Ar(abes)ques

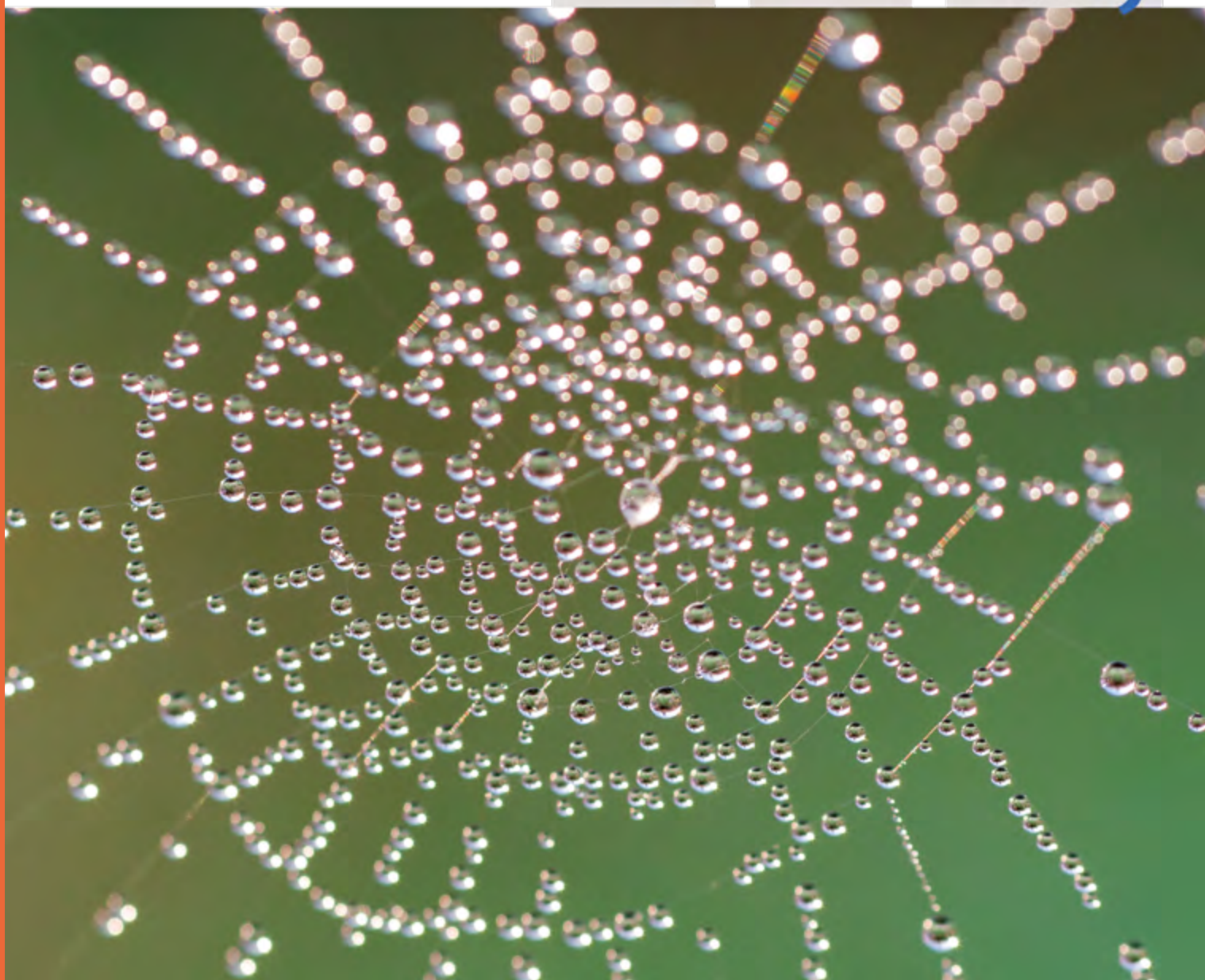
JANVIER - FÉVRIER - MARS 2024

DOSSIER

Autorités et référentiels : *le nouveau paradigme*

112

abes
agence bibliographique de l'enseignement supérieur



Sommaire

(Dossier)

