Le thésaurus de paléoclimatologie : création et perspectives d'utilisation

Gilles Banzet, Franck Bassinot, Henri Bretel, Valérie Daux, Ludovic Hamiaux, Christine Hatté, Catherine Kissel, Cédric Mercier, Tiphaine Penchenat, Denis-Didier Rousseau, et al.

LE THESAURUS DE PALÉOCLIMATOLOGIE : CRÉATION ET PERSPECTIVES D'UTILISATION

Banzet, Gilles [1], Bassinot, Franck [2], Bretel, Henri [3], Daux, Valérie [2], Hamiaux, Ludovic [1], Hatté, Christine [2], Kissel, Catherine [2], Mercier, Cédric [3], Penchenat, Tiphaine [2], Rousseau, Denis-Didier [4], Sempéré, Julien [3], Sepulcre, Sophie [5], Suhan, Stela [3], Vedovotto, Nathalie [1].

1. CNRS, Inist
2. Université Paris-Saclay, CNRS, CEA, UVSQ, LSCE
3. Université Paris-Saclay, DiBISO
4. Université de Montpellier, Géosciences Montpellier
5. Université Paris-Saclay, CNRS, GEOPS

Résumé

Fruit d'une collaboration entre paléoclimatologues et experts de l'information scientifique et technique, le thésaurus de paléoclimatologie, dénommé Paleosaurus, décrit et structure pour la première fois environ 2000 concepts de la discipline, en français et en anglais.

Destiné à offrir à la communauté scientifique un vocabulaire de référence, le thésaurus vise à faciliter le partage, sur le web, des données et des productions scientifiques en paléoclimatologie. Ses caractéristiques lexicales et normatives sont conformes aux standards du web sémantique. Il permet ainsi d'aligner les concepts de paléoclimatologie sur les systèmes internationaux d'organisation des connaissances.

Dans cet article, nous présentons les étapes de la création du thésaurus de paléoclimatologie, les outils logiciels employés, ainsi que ses utilisations potentielles.

Plan

Introduction : l'origine du projet et l'intérêt de créer un thésaurus de paléoclimatologie

1. Sélection des termes : méthodes et outils logiciels

1.1. Constitution d'un corpus de références bibliographiques

1.2. Extraction de termes et élaboration d'une liste de termes à examiner

1.3. Choix des termes à structurer et valider

2. Construction du thésaurus

2.1. Conversion au format SKOS, puis import dans VocBench

2.2. Premier essai de structuration

2.3. Enrichissement du thésaurus par des définitions

3. Publication et valorisation de Paleosaurus

3.1. Structuration finale et publication sur la plateforme Loterre

3.2. Le thésaurus et l'interopérabilité des données en paléoclimatologie

3.3. Utilité et valorisation de Paleosaurus auprès de la communauté scientifique

Conclusion : un travail de collaboration entre chercheurs et professionnels de l'information scientifique et technique

Références

Abréviations

Introduction : l'origine du projet et l'intérêt de créer un thésaurus de paléoclimatologie

L'Institut de l'information scientifique et technique (Inist) ainsi que l'Université Paris-Saclay, à travers sa Direction des Bibliothèques, de l'Information et de la Science Ouverte (DiBISO) et le Laboratoire des Sciences du Climat et de l'Environnement (LSCE) se réunissent au printemps 2019 pour répondre à l'appel à projet ANR - Flash « Science Ouverte ».

Le projet, dénommé « Paleosaurus », par contraction des termes paléoclimatologie et thésaurus, prévoyait de créer un vocabulaire commun et normé en paléoclimatologie afin de mieux lier et rapprocher les corpus du domaine. Il s'est proposé également d'élaborer des normes méthodologiques pour la rédaction des Plans de Gestion de Données (DMP) spécifiques à la discipline.

L'équipe de travail, composée d'experts de l'information scientifique et technique et d'experts scientifiques, s'est focalisée sur la création d'un thésaurus de paléoclimatologie[1]. Trois phases ont été identifiées : sélection de termes, construction proprement dite du thésaurus, publication et valorisation. Dans cet article nous présentons les étapes du travail effectué, les outils employés, ainsi que les usages potentiels de Paleosaurus.

Un thésaurus documentaire fait partie des référentiels utilisés en documentation, dénommés KOS (*Knowledge Organization Systems)*. C'est un vocabulaire normalisé qui comporte un ensemble structuré de termes, concepts ou classes représentant un domaine de la connaissance. Ces éléments sont organisés et reliés entre eux par des relations terminologiques (synonymie) ou sémantiques (association, hiérarchie). Le thésaurus permet ainsi l'échange et l'exploitation des

[1] Le projet n'a pas été retenu par l'Agence nationale de la recherche (ANR), mais il a cependant été maintenu sur les ressources propres des institutions engagées et s'est élargi à d'autres laboratoires d'experts : GEOPS et Géosciences Montpellier.

données par un opérateur humain ou un système d'information approprié[2]. C'est une référence pour indexer, classer et rechercher des documents et des données[3]. C'est également le résultat d'un travail soutenu et collaboratif de collecte, sélection et mise en forme de connaissances par des personnels spécialisés.

Les sciences de l'Environnement disposent de plusieurs registres de vocabulaire qui facilitent l'organisation des documents et des données et rendent les recherches plus efficaces dans ce domaine[4].

Cependant, la paléoclimatologie n'a pas encore déployé une manière normée d'indexer les contenus publiés et les données collectées et/ou produites. La discipline ne bénéficie pas d'un thésaurus ouvert et partagé disponible pour un usage scientifique et public.

La création d'un thésaurus de paléoclimatologie répond ainsi à une double demande : d'une part, renforcer le développement de la discipline et, d'autre part, standardiser l'indexation des prochaines publications et futurs dépôts de données. C'est dans une telle perspective que le développement du thésaurus a débuté au printemps 2020 avec un renouvellement des accords des parties entraînées dans le projet.

---

[2] Antoine Isaac, « Entre thésaurus et ontologies : une affaire d'interopérabilité et d'alignement » in *Documentaliste - Science de l'Information* (Dossier Web sémantique, web de données), vol. 48, n°4, 2011, pp. 48-49.

