Référentiels et vocabulaires

Pistes méthodologiques pour la construction d'un vocabulaire contrôlé

Blandine NOUVEL, Centre Camille Jullian (Aix-Marseille Univ, CNRS) & Frantiq

7 novembre 2023

Méthode

Concepts & termes

Relations sémantiques

Outils

Normalisation & FAIR

Rappel sur ce qu’est un référentiel

- Types de Référentiels et vocabulaires contrôlés
  - Du dictionnaire au **thésaurus**
  - Organisation +/- structurée
- Objectifs
  - Normaliser une terminologie pour éviter les ambigüités du langage naturel (synonymie, polysémie)
  - Partager un langage validé par une communauté
- Usages
  - Enrichir les métadonnées d’une ressource (indexation) en qualifiant son contenu (avec des mots-clés)
  - Faciliter la recherche documentaire en évitant le bruit et le silence (index et filtres)

Une affaire de longue haleine

PACTOLS 1987-2019
59727 concepts

- Anthroponymes
- Chronologie
- Lieux
- Oeuvres
- Peuples
- Sujets

Pactols 2022

Sujets
12037 concepts

- activités
- entités immatérielles
- entités matérielles
- entités nommées
- entités sociales collectives
- entités spatiales
- entités temporelles
- matériaux
- organismes vivants
- processus naturels
- rôles
- unités géopolitiques
- ~[termes dépréciés]

Lieux
50305 concepts

- entités actuelles
- entités du passé
- espaces aquatiques
- espaces terrestres
- sites archéologiques
- ~[termes dépréciés]

Par quel bout commencer ?

- Etat des lieux : qu'existe-il comme outil sur mon sujet ?
  - Les dictionnaires et lexiques spécialisés…
  - Les sites de référencement

    FAIRsharing.org standards, databases, policies

    BARTOC.org Basic Register of Thesauri, Ontologies & Classifications
  - Les vocabulaires numériques disponibles sont-ils normalisés, évolutifs, ouverts, réutilisables (FAIR) ?
  - Les adopter, y participer ou les utiliser comme source d'inspiration ?

- Le cas Pactols :
  - Recréer un nouveau à partir d'un lexique (mettre à jour et réorganiser)
  - Conserver le lien avec le Catalogue Collectif Indexé
  - Exploiter les fonctions de Opentheso

Définir le·s public·s, le·s usage·s

La transformation de Pactols d'un thésaurus documentaire à un référentiel du web pour l'archéologie a eu des incidences sur son contenu et son organisation

Évolution des ressources à indexer

De nouveaux utilisateurs

Constituer une équipe

- Des compétences
  - Compétences sur le web sémantique et la structuration de terminologies
  - Compétences techniques et scientifiques sur le domaine
  - Utilisateurs finaux (métiers IST et Données, archéologues)

- Un calendrier
  - Objectifs tenables
  - Séances de travail régulières
  - Processus itératif de discussion/validation
    - Lexique (ateliers de spécialistes)
    - Structuration (formalisation, définition, relations sémantiques)

Constituer son réservoir de mots-clés

- Extraction (automatique) de mots-clés dans un corpus (publications scientifiques) ou un ou des vocabulaires existants
- Viser l'exhaustivité terminologique du domaine?
  - Lien avec un fonds documentaire = pas de concept s'il n'y a pas au moins une ressource correspondante dans le catalogue
- Combien de termes ? de 100 à +10 000 aines

« La collecte des termes ne peut être efficace que si elle prend aussi en compte le vocabulaire des utilisateurs. » (M. Hudon, 2013)

Ré-organiser un vocabulaire

- Top-down:
  - Définir les branches principales, puis les sous-branches
  - Y positionner les concepts

- Down-top :
  - Rassembler les concepts par « famille »
  - Identifier des thématiques, de la plus précise à la plus large

- Un mélange des 2…

Les concepts & les termes

- Un terme représente un concept
    - réception de l'Antiquité ; vente d'esclave ; nucléus
- Se méfier des termes apparemment sans difficultés
    - Préférer des concepts plébiscités par les spécialistes
        - crémation plutôt qu'incinération
        - inhumation ≠ sépulture à inhumation
    - Mais quand la terminologie n'est pas [encore] stabilisée ou que le terme a des acceptions différentes selon les époques…
        - sarcophage *vs* cercueil *vs* coffre funéraire
        - charnier médiéval ≠ charnier moderne
- Exploiter les variantes et les notes d'application

L’écriture du concept

- Un substantif
  - simple ou composé
  - au singulier, sauf cas particulier (catacombes)
  - ni adjectif : pas gravé, mais galet gravé, plaquette gravée, signe gravé
  - ni verbe : pas fouiller, mais fouille, fouille subaquatique, fouille programmée

- Justesse des libellés et de leurs variantes
  - skos:prefLabel = la forme choisie/préférée
  - skos:altLabel = formes rejetées (synonymes, équivalent (ou antonymes) dans le thésaurus, formes obsolètes)
  - skos:hiddenAltLabel = formes erronées