Autorités, identifiants : le nouveau paradigme

Ar(abes)ques REVUE TRIMESTRIELLE DE L'AGENCE BIBLIOGRAPHIQUE DE L'ENSEIGNEMENT SUPÉRIEUR,

227, avenue du Professeur Jean-Louis Viala, CS 84308, 34193 Montpellier cedex 5. Tél. 04 67 54 84 10 - https://abes.fr / La revue Arabesques est accessible en ligne via la plateforme Prairial : https://publications-prairial.fr/arabesques
Directeur de la publication : Nicolas Morin. Coordination éditoriale et secrétariat de rédaction : Véronique Heurtematte. Comité de rédaction : Christophe Arnaud, Jean-Marie Feurtet, Christine Fleury, Etienne Naddeo, Laurent Piquemal, Marie-Pierre Roux (Abes), Yann Marchand (SCD université de Nantes).
Iconographie rassemblée par Christophe Arnaud.
Conception graphique : Anne Ladevie / Atelier à suivre www.anneladevie.com
Impression : Pure Impression.

ISSN (papier) 1269-0589/ISSN (web) 2108-7016

(Le billet du directeur...)

J'évoquais dans mon billet de juillet dernier le travail accompli tout au long du premier semestre 2023 de recueil des besoins et des attentes exprimés par les uns et les autres à l'égard de l'Abes. Il s'agissait d'analyser notre offre et nos services existants, ainsi que les besoins des organisations de l'ESR que nous desservons. Un certain nombre de sujets avaient ainsi émergé, par exemple autour des référentiels, de la nécessité de disposer de plus grands volumes de métadonnées, en particulier pour la documentation électronique, ou d'une meilleure interconnexion entre les systèmes informatiques de l'Abes et ceux des établissements.

J'indiquais en juillet qu'une fois ce travail de questionnement préparatoire achevé, le second semestre 2023 serait consacré à imaginer des réponses et à faire de premiers choix pour l'établissement. Cet effort a abouti fin novembre 2023 au vote, par le Conseil d'administration de l'Abes, du projet d'établissement 2024-2028.

Ce document[1] fixe les lignes directrices de notre action pour les cinq années qui viennent, mais il propose aussi une orientation de plus long terme, alors que doit s'engager bientôt un travail sur la refonte de l'article 2 du décret de création de l'Abes, qui définit nos missions.

L'Abes fournit des métadonnées aux établissements de l'ESR : cette mission, qui s'incarnait historiquement dans un catalogue collectif, ainsi que dans des outils de production et de diffusion des métadonnées à destination des établissements, est en train de changer profondément.

Les acteurs du secteur documentaire se sont, depuis la LRU en particulier, intégrés toujours plus dans la vie de leur établissement. Dans le même temps, les données issues du secteur documentaire sont désormais exploitées pour des usages de plus en plus variés, pour la recherche, le pilotage des établissements ou les ressources pédagogiques par exemple.

L'Abes se doit d'accompagner ces transformations : ce projet d'établissement cherche à traduire concrètement cette ambition.

Voici quelques-unes des actions phares qui, dans les cinq prochaines années, contribuent à cet objectif général.

L'Abes doit proposer des volumes plus importants de métadonnées, qu'il s'agisse des bouquets commerciaux de documentation électronique ou des imprimés acquis par les établissements.

L'Abes doit également proposer des flux de données plus complets, systématiques et automatisés entre les systèmes d'information locaux (outils documentaires, archives ouvertes, ERMS d'établissement ou de Couperin, bibliothèques numériques, outils de pilotage...) et ses propres systèmes.

Cette politique de volumes et de flux doit enfin s'appuyer sur une politique de qualité des données, qui doit permettre d'indiquer la fiabilité des métadonnées proposées et favoriser ainsi leur meilleure exploitation.

Notre valeur ajoutée peut être plus importante sur certaines catégories de métadonnées que sur d'autres, et il est proposé par exemple d'investir tout particulièrement dans les données de référentiels Agent, qui englobent les personnes (ORCID et IdRef) et les collectivités (en particulier les organisations de l'enseignement supérieur et de la recherche).

Cette politique des données est au cœur du projet d'établissement parce qu'elle est au cœur du service que nous apportons aux organisations de l'ESR. Cette ambition nécessitera la mise en œuvre de nouveaux moyens. En particulier, elle implique le renouvellement du système de gestion des métadonnées de l'Abes qui, pour sa partie centrale, le Sudoc, est en production depuis la fin des années 1990. Cette réinformatisation représente, pour la période 2024-2028, la part la plus importante du travail que nous devrons fournir, et des sommes que nous devrons investir, mais elle n'est pas une fin en soi et reste un moyen au service d'une politique des métadonnées pour les établissements.