[3] Enssib, *Thésaurus*, notice créée le 16 septembre 2013 : https://www.enssib.fr/le-dictionnaire/thesaurus, consulté le 15 juillet 2022 ; Sylvie Dalbin, « Thésaurus et informatique documentaires. Partenaires de toujours ? », *Documentaliste - Sciences de l'Information*, vol.44, n°1, 2007, pp.42-55. L'auteure délimite l'activité d'*indexation* (la recherche d'un terme dans un thésaurus) de celle de *recherche d'information* (voir le tableau A pages 46-47 de l'article cité).

[4] Dominique Vachez, « Étude comparative de thésaurus en Sciences de l'Environnement. Bonnes pratiques de conception et FAIRisation de thésaurus », 2021, https://hal.science/hal-03264803, consulté le 15 février 2023 ; DoRANum, « Principaux vocabulaires contrôlés dans le domaine de l'Environnement », 2022, https://doranum.fr/environnement/principaux-vocabulaires-controles-dans-le-domaine-de-lenvironnement_10_13143_sd1c-9a43/, consulté le 15 février 2023.

Les étapes de la création de Paleosaurus suivent un processus itératif et collaboratif. A chaque phase, des activités sont déployées et des outils spécifiques sont employés, comme l'illustre la figure 1.

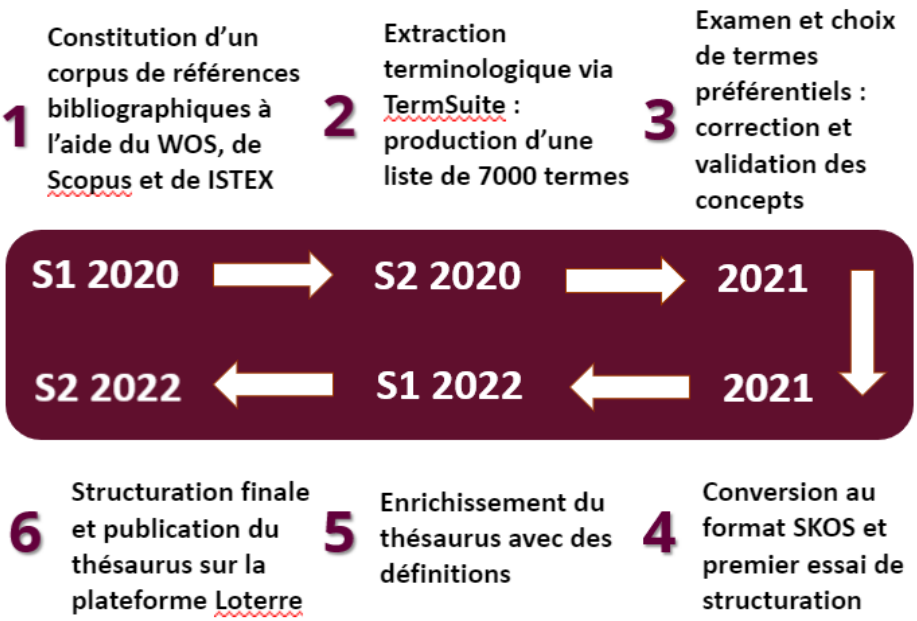


Figure 1 : Frise chronologique des étapes de création du thésaurus de paléoclimatologie, élaborée par Cédric Mercier (DiBISO)

Afin de retracer les fondations de Paleosaurus, la démarche méthodologique et le cheminement conceptuel suivi, il nous a semblé important de décrire les opérations successives, allant de la constitution du corpus initial de références bibliographiques en 2020 à la publication du thésaurus sur la plateforme *Loterre (Linked Open TERminology REsources)*[5] en 2022.

[5] https://www.loterre.fr/

1. Sélection des termes : méthodes et outils logiciels

1.1. Constitution d'un corpus de références bibliographiques

La création de Paleosaurus débute avec l'organisation d'un corpus bibliographique mis à disposition en vue de son exploitation pour une extraction terminologique.

Ce corpus a été constitué suivant deux méthodes distinctes. La première méthode, utilisant ISTEX[6], a permis de sélectionner 7 246 articles en texte intégral, couvrant la période 1890-2016. En effet, ISTEX étant une archive de documents scientifiques, les articles les plus récents n'y figurent pas systématiquement. Cette extraction a été menée par l'Inist. La deuxième, utilisant les références de la base de données *Scopus*[7] et *Web of Science*[8], a fait ressortir 30 418 références, pour la période 2000-2020, afin d'enrichir le corpus en vocabulaire récent. En revanche, l'architecture de *Scopus* et du *Web of Science* ne permettant pas d'extraire le texte intégral des publications qui y sont archivées, seuls les résumés d'articles *(abstracts)* et les mots-clés fournis par les auteurs ont été exploités dans ce cadre. Ce travail a été mené à l'Université Paris-Saclay par les experts de la DiBISO.

La recherche a porté sur des documents écrits en langue anglaise, très majoritairement utilisés dans les articles scientifiques du domaine.

---

[6] La plateforme ISTEX (Information scientifique et technique d'excellence), résultat d'un partenariat entre le CNRS, ABES, le consortium Couperin et l'Université de Lorraine, offre un accès à plus de 25 millions d'articles, en texte intégral, de toutes les disciplines scientifiques. https://www.istex.fr/

[7] *Scopus* est une base de références bibliographiques scientifiques, lancée en 2004 par l'éditeur commercial néerlandais *Elsevier*. Les utilisateurs ont accès à des milliers de titres, à des millions de profils d'auteurs et à 1,7 milliard de références citées. https://www.scopus.com/home.uri

[8] *Web of Science* est une plateforme de bases de données bibliographiques pluridisciplinaire en langue anglaise, gérée par la société *Clarivate Analytics*. La plateforme donne accès à de nombreuses références d'articles scientifiques, d'actes de conférences et de livres. Des articles en texte intégral sont également accessibles.
https://clarivate.com/webofsciencegroup/solutions/web-of-science/