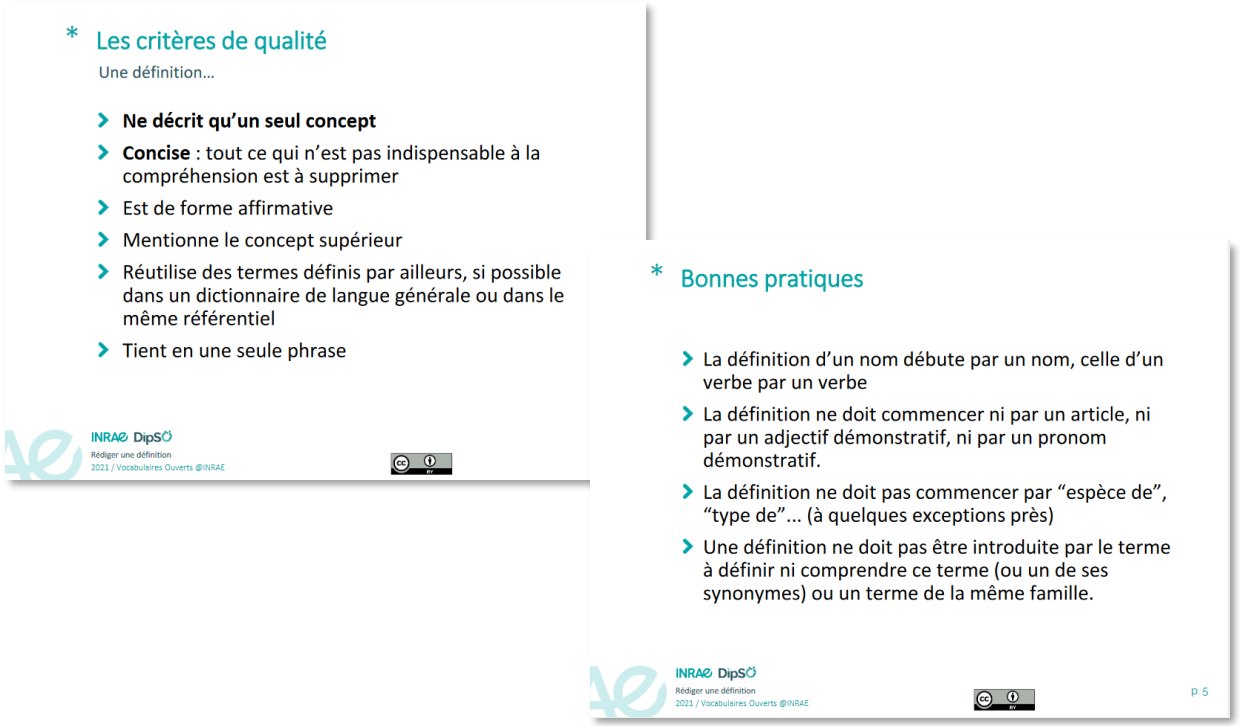
  bouton de préhension // bouton (récipient) *vs* bouton de costume // bouton (vêtement)
  instrument de musique // musique (instrument)
  oppidum plutôt qu’habitat fortifié de hauteur protohistorique
  archéologie préventive // archéologie de sauvetage

Définition et notes d'application

- Norme ISO 1087: 2019

Terminology work and terminology science — Vocabulary

Travaux terminologiques et science de la terminologie — Vocabulaire



Péloponèse
dans Pactols

Péloponèse (périphérie)

- **Définition :** Région administrative de Grèce ou périphérie comprenant cinq des sept districts régionaux de la péninsule du Péloponnèse.
- **Note d'application** : À différencier du concept "péninsule du Péloponnèse" qui est un espace géographique et qui ne couvre pas la même étendue que la périphérie du Péloponnèse.

Péninsule du Péloponèse

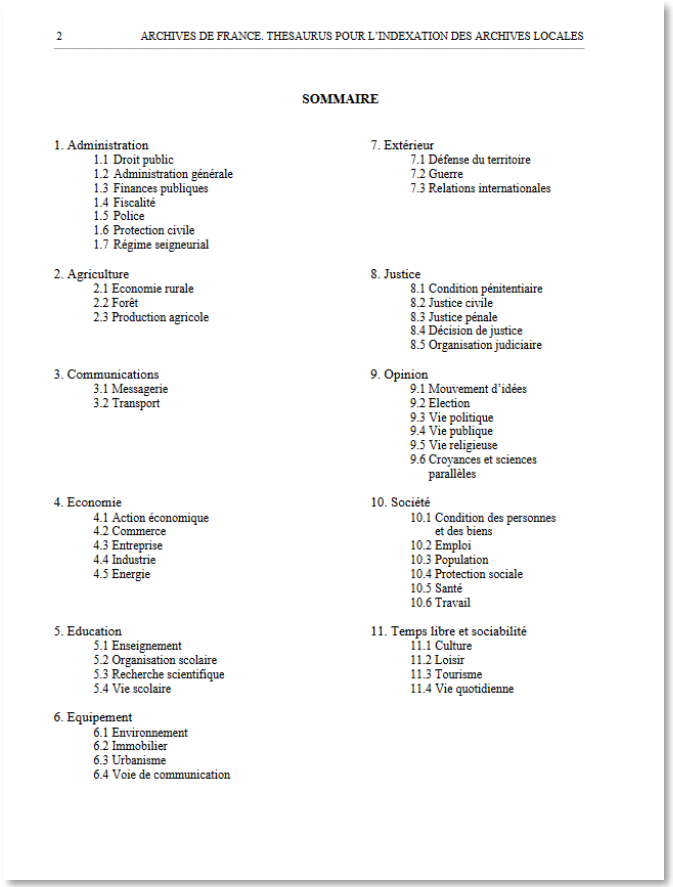
- **Définition** : Espace géographique désignant la péninsule du Péloponnèse, bordée par la mer Ionienne et la mer Égée et reliée à la Grèce continentale par l'isthme de Corinthe.
- **Note d'application** : À différencier du concept "Péloponnèse (périphérie)" qui est une région administrative de Grèce et qui ne couvre pas la même étendue géographique que la péninsule du Péloponnèse.