De même que l'Abes devra adapter son organisation interne pour permettre la réalisation des objectifs du projet, les modalités de dialogue et de collaboration avec tous ceux qui utilisent nos services et données devront aussi être repensées : les sujets que nous abordons aujourd'hui ont évolué, nos interlocuteurs aussi.

Les établissements attendent également de l'Abes qu'elle éclaire, dans ses domaines d'expertise, les choix technologiques et stratégiques de demain. Le projet 2024-2028 répond à cet enjeu par un renforcement significatif de notre activité de recherche et développement autour des thématiques de l'intelligence artificielle.

Enfin, parallèlement à son activité autour des données et logiciels documentaires, l'Abes joue de longue date maintenant un rôle crucial dans les acquisitions de documentation électronique. Ce travail, qui a crû au fil des années, conserve une part d'informel auquel ce projet tentera de répondre. Trois axes de progression ont été identifiés avec Couperin et le ministère : la soutenabilité économique de la mission, la définition de règles permettant d'identifier les négociations d'envergure nationale que l'Abes devrait porter, et enfin un renforcement de la coopération avec Couperin sur l'ensemble du processus menant de la négociation initiale à la contractualisation et au suivi des marchés.

J'espère que vous retrouverez dans ce projet 2024-2028 non pas nécessairement une réponse à chaque attente, mais une clarification des missions et actions, une ambition pragmatique qui soit adaptée à la réalité des établissements, et une vision partagée de l'avenir des métadonnées dans l'ESR.

Nicolas Morin
Directeur de l'Abes

[1] «Consulter le Projet d'établissement : **https://abes.fr/publications/publications-institutionnelles/projet-etablissement-2024-2028**

OUVERTURE DU SITE « code.rdafr.fr »

Reflet d'une collaboration aboutie entre l'Abes et la Bibliothèque nationale de France, la mise en ligne de ce site renouvelle les modalités de consultation du code de catalogage RDA-FR.

Le standard RDA-FR est développé depuis 2015 dans le cadre du programme national Transition bibliographique pour permettre le catalogage entités-relations, conformément au modèle international IFLA LRM.

DES CONTENUS RESTRUCTURÉS ET ENRICHIS

Reprenant tout le contenu publié à ce jour sous forme de fichiers PDF, le nouveau site web[1] a pour ambition de faciliter la consultation de la norme, la navigation dans ses règles et la recherche d'informations pour les catalogueurs, les formateurs et toute personne qui aurait besoin de consulter ce texte. Le contenu a été réorganisé selon une logique de présentation par entité (œuvre, expression, agent, etc.).
Il dispose naturellement d'un moteur de recherche, mais il fait aussi bénéficier ses utilisateurs de liens hypertextuels nombreux afin de diminuer le temps nécessaire pour accéder à une règle précise, chacune disposant de son permalien.

Le site propose aussi des contenus enrichis :

- Les alignements avec l'ontologie RDA-FR sont affichés pour chaque élément et seront ajoutés au fur et à mesure de l'accroissement du code et de l'ontologie
- Les référentiels sont accessibles sur des pages dédiées et peuvent être téléchargés
- Des ressources et des schémas sont publiés dans les sections dédiées, en fonction de leur disponibilité
- Des actualités rendent compte du rythme des publications, mais aussi des mises à jour ponctuelles inévitables dans la vie d'un tel corpus de textes.

UNE COLLABORATION ENTRE ABES ET BNF AUTOUR D'UN PROJET COMPLEXE

L'ampleur du travail était réelle : il s'agissait de convertir 1 200 pages PDF en 450 pages HTML environ, en 6 mois, selon une architecture ne correspondant pas complètement à la structuration du code initial. Ce projet a nécessité de mettre en place une équipe de 7 personnes (6 pour l'Abes, 1 pour la BnF), dont 2 stagiaires recrutées par l'Abes sur le projet.
Dans une phase initiale, une étude menée selon les principes de l'*UX Design* avait permis en 2022 de maquetter le fonctionnement général du site. Le projet est entré dans sa phase de réalisation au printemps 2023. Après avoir préparé le découpage intellectuel du texte avec l'aide des pilotes des groupes de normalisation responsables de la rédaction de RDA-FR, l'équipe a préparé et migré les fichiers vers le site. La réalisation de scripts automatisant le nettoyage des fichiers originaux a permis d'accélérer le processus de migration. Après la mise en place de l'arborescence du site, son ergonomie et ses fonctionnalités (recherche, citabilité, accessibilité, téléchargement de fichiers…) son ergonomie et ses fonctionnalités ont été améliorées.
Cette conversion sans réécriture de RDA-FR, si elle permet un bien meilleur accès au texte du code, donne aussi à voir la complexité de son contenu et parfois des incohérences liées à la longueur du processus de rédaction. C'est l'occasion, dans les prochains mois, en même temps que s'achèvera la rédaction du cœur du code RDA-FR, de simplifier celui-ci pour y faire mieux ressortir les règles de catalogage.
N'hésitez pas à nous faire part de vos retours via le formulaire prévu à cet effet sur ce site.