La requête lancée pour la récupération de données sur ISTEX fut la suivante :

title:(palaeoclimat* OR pal?oclimat* OR "past climate"~3 OR "climate of the past"~3 OR "past climatic"~3 OR "past climatological"~3 OR "past climatology"~3 OR "past climates"~3 OR "climatic of the past"~3 OR "climatological of the past"~3 OR "climatology of the past"~3) OR abstract:(palaeoclimat* OR pal?oclimat* OR "past climate"~3 OR "climate of the past"~3 OR "past climatic"~3 OR "past climatological"~3 OR "past climatology"~3 OR "past climates"~3 OR "climatic of the past"~3 OR "climatological of the past"~3 OR "climatology of the past"~3) OR subject.value:(palaeoclimat* OR pal?oclimat* OR "past climate"~3 OR "climate of the past"~3 OR "past climatic"~3 OR "past climatological"~3 OR "past climatology"~3 OR "past climates"~3 OR "climatic of the past"~3 OR "climatological of the past"~3 OR "climatology of the past"~3)

La requête utilisée sur le WOS et Scopus a suivi la même logique :

(TITLE (palaeoclimat* OR pal?oclimat* OR "past climate" OR "climate of the past" OR "past climatic" OR "past climatological" OR "past climatology" OR "past climates" OR "climatic of the past" OR "climatological of the past" OR "climatology of the past") OR ABS (palaeoclimat* OR pal?oclimat* OR "past climate" OR "climate of the past" OR "past climatic" OR "past climatological" OR "past climatology" OR "past climates" OR "climatic of the past" OR "climatological of the past" OR "climatology of the past") OR KEY (palaeoclimat* OR pal?oclimat* OR "past climate" OR "climate of the past" OR "past climatic" OR "past climatological" OR "past climatology" OR "past climates" OR "climatic of the past" OR "climatological of the past" OR "climatology of the past")) AND PUBYEAR > 1999

1.2. Extraction terminologique et élaboration d'une liste de termes à examiner

Le corpus de références bibliographiques et d'articles a permis une extraction terminologique, lancée en juin 2020, à l'aide de l'outil *TermSuite*[9]. La liste résultante comportait 119 308 termes et 23 983 variantes, rattachées à certains termes.

Un tri initial a été réalisé pour exclure des formes inconvenantes et corriger celles mal reconnues par le logiciel *TermSuite*. D'autres considérées comme non pertinentes ont également été écartées : les adjectifs et les adverbes seuls, les formes trop génériques. Les termes rattachés aux catégories « famille/genre/espèce animale ou végétale » (noms latins ou vernaculaires) et « géographie » ont également été mis de côté pour un examen ultérieur.

Environ 45 700 formes (termes et leurs variantes) sont ainsi ressorties de ce premier tri. Parmi cette liste, seules les formes repérées avec une fréquence supérieure ou égale à 100 (seuil établi par les membres de l'équipe de travail) ont été conservées. Ce tri a permis d'aboutir à une liste d'environ 7 000 formes en anglais.

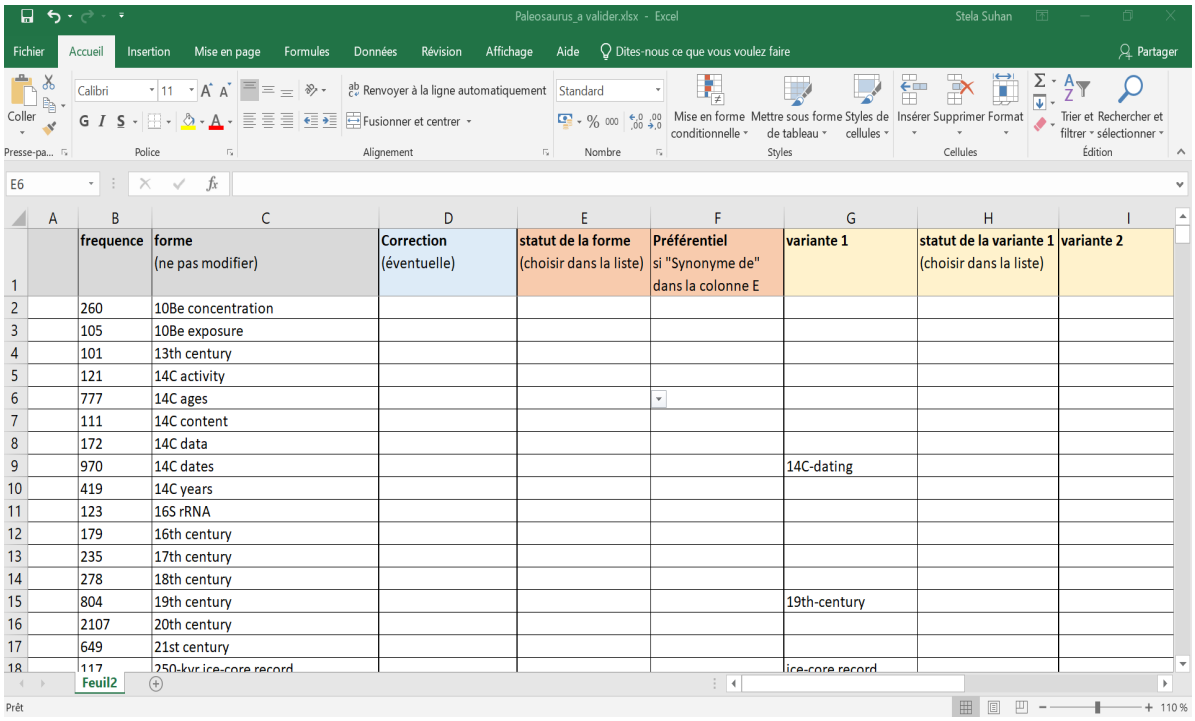
Lors de ces deux premières phases, le travail a été accompli notamment par les experts de la donnée de la DiBISO, en collaboration avec les experts de l'Inist.

[9] *TermSuite* est un outil d'extraction terminologique, conçu par le Laboratoire des Sciences du Numérique de Nantes (LNSN, CNRS-Université de Nantes). http://termsuite.github.io/

1.3. Choix des termes à structurer et valider

La liste des termes extraits a été soumise aux paléoclimatologues pour examen, sélection et validation. Ces termes, incorporés dans un fichier Excel (voir la figure 2), ont été évalués, un par un, par les scientifiques pour choisir ceux qui devaient :

- Intégrer le futur thésaurus en tant que terme préférentiel,
- Être mis en synonyme de tel ou tel terme préférentiel,
- Être rejetés.