Rédiger une définition, quelques clés / S. Aubin, https://hal.inrae.fr/hal-03153824

Structurer le thésaurus

- Norme ISO25964 -1 et 2, *Information and documentation – Thesauri and interoperability with other vocabularies*
- Langage SKOS, *Simple Knowledge Organisation System (*Modèle FAIR développé par le W3C) Version FR
  - des objets : thésaurus, concepts, groupes
  - des propriétés : les relations sémantiques
- Bonnes pratiques pour structurer un thésaurus (https://opentheso.hypotheses.org/67)

Branches thématiques *vs* organisation ontologique



2 ARCHIVES DE FRANCE. THESAURUS POUR L'INDEXATION DES ARCHIVES LOCALES

SOMMAIRE

https://francearchives.gouv.fr/file/efb9242a3d6d4e5557940d65a95d4f830dc1a30b/static_5372.pdf

https://vocabs.dariah.eu/bbt/en/?clang=fr

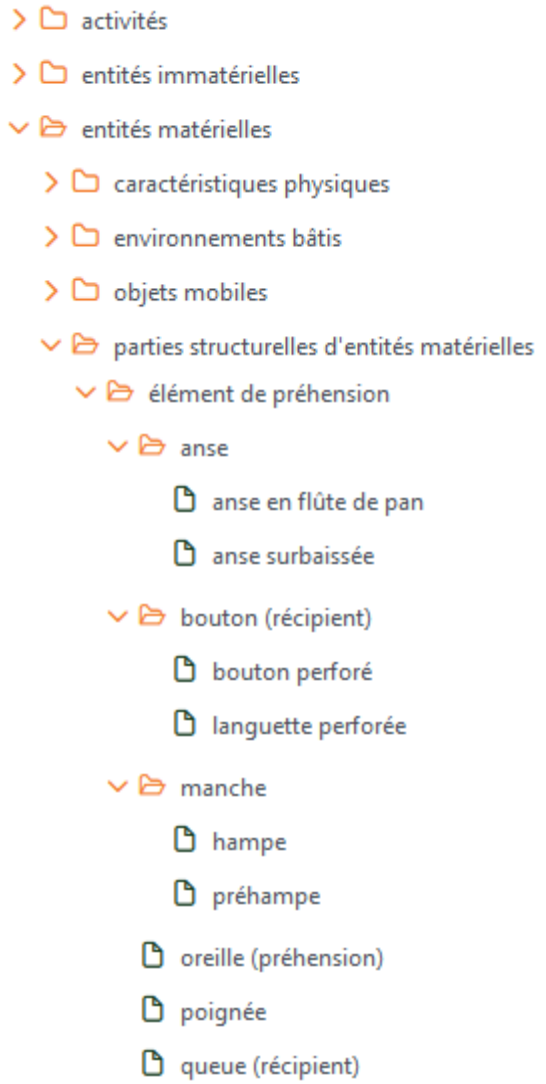
La subordination

- Caractéristique du thésaurus, représenté en arbre hiérarchique
  terme générique (TG) / terme spécifique (TS & TSI)
- Subordination en ISO :
  - Générique = IS A
    plan d'eau *est un* lac / lac *est un* plan d'eau
  - Instantielle = IS A (déclinaison)
    lac Érié *est une instance de* (iso-thes:narrowerInstantial) lac
  - Partitive = IS PART OF
    lac Érié *est une partie de* (iso-thes:narrowerPartitive) Grands lacs
- SKOS définit la direction de la subordination:
  - skos:broader / skos:narrower
    plan d'eau <skos:narrower = lac > / lac <skos:broader = plan d'eau >

https://upload.wikimedia.org/wikipedia/commons/c/c3/Canada_relief_map_2.svg

Les niveaux de profondeur

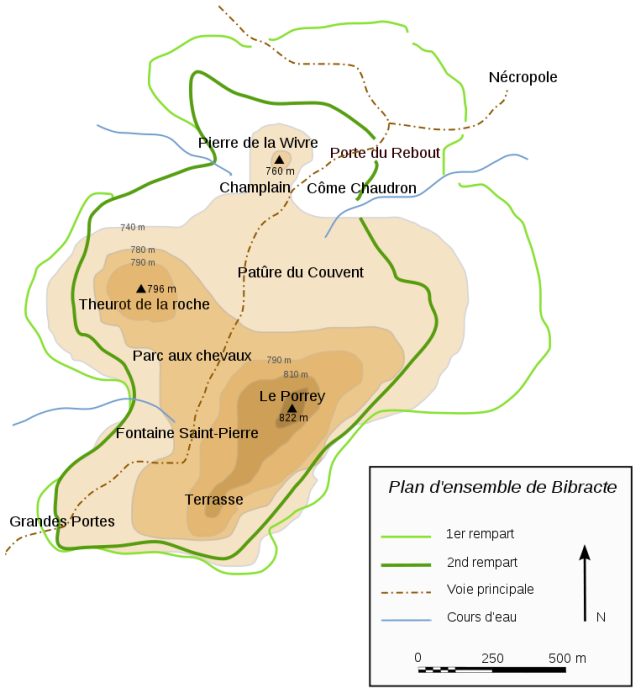
- La norme :
  - 4 niveaux ou plus
- L’usage et la visibilité,
  - Un compromis entre le nombre de niveaux et le nombre de branches
- La nécessité :
  - Respecter les besoins du public (spécialiste)



La polyhiérarchie

- Un concept a plusieurs génériques

France
Bourgogne-Franche-Comté
Département de la Nièvre
Glux-en-Glenne
Bibracte
Bibracte EPCC
Larochemillay
Bibracte
Saint-Léger-sous-Beuvray
Bibracte



