Projet collaboratif CNR/INIST : bilan et perspectives

Rédacteurs : Florence Menard, Virginie Triboulin, Thierry Bouchet

1. Introduction

Le projet d'enrichissement terminologique collaboratif mené de 2016 à 2020 par l'INIST[1] et le CNR[2] vient de s'achever. Conçu d'emblée comme preuve de concept, comme test pour la révision du vocabulaire Rameau par apport massif de données extérieures et retraitement de ces données pour une exposition en open data, il a été pour ses deux protagonistes une expérience tout à fait neuve et, pour le moment du moins, unique. Il nous semble donc nécessaire à ce stade d'y jeter un regard rétrospectif et d'en proposer un bilan qui soit le plus factuel possible. Nous situerons tout d'abord brièvement le contexte et les enjeux du projet puis son périmètre et son déroulement. C'est seulement à l'issue du bilan quantitatif et qualitatif qui en découle que nous pourrons envisager l'intérêt de collaborations du même type et, le cas échéant, les pistes et voies d'amélioration les plus pertinentes.

2. Contexte et enjeux

Les acteurs :

L'INIST articule son action autour de trois axes principaux : l'analyse et la fouille textuelle de la production scientifique, la valorisation des données de la recherche et l'accès à l'information scientifique pour les chercheurs.

Les ressources terminologiques de l'INIST sont :

- Multidisciplinaires
- Riches : plus de 200 000 concepts
- Bilingues (anglais) voire tri ou quadrilingues (espagnol et allemand)
- Mises à jour et enrichies en continu jusqu'en 2014
- Destinées à un usage interne à l'origine, mais certaines sont désormais proposées en téléchargement pour un usage externe, sur le portail Loterre

Mais

- Peu structurées, avec une arborescence minimale dans certains domaines, et parfois redondant et non sourcé

[1] Institut de l'information scientifique et technique (CNRS)
[2] Centre national Rameau (Bibliothèque nationale de France / département des Métadonnées)

Le vocabulaire Rameau est :

- Multidisciplinaire
- Riche : un peu moins de 200 000 concepts
- Bilingue pour un grand nombre de concepts, avec une équivalence LCSH et/ou MeSH
- Mis à jour et enrichi en continu
- Destiné à un réseau
- Très structuré, présentant une arborescence forte notamment dans les domaines revus

Mais

- Présentant de nombreuses redondances et des sources lacunaires dans certains domaines

L'INIST et le Centre national Rameau ouvrent également leurs données :

- Via le Catalogue général et data.bnf pour Rameau
- Via différentes plateformes pour l'INIST : TermSciences (ouverture en 2005 mais les ressources ne sont plus actualisées depuis 2007), Ortolang (depuis 2016) et Loterre (depuis 2018).

Le CNR a mis en place en 2007 un programme de révision systématique du vocabulaire qui a permis de revoir plusieurs sous-domaines, notamment scientifiques et plus particulièrement dans le domaine de la taxonomie : ces domaines sont intéressants car ils sont nativement très structurés, avec une forte volumétrie, mais avec des données (noms vernaculaires et noms latins) assez pauvres.

3. Périmètre et déroulement

Dans le *domaine du vivant*, le choix s'est porté sur la taxonomie des poissons car ce domaine n'avait pas été revu et compléterait d'autres taxonomies revues telles que celles des mammifères et des oiseaux.

L'INIST a ensuite fourni un fichier SKOS contenant 5564 concepts (alignement entre Catalogue of Life et Uniprot). Ce fichier a été complété par des espèces fossiles issues de l'alignement entre les vocabulaires de Taxonomicon et de Vertebrate Taxonomy Ontology.

Après exclusion des espèces, le fichier de travail final comporte 5547 concepts. La hiérarchie du domaine a ensuite été remise à plat car elle avait évolué et le vocabulaire a été enrichi de nombreuses sources manquantes, de noms vernaculaires et de formes latines.