[1] https://code.rdafr.fr

LE PROJET FNE reporté

L'objectif du Fichier national d'entités (FNE), tel qu'il avait été initialement conçu lors de son lancement en 2018, était de mettre à disposition des professionnels de la culture et de l'enseignement supérieur une base centralisée de coproduction d'entités conformes au modèle conceptuel international IFLA LRM. Si le Comité stratégique bibliographique, dans sa séance du 28 juin 2023, a réaffirmé l'importance stratégique des ambitions portées par ce projet, il en a toutefois lié la réalisation au renouvellement des outils de production des deux agences bibliographiques. Un report de plusieurs années a donc été décidé, pour assurer au FNE une assise stable et durable. Le matériel documentaire accumulé lors des étapes du projet a été archivé par la Bibliothèque nationale de France et l'Abes de manière à pouvoir être réutilisé lors de la relance du projet.

AUTORITÉS ET RÉFÉRENTIELS : LE NOUVEAU PARADIGME

« Parler avec Autorités » nous invite l'ADBU dans son Manuel d'instruction à l'usage du bibliothécaire débarquant en infodoc en 2023[1] ! Beaucoup se réjouiront donc de la thématique de ce numéro d'*Arabesques*. Un précédent dossier[2] en 2017 donnait déjà à voir l'importance des identifiants, les usages d'IdRef hors de l'Abes, et la complémentarité homme/machine dans les tâches de liage des ressources documentaires à des notices d'autorité. En 2023, en dressant les 6 points forts de l'Abes, l'évaluation du Hcéres saluait « la qualité et la pertinence de l'offre de services, en particulier des référentiels, dont IdRef »[3].

De fait, grâce à une politique d'alignement entre identifiants et d'agrégation de ressources, IdRef est au coeur d'un graphe de connaissance intéropérable, dont les frontières sont étendues et extensibles. Ce graphe peut être exploité par des établissements pour nourrir en interne leur propre système d'information local en l'adossant à des données valides et contrôlées. Et bien entendu, il est ouvert et réutilisable par des chercheurs pour rendre leurs jeux de données FAIR.

Décrire l'Enseignement supérieur et la Recherche d'aujourd'hui dans IdRef à travers la production scientifique, c'est contribuer à la valorisation de la France à l'international, conformément à la stratégie ministérielle des données, algorithmes et code sources, en s'appuyant sur ORCID et ROR, deux identifiants internationaux compatibles avec la science ouverte.

Cependant, les données de référence, ce ne sont pas que des personnes ou des structures ! Comme le montre le dynamisme de Geonames, la description des toponymes n'est pas l'apanage des bibliothécaires. Enfin, n'oublions pas l'indexation Rameau. Si la syntaxe tend à se simplifier, la richesse du vocabulaire est un défi à relever et exploiter pour les machines, dans l'optique de seconder les humains.

Ce numéro d'*Arabesques* se clôt par le portrait d'un correspondant Autorités. À travers lui, l'Abes remercie l'ensemble des collègues qui produisent et enrichissent quotidiennement IdRef. Vous êtes les infatigables artisans d'un bien commun essentiel. Merci !

[1] https://adbu.fr/actualites/parler-linfo-doc-vocabulaire-et-notions-de-base

[2] *Arabesques* n° 85, avril - mai - juin 2017, Autorités, identifiants, entités : l'expansion des référentiels https://publications-prairial.fr/arabesques/index.php?id=201

[3] Rapport d'évaluation de l'Agence bibliographique de l'enseignement supérieur (Abes), Rapport publié le 10/05/2023, p. 31. https://www.hceres.fr/fr/rechercher-une-publication/agence-bibliographique-de-lenseignement-superieur-abes-1

IdRef : un réseau de professionnels pour des référentiels *au service de l'Enseignement supérieur et la Recherche*

Fort d'une communauté professionnelle de plus de 240 correspondants Autorités coordonnée par l'Abes, IdRef est le fruit d'un travail collectif qui produit des référentiels de qualité avec l'objectif de s'ouvrir toujours plus largement au-delà du monde des bibliothèques.