La grande *domus* de Bibracte
Photo et plan du site https://www.bibracte.fr

Les relations d’association

Équivalent à un « Voir aussi »

- Lier des termes de niveau équivalent

entités matérielles > objets mobiles > **matériel de broyage** > meule > meule va-et-vient
TA
entités immatérielles > méthodes > technique de fabrication > **broyage (technique)**

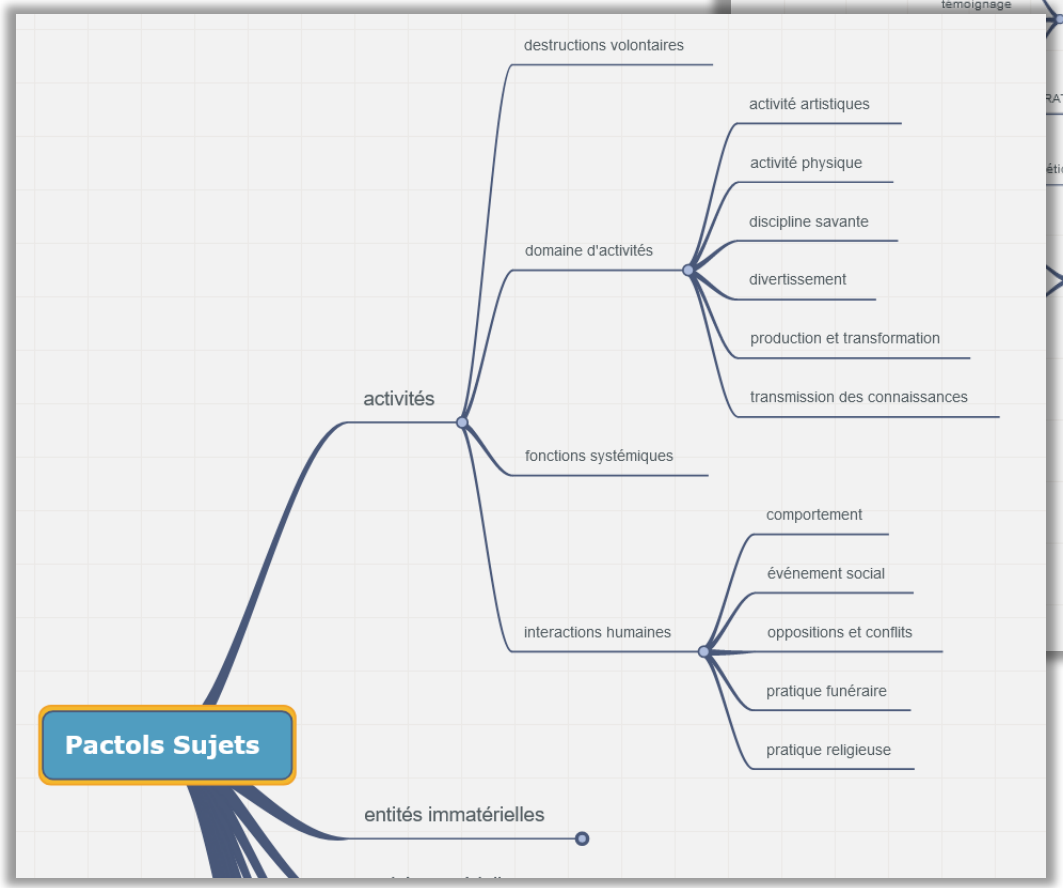
- Une solution alternative à la polyhiérarchie ou à des relations génériques transitives fautives