	A	B	C	D	E	F	G	H	I
1		**frequence**	**forme** (ne pas modifier)	**Correction** (éventuelle)	**statut de la forme** (choisir dans la liste)	**Préférentiel** si "Synonyme de" dans la colonne E	**variante 1**	**statut de la variante 1** (choisir dans la liste)	**variante 2**
2		260	10Be concentration						
3		105	10Be exposure						
4		101	13th century						
5		121	14C activity						
6		777	14C ages						
7		111	14C content						
8		172	14C data						
9		970	14C dates				14C-dating		
10		419	14C years						
11		123	16S rRNA						
12		179	16th century						
13		235	17th century						
14		278	18th century						
15		804	19th century				19th-century		
16		2107	20th century						
17		649	21st century						
18		117	250-kyr ice-core record				ice-core record		

Figure 2 : capture d'écran du fichier Excel utilisé pour le choix des termes

L'expertise scientifique a été indispensable à cette étape. Les scientifiques se sont répartis entre eux les termes à structurer et à valider. Ils se sont organisés en sessions de travail individuelles ou collectives.

Au cours de cette phase, une réflexion s'est engagée sur le début de structuration du thésaurus. A titre d'exemple, « Arctic ice sheet » est-il à insérer en tant que spécifique de « ice sheet » ? Quelle est la pertinence de l'ajout en bloc de certains « groupes » de données (âges géologiques, roches, taxonomie animale/ végétale, toponymes etc.) ?

Si les termes sélectionnés étaient initialement tous en anglais, il a été convenu à cette étape (2021) de concevoir un thésaurus bilingue (anglais et français), afin qu'il réponde à la fois aux besoins des chercheurs, des enseignants ou étudiants et à ceux d'un public potentiellement plus large.

2. Construction du thésaurus

2.1. Conversion au format SKOS, puis import dans VocBench

Pour représenter et rendre les informations réellement partageables et exploitables, les experts de l'Inist ont réalisé en 2021 la conversion en SKOS de la liste des termes sélectionnés et organisés initialement dans un document Excel.

SKOS (*Simple Knowledge Organization System*)[10] est un format qui se veut simple pour faciliter l'accès et les échanges de données terminologiques dans le WEB sémantique. Il permet de représenter des concepts avec les descriptions qui leur sont attribuées : informations terminologiques, liens sémantiques entre concepts, notes d'application, définitions, exemples[11].

L'adoption d'un index standardisé et précis en SKOS permet de représenter les termes liés à leur description. Elle aide à interroger les liens et à corriger les anomalies éventuellement détectées. Elle est également en accord avec les principes FAIR (Facile à trouver, Accessible, Interopérables, Réutilisable)[12]. Le format SKOS est enfin celui qui est utilisé à terme pour exposer la ressource terminologique finalisée sur la plateforme Loterre.

---

[10] Alistair Miles, Sean Bechhofer (eds.), *SKOS Reference*. W3C Recommendation, 2009. https://www.w3.org/TR/skos-reference/, consulté le 18 juillet 2022.

[11] Antoine Isaac, « Les référentiels : typologie et interopérabilité », pp. 85-104, in Lisette Calderan, Pascale Laurent, Hélène Lowinger et Jacques Millet (coord.), *Le document numérique à l'heure du web de données*. Séminaire Inria, Carnac, 1er-5 octobre 2012, ADBS Editions, Paris, 2012, Version HAL : https://hal.inria.fr/hal-00740282v2, consulté le 18 juillet 2022. Voir également Antoine Isaac et Thierry Bouchet, « Rameau et Skos », *Arabesques*, 54 | 2009, 13-14, référence électronique : DOI : 10.35562/arabesques.2109, consulté le 18 juillet 2022.

[12] Pour mieux comprendre comment SKOS correspond aux principes FAIR, à voir la ressource : Jérémy Yon, Sophie Aubin, « SKOS : un standard pour une ressource sémantique simple et FAIR », https://vocabulaires-ouverts.inrae.fr/nos-fiches/rendre-son-vocabulaire-plus-fair/skos-un-standard-pour-une-ressource-semantique-simple-et-fair/, consulté le 22 février 2023.

Afin de poursuivre l'enrichissement et la structuration du thésaurus, ce fichier SKOS a été importé dans VocBench, outil libre multilingue d'édition et de gestion collaborative de terminologies[13]. Le travail sur VocBench a servi à enrichir les concepts et à créer une hiérarchie qui représente le format idéal pour vérifier le contenu du thésaurus et en détecter les manques éventuels.

A partir de cette étape, juin 2021, la vérification des entrées sélectionnées dans le fichier partagé a été exécutée en deux temps. Les experts de l'Inist ont tout d'abord traduit en français les termes retenus par les experts scientifiques et ont effectué le tri des entrées insérées. Par la suite, le travail commun (experts de l'Inist et scientifiques) a consisté à compléter les hiérarchisations, à supprimer les entrées non pertinentes, à combler les manques et à implémenter des mots (génériques ou déclinaisons).

2.2. Premier essai de structuration

Un premier essai de structuration a été réalisé par l'Inist en septembre 2021 afin d'insérer les concepts validés au sein d'une hiérarchie de thésaurus[14]. En parallèle, des traductions en français et certains synonymes ont été proposés, ainsi que des liens vers des concepts équivalents antérieurement publiés sur Loterre au sein de la collection « Sciences de la Terre »[15].

---

[13] *VocBench (Acquisition et Gestion des connaissances)* est conçu pour aider les institutions publiques à maintenir et à publier leurs vocabulaires contrôlés de manière ouverte et interopérable Cf. https://op.europa.eu/fr/web/eu-vocabularies/vocbench, consulté le 21 juillet 2022

[14] Nathalie Vedovotto, responsable du service ingénierie terminologique de l'Inist, « Paleosaurus : premier essai de structuration », courriel envoyé à l'équipe du travail le 27 septembre 2021, en annexant des exports en pdf (fichiers compressés), en anglais et en français, pour une visualisation du thésaurus.