Dans le *domaine des sciences et techniques*, le choix s'est fait par défaut : les techniques ont d'abord été éliminées, ce domaine étant dans Rameau encore assez mal structuré et défini, puis l'optique fut écarté au profit des transferts de chaleur.

Le fichier SKOS fourni par l'INIST concernant ce domaine contenait 1642 concepts.

Une extraction des données du catalogue de la BnF concernant toutes les notices typées « Physique » a retourné 1927 notices. Les données fournies par l'INIST débordant très largement le cadre des transferts de chaleur (comme des « équipements », des « techniques », etc.), il a été décidé de revoir l'intégralité du domaine des transferts de chaleur au sein des autorités Rameau et d'y mener un travail d'ajout de sources et de clarification des liens sémantiques.

4. Bilan

Si le bilan quantitatif est relativement simple à présenter et montre assez clairement le niveau volumétrique auquel se situe l'apport terminologique du projet :

Taxonomie des poissons : 2282 taxons ont été créés

Thermodynamique/Transfert de chaleur : 66 notices révisées

Le bilan qualitatif nécessite de faire intervenir plusieurs critères d'analyse qui ne se situent pas tous au même niveau.

Il convient tout d'abord de souligner que les principaux objectifs du projet ont été atteints :

- Alignement terminologique entre référentiels
- Enrichissement sur un mode collaboratif
- Amélioration globale de la qualité des domaines revus. Ce processus d'amélioration a également touché les autorités périphériques qui n'entraient pas dans le périmètre précis du projet (assainissement du sous-domaine)

Pour le CNR :

- De nouvelles compétences, portant sur les processus d'extraction, d'alignement via Excel, d'échange de fichiers SKOS ont été acquises et pourront, le cas échéant, être mises à profit dans des projets du même type ou dans d'autres contextes ;
- La révision du vocabulaire Rameau par apport et retraitement de données exogènes a permis un enrichissement qui n'aurait jamais pu avoir lieu à un tel niveau dans le cadre du processus habituel de mise à jour et de réorganisation de ses données propres ;
- La relecture des formes latines a constitué un contrôle qualité particulièrement pertinent ;
- L'exposition des données en lod sur Loterre (INIST) permet un rebond "vertueux" vers les pages data.bnf.fr et offre une visibilité accrue des autorités Rameau, en particulier auprès d'un public scientifique plus habitué à TermSciences qu'à Rameau ;

- La circulation et l'échange de données permis par ce projet ont toute leur place dans l'économie globale de la Transition bibliographique et potentiellement du FNE[3] ;
- Collaborer avec une institution telle que l'INIST pour la révision du vocabulaire Rameau est un atout supplémentaire dans la politique d'ouverture et de coopération du CNR

Pour l'INIST

- Sourçage des autorités, nettoyage des redondances, hiérarchisation des concepts
- Réutilisation par les chercheurs de données alignées avec un vocabulaire national contrôlé
- Exposition des données sur Loterre en open linked data pour une circulation et une visibilité accrues
- Intérêt d'une coopération avec la BnF

5. Perspectives : poursuivre, dans quelles conditions ?

S'il est clairement envisageable de poursuivre des projets d'enrichissement collaboratifs entre le CNR et l'INIST, il n'est pas interdit d'en améliorer le cadre et le périmètre.

Le cadre, ce pourrait être par exemple, une convention permettant de formaliser ce qui ne pourrait plus être désormais une simple preuve de concept.

Il faudrait également réfléchir au choix du ou des domaines concernés et envisager de traiter un domaine revu, moins risqué en termes d'inflation volumétrique, ou, s'il s'agit d'un domaine Rameau non-revu, de le choisir d'une taille plus modeste.

Un élément-clé de maîtrise du temps réside dans la garantie d'un interlocuteur unique et stable (pour un domaine donné) qui puisse consacrer une part régulière et dédiée de son activité à la poursuite du projet.

[3] Fichier national d'entités