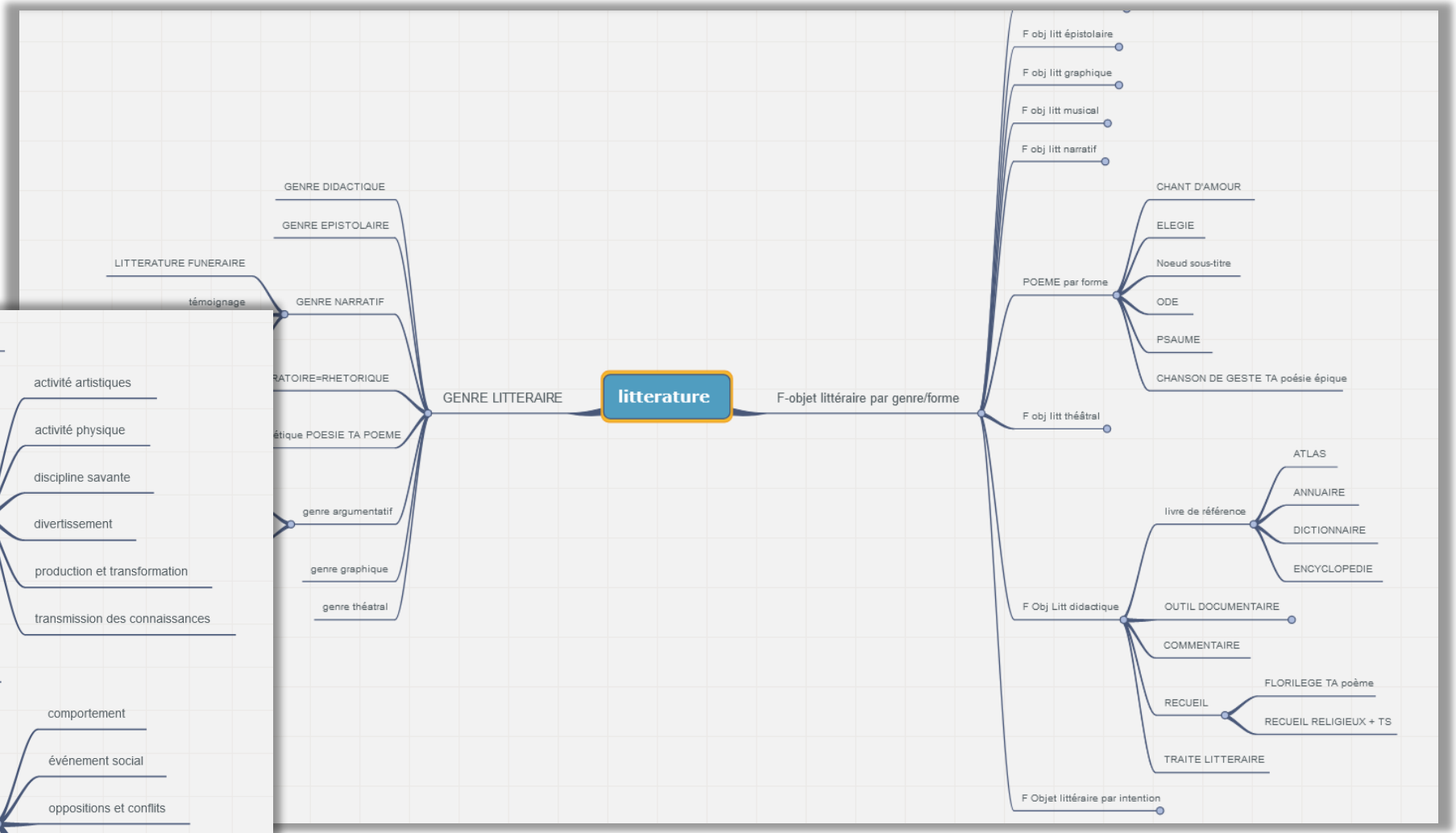
entités actuelles > … > France > … > Département de la Nièvre > Glux-en-Glenne > **Bibracte**
entités actuelles > … > France > … > Département de la Nièvre > Larochemillay > **Bibracte**
entités actuelles > … > France > … > Département de la Nièvre > Saint-Léger-sous-Beuvray > **Bibracte**
TA
espaces terrestres > … > Massif central > Morvan > **mont Beuvray**

ET PAS (dans l’organisation Pactols)

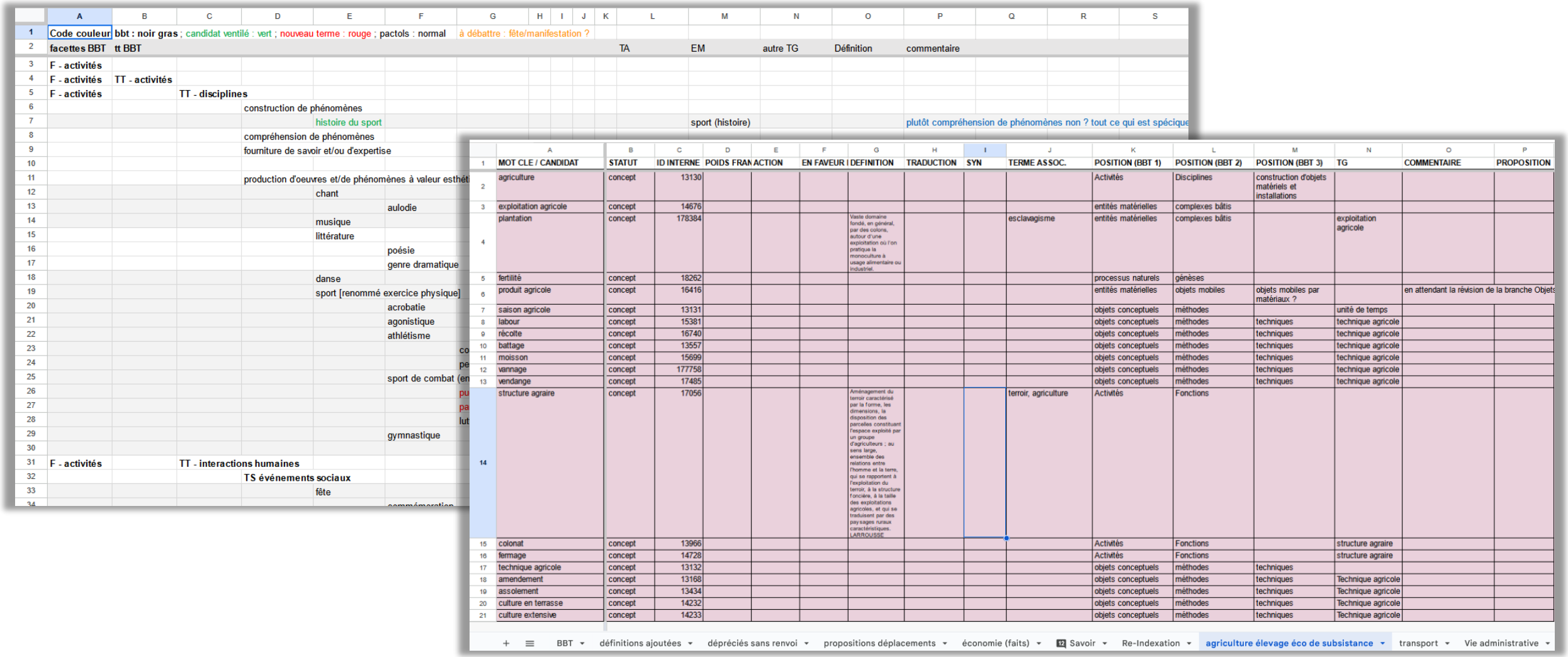
entités actuelles > … > France > … > Département de la Nièvre > Glux-en-Glenne > **Bibracte**
entités actuelles > … > France > … > Département de la Nièvre > Larochemillay > **Bibracte**
entités actuelles > … > France > … > Département de la Nièvre > Saint-Léger-sous-Beuvray > **Bibracte**
espaces terrestres > … > Massif central > Morvan > mont Beuvray > **Bibracte**