[15] Cf. https://skosmos.loterre.fr/26L/fr/, consulté le 21 juillet 2022.

Le thésaurus de paléoclimatologie comptait, à cette phase, 1939 entrées, issues des formes sélectionnées par les scientifiques, en incluant les âges géologiques qui furent insérés à partir d'un fichier maintenu à l'Inist.

Ces entrées ont été organisées selon la structure classique d'un thésaurus :

- Relations hiérarchiques entre concepts : *concepts génériques* et *concepts spécifiques*[16],
- Relations d'équivalence entre termes : *synonymes* (et équivalences linguistiques),
- Relations associatives entre concepts : concepts associés.

Les relations entre concepts sont présentées selon la forme d'un thésaurus multi-hiérarchique, ce qui signifie qu'un concept peut être rattaché à plusieurs termes génériques. A l'opposé, dans un référentiel mono-hiérarchique, un concept n'est lié qu'à une seule branche.

Cette première structuration du thésaurus a mis en évidence des interrogations, des réflexions et des points à clarifier. Ces éléments concernaient en l'occurrence les concepts des catégories : toponymie, espèces animales et végétales, composés chimiques, instruments et méthodes d'analyse et termes de stratigraphie (étages géologiques et autres périodes de référence telles les glaciations). Ainsi, la sélection des termes stratigraphiques, par exemple, s'est appuyée sur la classification internationale (en dehors de certains termes spécifiques se rapportant à des événements ou périodes d'intérêt particulier). La liste des espèces animales et végétales a été retravaillée par les experts scientifiques en partant du constat que le même niveau de précision n'était pas forcément nécessaire selon les branches concernées (ex. pour les foraminifères, on utilise au quotidien le nom d'espèce plutôt que le nom de genre). Des formulations inappropriées ont par ailleurs été détectées dans les articles pour les espèces chimiques, les instruments et les méthodes

[16] Un concept spécifique répond aux propositions « est une partie de » ou « est une sorte de » par rapport à son concept générique.

d'analyse. Dans ce dernier cas, la solution adoptée a consisté à les exclure de l'affichage du thésaurus, tout en les conservant en vue des indexations sous le statut *skos:hiddenLabel* (termes cachés dans les interfaces de visualisation mais disponibles pour les traitements automatiques).

La prévisualisation structurée du thésaurus a permis de mieux détecter certaines anomalies (termes ou groupes trop spécifiques, trop génériques ou hors sujet), de repérer des lacunes et des manques (listes incomplètes de concepts relatifs à un sujet particulier).

2.3. Enrichissement du thésaurus par des définitions

L'Inist a testé deux méthodes pour extraire automatiquement des champs lexicaux au sein d'articles et autres ressources du web pouvant correspondre à des définitions[17].

La première méthode est fondée sur l'utilisation de l'outil OpenRefine[18] et consiste à « réconcilier » (aligner) les entrées du thésaurus avec wikipédia ou DBpédia, puis à collecter tout ou partie des informations disponibles dans le champ « description » de ces réservoirs. Une liste de propositions d'esquisses de définitions (en français et/ou anglais) et leurs liens respectifs (URL) concernant de l'ordre de 600 concepts (sur les 1921 retenus initialement) a ainsi été fournie en avril 2022[19].

La seconde méthode consiste à utiliser l'outil Unitex/Gramlab[20], une suite logicielle libre fondée sur la constitution de graphes permettant de rechercher des

[17] Patricia Fener, Claude Dahdouh, « Repérage automatique d'énoncés définitoires avec Unitex pour l'aide à l'enrichissement de ressources terminologiques : retour d'expérience », 2021, https://hal.archives-ouvertes.fr/hal-03390661, consulté le 26 octobre 2022.

[18] Logiciel libre de nettoyage et mise en forme de données. https://openrefine.org/

[19] Gilles Banzet, Ludovic Hamiaux, Nathalie Vedovotto, « Liste de propositions de définitions », message envoyé par Inist le 21 avril 2022.

[20] Suite logicielle libre, multiplateforme, multilingue, fondée sur des dictionnaires et des grammaires pour l'analyse de corpus. https://unitexgramlab.org/fr

contextes textuels dans des documents. Un graphe a ainsi été élaboré à l'Inist, s'appuyant sur la phraséologie des définitions en langue anglaise et sur un dictionnaire constitué à partir des concepts du thésaurus, afin de détecter des parties de phrases susceptibles de contenir des énoncés définitoires. Un fichier d'une cinquantaine de propositions de définitions (en anglais) ainsi que leurs sources a été proposé en juin 2022, en appliquant le graphe en question sur le texte intégral des 7 200 articles extraits d'ISTEX au début du projet. La figure 3 décrit schématiquement l'enrichissement du Paleosaurus par des définitions.

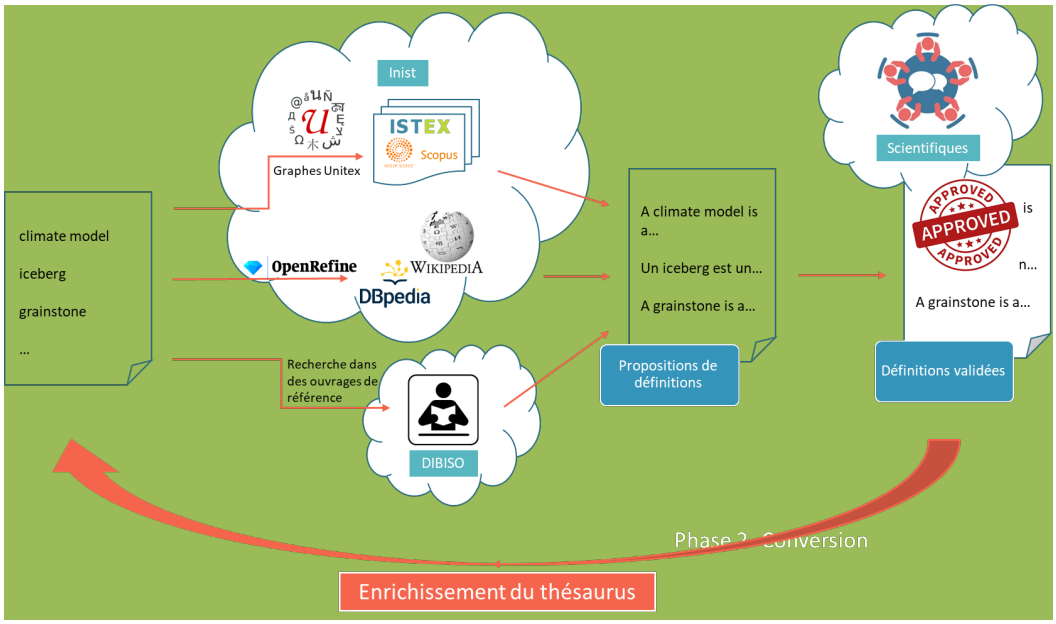


Figure 3 : Enrichissement des concepts sur l'exemple des définitions. Schéma élaboré par Ludovic Hamiaux et Gilles Banzet (Inist)