Saviez-vous que le groupe *Ouest-France* exploite IdRef pour repérer des chercheuses contemporaines afin de féminiser ses colonnes quand le recours à un expert est requis ? Et que, depuis cet été, de discrets liens hypertextes sont apparus sous le nom de chercheurs français cités dans des articles[1] de la rubrique Sciences du journal *Le Monde* ?
L'avenir dira si cet usage d'IdRef par des médias de renom est une préfiguration notable ou un simple feu de paille. Une chose est sûre : la communauté des professionnels de l'information qui œuvrent tous les jours pour décrire avec justesse les ressources qui constituent leurs collections représente la plus grande richesse d'IdRef. Ce gage de confiance a autorisé IdRef à mener plus loin que tout autre fichier d'autorités européen ou international l'ouverture et la coproduction de données, devenant une véritable singularité dans le paysage actuel.

DU MONDE DES BU…

En 2009, en parallèle du Sudoc, l'Abes développe deux nouvelles bases de signalement pour des objets spécifiques : Calames pour les archives et manuscrits ; STEP/STAR (puis theses.fr) pour les thèses de doctorat. IdRef nait ainsi pour urbaniser ces applications et offrir aux « données qui font autorité » l'autonomie et la réutilisabilité nécessaires pour y adosser toutes les autres.
Application *full web,* IdRef offre une vaste gamme de fonctionnalités. Pour le catalogueur[2], IdRef se branche directement à son environnement de signalement (ressource éditoriale dans le Sudoc, thèse électronique dans STAR, article numérisé dans Persée, vidéo dans Canal-U, etc.). Le *workflow* est fluide et les informations bibliographiques sont récupérées pour éviter les doubles saisies.
Pour les machines, IdRef fournit des identifiants uniques et pérennes (mine de rien, c'est un « service »[3] !) et des API qui optimisent le transit des données.
Du point de vue juridique, les données d'IdRef sont ouvertes sous licence Etalab et respectueuses du RGPD.
L'interopérabilité découle des fonctionnalités précédentes : reposant sur l'interfaçage applicatif et sur le mécanisme de l'identifiant pivot, les échanges d'informations normalisées, fiables et actualisées sont simplifiés et la cohérence globale des données renforcée.

Vue aérienne d'une plantation de thé

Crédit Adobe stock, par Sunbrothers

[1] https://www.lemonde.fr/series-d-ete/article/2023/08/15/guillaume-cabanac-le-sisyphe-de-la-depollution-de-la-science_6185437_3451060.html

[2] Le vocabulaire peut varier : catalogueur dans le Sudoc, contributeur dans Canal-U, convertisseur chez cairn.info… Derrière cette variété de libellés existent des compétences similaires : le signalement.

[3] Cf. article d'Isabelle Blanc.

[4] Avec abandon de l'ancienne gestion des autorités dans Koha

[5] Cf. article sur les alignements d'Isabelle Mauger Pérez.

Rappelons que le réseau Sudoc est la matrice constitutive d'IdRef : à l'origine de bon nombre des 6 millions de notices d'autorité actuelles, il continue d'alimenter l'immense réservoir des personnes physiques et morales, de concepts et des lieux mentionnés dans les documents possédés par les BU. Pour sa part, le réseau des thèses de doctorat, dont le circuit national de signalement est modernisé et fiabilisé depuis 2010, est le second vecteur de données d'autorité en volume et en « qualité » via le signalement de plus de 12 000 thèses de doctorat par an. Points d'entrée dans la carrière académique, les thèses sont des condensés de relations entre personnes, structures et travaux universitaires, soit un poste d'observation privilégié sur la recherche française contemporaine.

… AU MONDE DOCUMENTAIRE FRANCOPHONE…

Si la démarche d'urbanisation d'un système d'information tire profit du recours à IdRef, les modalités en sont variées : simple réutilisation pour Univ-Droit – le portail universitaire du droit ; contribution active sans perte des indispensables données de gestion des droits pour Persée ; adoption pleine et entière pour Frantiq[4], etc.

Au gré des intégrations d'IdRef au sein d'opérateurs nationaux puis européens, l'Abes s'est dotée d'une boite à outils complète. Quand un potentiel contributeur se manifeste, l'un des gros défis réside dans la reprise du gisement documentaire qu'est sa base de données : c'est à cette fin qu'un service d'alignement pour les identifiants Personnes a été développé, qui fonctionne tous azimuts[5].