Préparer-1


Pactols Sujets
activités
destructions volontaires
domaine d'activités
activité artistiques
activité physique
discipline savante
divertissement
production et transformation
transmission des connaissances
fonctions systémiques
interactions humaines
comportement
événement social
oppositions et conflits
pratique funéraire
pratique religieuse
entités immatérielles


litterature
GENRE LITTERAIRE
GENRE DIDACTIQUE
GENRE EPISTOLAIRE
LITTERATURE FUNERAIRE
témoignage
GENRE NARRATIF
RATOIRE=RHETORIQUE
étique POESIE TA POEME
genre argumentatif
genre graphique
genre théatral
F-objet littéraire par genre/forme
F obj litt épistolaire
F obj litt graphique
F obj litt musical
F obj litt narratif
POEME par forme
CHANT D'AMOUR
ELEGIE
Noeud sous-titre
ODE
PSAUME
CHANSON DE GESTE TA poésie épique
F obj litt théâtral
F Obj Litt didactique
livre de référence
ATLAS
ANNUAIRE
DICTIONNAIRE
ENCYCLOPEDIE
OUTIL DOCUMENTAIRE
COMMENTAIRE
RECUEIL
FLORILEGE TA poème
RECUEIL RELIGIEUX + TS
TRAITE LITTERAIRE
F Objet littéraire par intention

Préparer-2



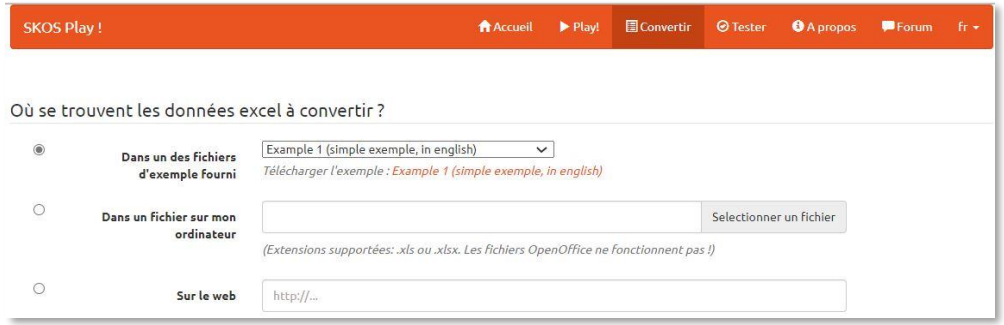
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Code couleur	bbt : noir gras ; candidat ventilé : vert ; nouveau terme : rouge ; pactols : normal					à débattre : fête/manifestation ?												
2	facettes BBT	tt BBT										TA	EM	autre TG	Définition	commentaire			
3	F - activités																		
4	F - activités	TT - activités																	
5	F - activités		TT - disciplines																
6				construction de phénomènes															
7					histoire du sport								sport (histoire)			plutôt compréhension de phénomènes non ? tout ce qui est spécique			
8				compréhension de phénomènes															
9				fourniture de savoir et/ou d'expertise															
10																			
11				production d'oeuvres et/de phénomènes à valeur esthéti															
12					chant														
13						aulodie													
14					musique														
15					littérature														
16						poésie													
17						genre dramatique													
18					danse														
19					sport [renommé exercice physique]														
20						acrobatie													
21						agonistique													
22						athlétisme													
23							co												
24							pe												
25						sport de combat (en													
26							pu												
27							pa												
28							lut												
29						gymnastique													
30																			
31	F - activités		TT - interactions humaines																
32				TS événements sociaux															
33					fête														
34						commémoration													

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	MOT CLE / CANDIDAT	STATUT	ID INTERNE	POIDS FRAN	ACTION	EN FAVEUR I	DEFINITION	TRADUCTION	SYN	TERME ASSOC.	POSITION (BBT 1)	POSITION (BBT 2)	POSITION (BBT 3)	TG	COMMENTAIRE	PROPOSITION
2	agriculture	concept	13130								Activités	Disciplines	construction d'objets matériels et installations			
3	exploitation agricole	concept	14676								entités matérielles	complexes bâtis				
4	plantation	concept	178384				Vaste domaine fondé, en général, par des colons, autour d'une exploitation où l'on pratique la monoculture à usage alimentaire ou industriel.			esclavagisme	entités matérielles	complexes bâtis		exploitation agricole		
5	fertilité	concept	18262								processus naturels	génèses				
6	produit agricole	concept	16416								entités matérielles	objets mobiles	objets mobiles par matériaux ?		en attendant la révision de la branche Objets	
7	saison agricole	concept	13131								objets conceptuels	méthodes		unité de temps		
8	labour	concept	15381								objets conceptuels	méthodes	techniques	technique agricole		
9	récolte	concept	16740								objets conceptuels	méthodes	techniques	technique agricole		
10	battage	concept	13557								objets conceptuels	méthodes	techniques	technique agricole		
11	moisson	concept	15699								objets conceptuels	méthodes	techniques	technique agricole		
12	vannage	concept	177758								objets conceptuels	méthodes	techniques	technique agricole		
13	vendange	concept	17485								objets conceptuels	méthodes	techniques	technique agricole		
14	structure agraire	concept	17056				Aménagement du terroir caractérisé par la forme, les dimensions, la disposition des parcelles constituant l'espace exploité par un groupe d'agriculteurs ; au sens large, ensemble des relations entre l'homme et la terre, qui se rapportent à l'exploitation du terroir, à la structure foncière, à la taille des exploitations agricoles, et qui se traduisent par des paysages ruraux caractéristiques. LARROUSSE			terroir, agriculture	Activités	Fonctions				
15	colonat	concept	13966								Activités	Fonctions		structure agraire		
16	fermage	concept	14728								Activités	Fonctions		structure agraire		
17	technique agricole	concept	13132								objets conceptuels	méthodes	techniques			
18	amendement	concept	13168								objets conceptuels	méthodes	techniques	Technique agricole		
19	assolement	concept	13434								objets conceptuels	méthodes	techniques	Technique agricole		
20	culture en terrasse	concept	14232								objets conceptuels	méthodes	techniques	Technique agricole		
21	culture extensive	concept	14233								objets conceptuels	méthodes	techniques	Technique agricole		