Ces propositions constituent la matière première pour l'enrichissement du thésaurus avec des définitions, qui ont ainsi été sélectionnées et vérifiées, retravaillées et reformulées. Dans le cadre d'une mission doctorale, cette tâche a été confiée à une doctorante en sciences du climat et de l'environnement qui a revu et retravaillé les définitions des 600 termes concernés.

3. Publication et valorisation du Paleosaurus

3.1. Structuration finale et publication de Paleosaurus sur la plateforme Loterre

En novembre 2022, une première version du thésaurus de paléoclimatologie a été publiée sur Loterre, plateforme multidisciplinaire d'exposition des ressources terminologiques scientifiques. Développée par l'Inist et fondée sur les technologies du web de données, la plateforme vise à faciliter les échanges et l'interopérabilité des ressources terminologiques. Avec Loterre, l'Inist contribue non seulement à publier des terminologies ouvertes mais aussi à les rendre le plus possible conformes aux principes FAIR d'accessibilité, de lisibilité, d'interopérabilité et de réutilisation[21].

Paleosaurus, publié sous licence libre CC-BY 4.0, est accessible, consultable et téléchargeable librement (en formats SKOS-RDF, csv, pdf et json-ld)[22]. Il compte environ 2000 entrées conceptuelles, hiérarchisées sous une trentaine de top-concepts (concepts de tête)[23]. Il est poly-hiérarchique, ce qui signifie qu'un concept peut être rattaché à plus d'un terme générique. Tous les concepts sont décrits par une forme préférentielle française et une anglaise, un grand nombre étant de surcroît enrichi par des synonymes, sachant qu'actuellement, de l'ordre de 600 comportent

[21] Majid Khayari, Véronique Reszetko, Dominique Vachez, Nathalie Vedovotto, Jérémy Yon, et al., *De TermSciences à Loterre : comment l'Inist-CNRS a rendu les terminologies ouvertes plus conformes aux principes FAIR*, 2021, https://hal.archives-ouvertes.fr/hal-03176063/

[22] https://skosmos.loterre.fr/QX8/fr/

[23] « Le concept de premier niveau ou top concept est le concept le plus haut placé dans la hiérarchie à laquelle il appartient. Il est représenté par un terme de tête ou *top term*. Situé à la "racine" du vocabulaire, le concept de premier niveau constitue un point d'entrée dans l'une des branches d'un thésaurus », Cf. Emmanuelle Perrin, *Construire un thésaurus*, 2020, https://opentheso.hypotheses.org/48, consulté le 8 avril 2024

des définitions (en français et/ou anglais), en plus d'éventuelles relations d'association vis-à-vis d'autres concepts.

Afin de l'illustrer, la figure 4 présente l'exemple du terme « glaciation ».

Figure 4 : capture d'écran de la présentation du concept "Glaciation" en français, https://skosmos.loterre.fr/QX8/fr/page/-5K8FD2VC-9, consulté le 10 janvier 2024

Le concept « glaciation » possède deux concepts génériques qui sont *période climatostratigraphique* et *phénomène climatique*. Il se définit comme : « (…) une phase paléoclimatique froide, en même temps qu'une période géologique de la Terre durant laquelle une partie importante des continents est englacée (…) ». Le concept a comme formes spécifiques : *anglien, dévensien, donau, elstérien, glaciation Baltique, glaciation Bull Lake, glaciation continentale et autres*. Les concepts associés à ce terme sont : *glaciaire, déglaciation, terminaison glaciaire*. « Glaciation » a comme synonymes : *âge de glace, âge glaciaire, cycle glaciaire, période glaciaire*. Le terme se traduit en anglais par : *glaciation, englaciation, glacial age, glacial cycle, glacial episode, glacial event, glacial installment, glacial period, ici age*. On peut retrouver ses équivalents exacts dans les ressources « Science de la Terre » et dans le

thésaurus de changement climatique, hébergées également sur la plateforme Loterre et dans l'encyclopédie libre Wikipédia. Via la plateforme ISTEX on peut consulter des articles qui vont usage de ce concept, 52626 articles sont recensés, toutes catégories[24].

3.2. Le thésaurus et l'interopérabilité des données en paléoclimatologie

Les questions d'interopérabilité et d'ouverture des données deviennent de plus en plus fondamentales dans le contexte de la Science Ouverte et dans l'environnement du Web sémantique[25]. L'utilisation d'un vocabulaire commun et partagé s'avère ainsi essentiel pour assurer une interconnexion entre les données, leur traitement informatique et documentaire[26]. De plus, l'emploi de termes normalisés est une condition préalable pour la description, la compréhension et la réutilisation des données de la recherche.

Dès lors, l'un des enjeux du projet a été de rendre les données produites en paléoclimatologie interopérables à l'aide de ce type de vocabulaire normalisé.

L'emploi d'un vocabulaire contrôlé est également une recommandation de l'Agence nationale de la recherche pour l'évaluation de la qualité des données dans les plans de gestion des données (ANR 2019)[27].

