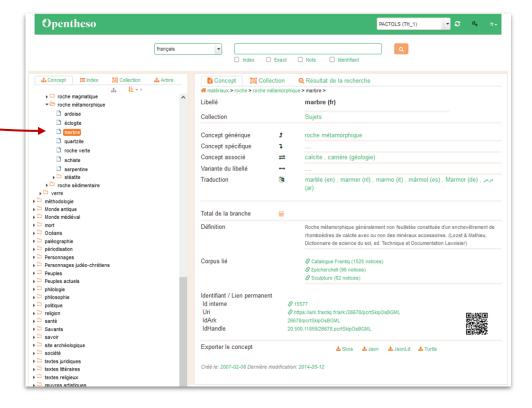




Q Texte inscription + photo



Vers le thésaurus



D'autres bases ou systèmes questionnant PACTOLS via l'API de Opentheso

EDArc de l'Inrap = données de terrain

OpenArchaeo = système d'exposition de données archéologiques structurées en CIDOC-CRM

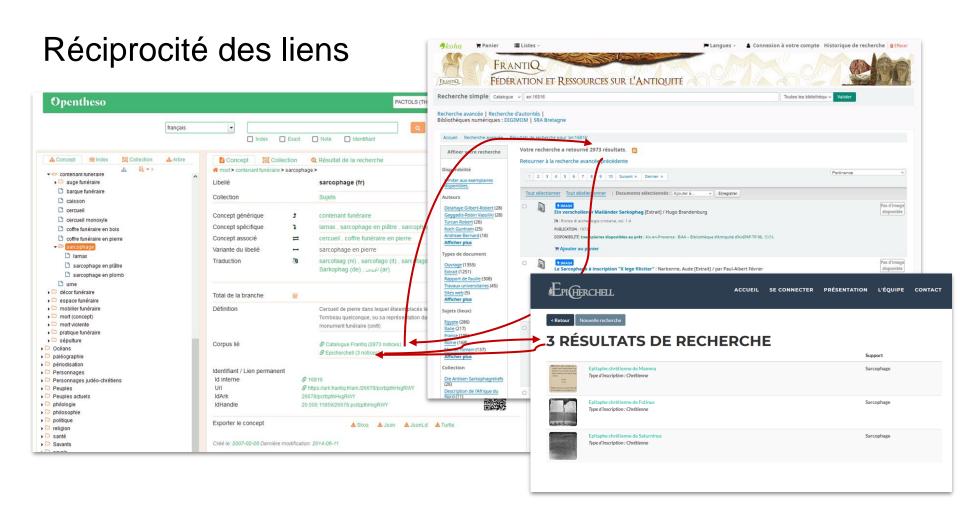
Métopes = production de fichiers en XML-TEI pour l'édition

. . .

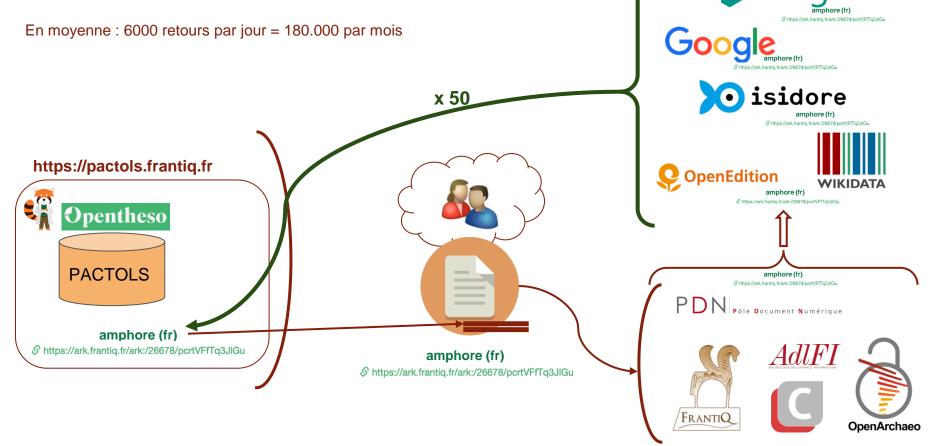
4. API et requête de retour

- L'interopérabilité est complète entre les réservoirs, base de données et thésaurus
- Les réservoirs de données hétérogènes sont liés grâce aux identifiants pérennes

Optimise la visibilité du thésaurus et des réservoirs liés.



Requêtes de retour



5. Participer à l'évolution de PACTOLS

Se tenir informé, échanger

La liste de diffusion :

pactols@services.cnrs.fr

34 abonnés Frantiq (GT Pactols, FLEN) et hors réseau



https://oceanindien.france-assos-sante.org

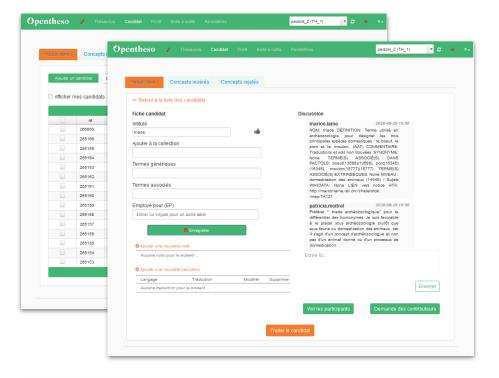
Enrichir les contenus

Suggérer des candidats

- via le module candidats :
 à nouveau accessible avec la v2,
 sur authentification
- via mail: <u>pactols@services.cnrs.fr</u> avec justification, traduction(s), positionnement

Proposer des corrections depuis la notice concept.

A venir avec la v2!





Les bonnes adresses

Consulter PACTOLS:

- version publique, opérationnelle : https://pactols.frantiq.fr
- version en développement pour consultation uniquement, url provisoire : https://pactols2.frantiq.fr

Télécharger PACTOLS : https://github.com/frantiq/PACTOLS

Télécharger Opentheso:

- logiciel : https://github.com/miledrousset/Opentheso2/tags
- Widget: https://github.com/miledrousset/Opentheso2/tree/master/src/main/webapp/scriptAutoComplete

Contacts: <u>blandine.nouvel@frantiq.fr</u>

miled.rousset@mom.fr

contact@frantiq.fr

Quelques ressources

Pour une politique publique de la donnée, Rapport demandé par le Premier ministre au député Eric Bothorel, déc. 2020 et "Les données de la recherche, un enjeux stratégique". Billet Ouvrir la science, 16 juil. 2021.

<u>Guide pratique sur l'harmonisation internationale de la gestion des données</u> de Science Europe, 2e édition, janv. 2021.

Traduction et science ouverte, Rapport du CoSo sur le multilinguisme, nov. 2020.

Europeana Archaeology, <u>projet CARARE</u>: #ConnectingArchaeology webinar, 30 avr. 2021: <u>Using Vocabularies and Linked data</u> enregistrement de la séance sur https://vimeo.com/547534057