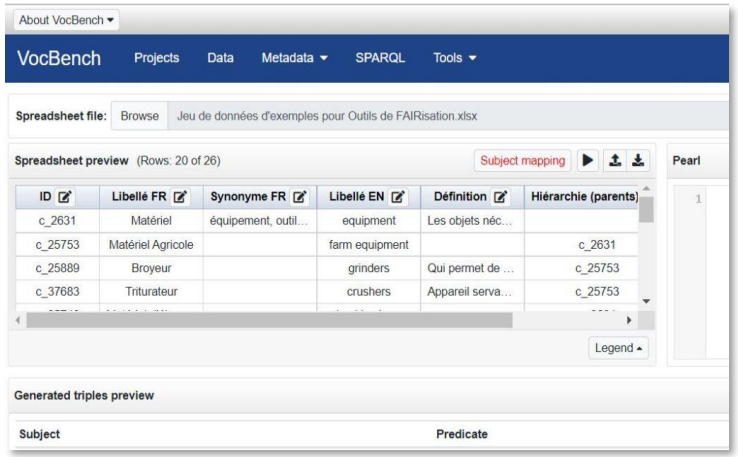
BBT | définitions ajoutées | dépréciés sans renvoi | propositions déplacements | économie (faits) | Savoir | Re-Indexation | agriculture élevage éco de subsistance | transport | Vie administrative

De CSV à SKOS

Skos Play! > « convertir une feuille de calcul Excel en fichier SKOS »



VocBench + Sheet2RDF



OpenRefine + extension RDF-openRefine

Opentheso

https://vocabulaires-ouverts.inrae.fr/2022/10/13/3-outils-transformer-tabule-skos/

FAIRiser un thésaurus

Un vocabulaire FAIR doit être

- **F**acile à trouver :
  - être enregistré (indexé, répertorié) dans un entrepôt dédié aux vocabulaires ou généraliste
  - avoir des identifiants uniques et pérennes (PID)
- **A**ccessible :
  - être accessible sur le web (https)
  - téléchargeable dans des formats standardisés (csv, rdf, Json)
  - requêtable via des APIs (open API)
- **I**nteropérable :
  - encodé dans une représentation standard, telle que le langage d'ontologie Web (OWL) ou SKOS
  - et les extensions spécifiques aux domaines
- **R**éutilisable :
  - avec une licence explicite
  - une documentation
  - et une organisation pour sa maintenance

https://vocabulaires-ouverts.inrae.fr/2022/10/13/rendre-son-vocabulaire-plus-fair/