---

[24] Cf. https://skosmos.loterre.fr/QX8/fr/page/-5K8FD2VC-9, consulté le 10 janvier 2024

[25] Selon la définition proposée par la norme ISO 25964 « Thésaurus et interopérabilité avec d'autres vocabulaires » (2011-2013), **l'interopérabilité se définit comme** la capacité de deux ou plusieurs systèmes à échanger des informations et à utiliser les informations qui ont été échangées avec une perte minimale de contenu. Cf. Pour la première partie (2011) : https://www.iso.org/fr/standard/53657.html et pour la deuxième partie (2013) : https://www.iso.org/fr/standard/53658.html, consulté le 2 septembre 2022

[26] On se réfère ici aux métadonnées, bases de données, catalogues de bibliothèques, bibliothèques numériques.

[27] https://anr.fr/fr/actualites-de-lanr/details/news/lanr-met-en-place-un-plan-de-gestion-des-donnees-pour-les-projets-finances-des-2019/ , dernière consultation le 11 décembre 2023

Dans quelle mesure Paleosaurus répond-t-il aux principes FAIR : Facile à Trouver, Accessible, Interopérable, Réutilisable[28] ?

Paleosaurus aligne les concepts de paléoclimatologie sur les systèmes internationaux d'organisation des connaissances. Ses caractéristiques lexicales et normatives sont conformes aux standards du web sémantique. Il est facile à trouver, documenté, lisible par les machines et exploitable par le public. Chaque concept est identifié par un identifiant pérenne, basé sur la norme URI (Uniform Resource Identifier)[29] et de format ARK (Archival Resource Key)[30]. Le thésaurus dans sa globalité est également associé à un identifiant pérenne de type DOI Datacite (https://doi.org/10.13143/lotr.1369). Il se veut donc en accord avec les principes FAIR.

En même temps, le thésaurus est un outil d'aide pour une meilleure « fairisation » et réplicabilité des ensembles de données. Il fournit un lexique scientifique structuré et hiérarchisé pour optimiser la recherche, cataloguer et récupérer l'information disponible en paléoclimatologie.

L'usage d'un vocabulaire contrôlé apporte de la clarté et de la précision, et peut faire correspondre les termes de l'indexation avec ceux de la description et de l'interrogation des données. Il contribue ainsi à une uniformisation et à une meilleure compréhension des termes utilisés car il vient lever d'éventuelles ambiguïtés au niveau des concepts qui peuvent présenter des significations et interprétations différentes selon le contexte. De cette manière, le thésaurus de paléoclimatologie garantit à minima une interopérabilité sémantique[31] des ressources pour lesquelles il sera utilisé.

---

[28] Mark D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, n°3, 2016, https://doi.org/10.1038/sdata.2016.18

[29] https://www.w3.org/Addressing/URL/uri-spec.html

[30] https://www.bnf.fr/fr/lidentifiant-ark-archival-resource-key

[31] En général, on distingue quatre formes d'interopérabilité : 1. Interopérabilité système au niveau du matériel et du système d'exploitation ; 2. Interopérabilité syntaxique liée au format des données et à leur encodage ; 3. Interopérabilité structurelle fondée sur le modèle des données ; 4. Interopérabilité sémantique, au niveau terminologique, que garantir l'emploi de vocabulaire contrôlés et partagés. Cf.

3.3. Utilité et valorisation de Paleosaurus auprès de la communauté scientifique

Paleosaurus est destiné à offrir à la communauté scientifique un vocabulaire de référence, censé faciliter le partage, sur le web, des données et des productions scientifiques en paléoclimatologie.

Il offre une vue d'ensemble de la discipline et permet d'appréhender au mieux les concepts scientifiques qui la caractérisent. Il contribue de même à une homogénéisation des termes utilisés et à l'assignation de ces mêmes termes au regard de contenus similaires (du fait de la reconnaissance des différentes variantes d'un même terme). Il est évolutif, des améliorations et des ajouts pouvant se faire en continu en fonction des retours des utilisateurs.

La paléoclimatologie dispose dorénavant de son propre vocabulaire contrôlé, pour explorer et partager plus aisément les connaissances et les données produites. Un tel référentiel devrait en l'occurrence permettre de relier beaucoup plus facilement les termes aux données qui leurs sont le plus couramment attribuées.

De surcroît, il pourra simplifier les échanges entre scientifiques, l'enregistrement des données pour une réutilisation ultérieure, l'appropriation par les étudiants et le transfert des résultats vers les autres communautés.

Il sera possible de consulter Paleosaurus pour vérifier l'exactitude syntaxique et sémantique des concepts employés, pour décrire les ensembles de données à l'aide de mots-clés, notamment dans des entrepôts de données de la recherche tels

---

Marcia Lei Zeng, "Interoperability", in Birger Hjørland and Claudio Gnoli (eds.), *Encyclopedia of Knowledge Organization,* https://www.isko.org/cyclo/interoperability#app1, consulté le 2 septembre 2022

que *Recherche Data Gouv*[32], pour préparer un plan de gestion des données ou pour une recherche bibliographique efficace.

L'expérience relatée dans cet article décrit et valorise le travail et les moyens mis en œuvre dans le projet de création de Paleosaurus. La démarche méthodologique adoptée et les outils employés montrent le processus collaboratif déployé pour parvenir à l'élaboration d'un tel référentiel nécessaire dans l'activité de recherche.

L'intérêt d'un thésaurus résidant dans son utilisation, nous nous employons à faire connaître Paleosaurus auprès de la communauté des paléoclimatologues, à l'échelle internationale. Le projet a notamment été présenté sous la forme d'un poster scientifique[33] au cours de la 7ème édition du colloque « Climats et impacts 2022 »[34] ayant eu lieu en novembre 2022 à l'Université Paris Saclay. Il a fait l'objet d'une présentation-discussion « Le Paleosaurus, vocabulaire normé en paléoclimatologie », lors de la journée de lancement de l'Open Science Month 2023 à l'Université Paris Saclay[35].