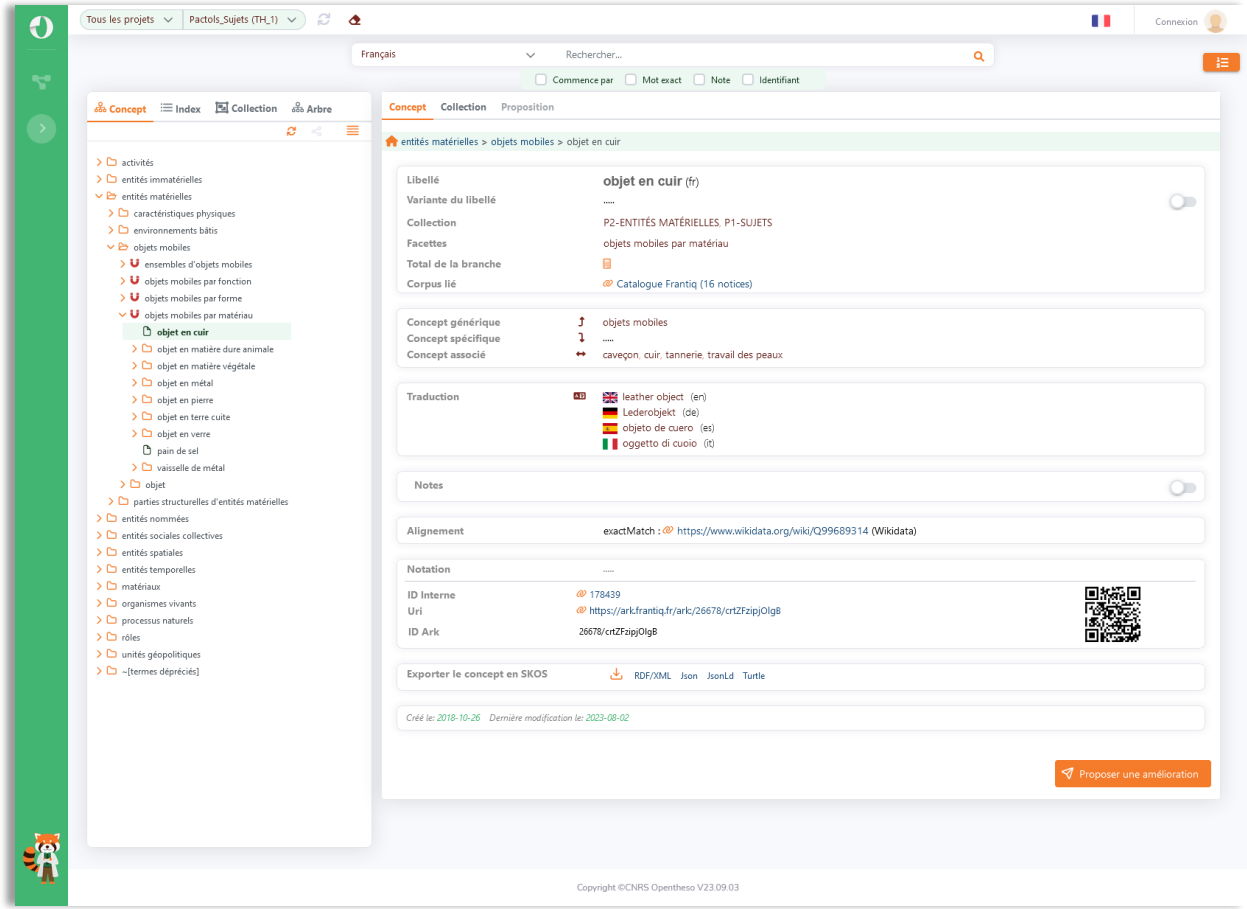
Outils spécialisés open-source

Construire et gérer « AtoZ »	• Intelligent Taxonomy Manager - https://mondeca.com/logiciels/ • TemaTres - https://sourceforge.net/projects/tematres/ • Protégé - https://protege.stanford.edu/ • Ginco, Gestion informatisée de nomenclatures collaboratives ouvertes - https://github.com/culturecommunication/ginco • Opentheso - https://github.com/miledrousset/Opentheso2
Contrôler	• SKOSPlay - https://skos-play.sparna.fr/play/
Publier et consulter	• Skosmos - https://www.skosmos.org/ • VocBench - https://bitbucket.org/art-uniroma2/vocbench3/downloads/

https://github.com/gbv/bartoc.org/wiki/Software-for-controlled-vocabularies

Opentheso & Pactols

- Depuis 2005 pour Frantiq
- Outil complet
  - Création
  - Gestion et administration
  - Collaboration (4 niveaux d'autorisation)
  - Publication
- Outil normalisé
  - ISO 25964
  - Multilingue
  - SKOS
  - API
- Diffusé
  - Github
  - IR* Huma-Num
  - Projets nationaux & internationaux
- Documenté
  https://opentheso.hypotheses.org/



Faire vivre le thésaurus

Mise à jour le lexique

- Gérer les candidats
- Administrer les propositions d'améliorations

Révision des branches

Mise à niveau des traductions

Suivi des alignements

- Wikidata
- IdRef
- Geonames, Pleiades
- PeriodO

Communication

- En interne GT Pactols
- utilisateurs-pactols@listes.services.cnrs.fr
- Web et RS

Formation

- Les catalogueurs Frantiq
- Les autres utilisateurs (MASA)

Expériences publiées

- MENG Fan, ZHOU Kaile, BU Yi, HUANG Win-Bin, ZHANG Pengyi, LONG Fei et LI Yan, « Keywords Extraction and Thesaurus Construction for Domain News », *Procedia Computer Science*, vol. 214, 2022, p. 837-844. https://www.sciencedirect.com/science/article/pii/S1877050922019597
- Building the DEFC thesaurus / Digitizing Early Farming Cultures project https://defc.acdh.oeaw.ac.at/blog/post03/
- KOMBOLO Moise, YON Jérémy, LANDRIEU François, RICHON Brigitte, AUBIN Sophie et HOCQUETTE Jean-François, « Le Thésaurus de la viande: un nouvel outil accessible à tous Une nouvelle ressource sémantique répondant aux principes de la science ouverte: le thésaurus de la viande comme outil informatique de dialogue entre les acteurs de la filière », *Viandes & produits carnés*, 4 avril 2022, p. 14. https://hal.inrae.fr/hal-03670992/document
- Thesaurus sur la diaspora canadienne sud-asiatique : https://doi.org/10.18357/kula.223

Quelques sites de référencement

- AgroPortal / Univ de Montpellier, CNRS, Inrae https://agroportal.lirmm.fr/ontologies
  - + 160 référentiels et ontologies sur l'agronomie
- BARTOC, Basic Register of Thesauri, Ontologies & Classifications https://bartoc.org/
  - Recense + 3400 vocabulaires
- FAIRsharing https://fairsharing.org/FAIRsharing
  - Plusieurs milliers de standards, terminologies, formats… tous domaines
- Glossary links / Term Coord (European Parliament, DG TRAD, Terminology coordination) https://termcoord.eu/glossarylinks/
  - Recense + 2000 glossaires sur tous les sujets traités par l'UE, évolutif
- Loterre / CNRS-Inist https://www.loterre.fr/
- TemaTres / Diego Ferreyra https://vocabularyserver.com/web/items/browse?collection=1
  - Recense les utilisateurs de l'application (690)
- Data Geo-science Vocabularies / BRGM https://data.geoscience.fr/ncl/