[32] https://recherche.data.gouv.fr/fr

[33] Voir le poster « Un nouveau thésaurus de paléoclimatologie : création et perspectives d'utilisation » dans HAL : https://hal.science/hal-03895146v1

[34] https://premc.org/climat-impacts-2022/, dernière consultation le 10 janvier 2023

[35] https://www.universite-paris-saclay.fr/open-science-month-2023, dernière consultation 11 décembre 2023

Conclusion : un travail de collaboration entre chercheurs et professionnels de l'information scientifique et technique

La création de Paleosaurus est le résultat d'une collaboration fructueuse entre des paléoclimatologues et des experts de l'information scientifique et technique, rattachés à l'Inist ou à la DiBISO de l'Université Paris-Saclay. Les rôles de chacun, durant les deux ans et demi du déroulement du projet, ont été tout à la fois complémentaires et bien définis :

- Les équipes de la direction des bibliothèques ont joué un rôle d'interlocuteur de proximité avec les chercheurs dès le début du projet, identifiant avec les scientifiques la problématique (en particulier l'absence de vocabulaire normé dans la discipline) et les mettant en lien avec les experts nationaux du service d'ingénierie terminologique de l'Inist. La bibliothèque universitaire a ensuite assuré en partie la constitution du corpus bibliographique initial puis l'organisation régulière des réunions regroupant l'ensemble des partenaires du projet. Elle a également coordonné la réalisation d'éléments de valorisation de ce thésaurus (poster scientifique et article notamment).
- Les paléoclimatologues ont apporté l'expertise scientifique disciplinaire indispensable au projet. Le long travail de sélection et de première hiérarchisation des termes leur a notamment incombé. Le travail de production des définitions a été réalisé par Tiphaine Penchenat, alors doctorante du LSCE, accueillie à la bibliothèque universitaire d'Orsay pour cette mission.
- Les spécialistes de l'Inist ont apporté les outils indispensables à la création du thésaurus (extracteurs de termes et de définitions, maquettes de tableaux pour réaliser le tri des termes, outils de gestion de thésaurus et de traitement des données) et l'expertise nécessaire pour les utiliser et en exploiter les résultats. A travers leur expérience de la création de vocabulaires contrôlés, ils

ont tenu un rôle de conseil méthodologique tout au long du projet. Une fois les termes validés par les scientifiques, les équipes de l'Inist ont travaillé à la structuration finale de Paleosaurus, réalisé la conversion des données aux formats du web sémantique (SKOS), puis l'ont publié sur la plateforme Loterre, conçue et maintenue par l'Inist.

Ce projet a donc mobilisé des compétences techniques et scientifiques. Les compétences techniques ont permis notamment d'automatiser au maximum le processus, là où l'expertise scientifique, et disciplinaire en particulier, est indispensable à la sélection des termes puis à la construction du thésaurus. Le succès de Paleosaurus dépend maintenant de son appropriation par la communauté scientifique visée. Là encore, les expertises et les réseaux complémentaires des professionnels de l'information scientifique et technique (IST) et des paléoclimatologues devront continuer de se mobiliser afin de développer les usages potentiels et tout à fait prometteurs de ce thésaurus.

Références bibliographiques

Dalbin, Sylvie, « Thésaurus et informatique documentaires. Partenaires de toujours ? », *Documentaliste-Sciences de l'Information*, vol. 44, no. 1, 2007, pp. 42-55

Durost, Sébastien ; Reich, Guillaume ; Girard, Jean-Pierre, *Terminologies, modèles de données archéologiques et thésaurus documentaires : réflexions à partir d'une typologie de céramique*, 2021, HAL Id : ⟨hal-03278684⟩, version 1

Isaac, Antoine, « Entre thésaurus et ontologies : une affaire d'interopérabilité et d'alignement » in *Documentaliste - Science de l'Information* (Dossier Web sémantique, web de données), vol. 48, n°4, 2011, pp. 48-49

Isaac, Antoine, « Les référentiels : typologie et interopérabilité », in Lisette Calderan, Pascale Laurent, Hélène Lowinger et Jacques Millet (coord.), *Le document numérique à l'heure du web de données*. Séminaire Inria, Carnac, 1er-5 octobre 2012, ADBS Editions, Paris, 2012, pp.85-104. Version HAL : https://hal.inria.fr/hal-00740282v2

Isaac, Antoine et Bouchet, Thierry, « Rameau et Skos », *Arabesques*, 54, 2009, 13-14, référence électronique : DOI : 10.35562/arabesques.2109

Khayari, Majid; Reszetko, Véronique; Vachez, Dominique; Vedovotto, Nathalie; Yon, Jérémy et al., « De TermSciences à Loterre : comment l'Inist-CNRS a rendu les terminologies ouvertes plus conformes aux principes FAIR », 2021, https://hal.archives-ouvertes.fr/hal-03176063/

Garnier, Eric; Stahl, Ulrike; Laporte, Marie-Angelique; Kattge, Jens; Mougenot, Isabelle, et al., « Towards a thesaurus of plant characteristics: An ecological contribution », *Journal of Ecology*, 2016, 105 (2), pp.298 – 309, https://doi.org/10.1111/1365-2745.12698

Nouvel, Blandine, « Le thésaurus PACTOLS, système de vocabulaire contrôlé et partagé pour l'archéologie », *Archéologies numériques*, vol 3, n°1, 2019, DOI: 10.21494/ISTE.OP.2019.0356

Perrin, Emmanuelle, « Thésaurus et interopérabilité des données archéologique : le projet HyperThesau », *Humanités numériques*, 4, 2021, DOI : https://doi.org/10.4000/revuehn.2384

Thébault, Vincent, « BiblioLabs, un outil au service du pilotage de l'université Paris-Saclay », *Arabesques*, 96, 2020, référence électronique : DOI : 10.35562/arabesques.1478

Vachez, Dominique, « Étude comparative de thésaurus en Sciences de l'Environnement. Bonnes pratiques de conception et FAIRisation de thésaurus », 18 juin 2021, *https://hal.science/hal-03264803*

Abréviations

ANR : Agence Nationale de la Recherche

DiBISO : Direction des Bibliothèques, de l'Information et de la Science Ouverte

FAIR : Findable, Accessible, Interoperable, Reusable

Inist : Institut de l'information scientifique et technique

ISO : International Organization for Standardization

IST : Information scientifique et technique

ISTEX : Information scientifique et technique d'excellence

GEOPS: Laboratoire Géosciences Paris-Saclay

KOS: Knowledge Organization Systems

LOTERRE : Linked open terminology resources

LSCE : Laboratoire des Sciences du Climat et de l'Environnement

SKOS: Simple Knowledge Organization Systems

WOS: Web of Science