Il fallait une telle palette de services pour satisfaire des besoins de professionnels plus éloignés du Sudoc (cairn.info ou Canal-U) ou pour attirer à l'international les collègues belges de l'université de Liège et les collègues suisses. L'Abes a ainsi réussi à intégrer dans IdRef des réseaux de catalogage dont le volume de données et le réseau sont de tailles conséquentes : Frantiq, et les trois réseaux francophones suisses – Rero+, SLSP et Renouvaud. Cette extension d'IdRef au-delà de l'Abes illustre le passage du « monde clos » des silos à « l'univers infini »[6] de la *big data* et du web sémantique.

PUIS AU MONDE ACADÉMIQUE…

IdRef se positionne comme un service incontournable ayant vocation à innerver la sphère académique dans sa double dimension de pilotage bibliométrique et de production de connaissances. Côté pilotage, IdRef fluidifie la navigation du local (annuaires LDAP, gestion des thèses de doctorat dans ADUM) à l'international (ORCID[7] et prochainement ROR[8]). IdRef dispose d'une couverture et d'une fiabilité très élevées, comme en atteste son utilisation massive dans l'outil national scanR, moteur de la Recherche et de l'Innovation. Côté scientifique, IdRef est utilisé dans des jeux de données de la recherche, notamment en humanités numériques. Les caractéristiques des autorités (normalisation, désambiguïsation, structuration) rendent davantage « FAIR » les connaissances produites. Et si face à la montagne des données de recherche, il faudra être sélectif, la communauté fourbit ses armes[10] en vue d'assurer à ces précieuses données la plus riche intégration possible dans le web sémantique, dont les promesses en termes d'accroissement de la connaissance sont colossales et seront bientôt tenues.

LA FORCE D'UN RÉSEAU POUR « IDREFISER » LES DONNÉES DE L'ESR

IdRef est un bien commun, fruit d'une communauté dont la vitalité prend racine dans quelques principes fondateurs : le principe de confiance, intrinsèquement lié au catalogage partagé, et celui de la souveraineté des données produites, qui a pour corollaire le principe de responsabilité, les 200 correspondants Autorités[11] jouent ici un rôle majeur.

Saluons l'expertise forgée par de solides pratiques métier et la volonté constante de se former[12].Surtout, réjouissons-nous de l'engagement à servir, non seulement pour son propre bénéfice mais aussi pour d'autres initiatives. Que l'on pense à l'allant régnant au sein du chantier collaboratif de curation des quelques milliers de notices élémentaires dites « Tp1 », ou au chantier de création d'éditeurs français pour le corpus Mir@bel, ou encore à l'usage de paprika.idref.fr pour rendre à chaque auteur la paternité de ses créations en cas d'imbroglio.

Ces valeurs soudent la communauté, ce qui est crucial (protecteur et prometteur) dès lors que nos données vont servir de corpus d'entraînement à des intelligences artificielles « métier » ou généralistes[13].

L'Abes souhaite maintenant œuvrer à la poursuite du déploiement d'IdRef auprès des archives ouvertes et institutionnelles, des entrepôts de données de recherche, des éditions universitaires, des services d'archives scientifiques, des bibliothèques numériques, etc.

Alors, à vous qui lirez cet article et qui produisez des données de manière isolée, n'hésitez pas à rejoindre la communauté.

Et vous qui êtes déjà membres de la communauté IdRef et êtes les meilleurs ambassadeurs, démarchez vos collègues hors de la BU, au sein de vos établissements, et vantez-leur les bénéfices à nous rejoindre !

François Mistral
Responsable IdRef – Autorités à l'Abes
mistral@abes.fr

[6] Du monde clos à l'univers infini / Alexandre Koyré, 1973 : La pensée philosophique et scientifique a accompli une révolution profonde aux XVI^e^ et XVII^e^ siècles.

[7] Cf. entretien avec Nicolas Fressengeas

[8] Cf. article de Carole Melzac

[9] Cf. article du comité SoViSu+

[10] Cf. article de F. Beretta

[11] Le portrait de ce numéro d'*Arabesques* met en avant l'un d'entre eux !

[12] Cf. article d'Aurélie Faivre et Héloïse Lecomte

[13] Cf. article Yann Nicolas

INTERVIEW

« LA RECHERCHE FRANÇAISE NE PEUT FONCTIONNER EFFICACEMENT QU'EN RESTANT CONNECTÉE AU RESTE DU MONDE. »

Entretien avec Nicolas Fressengeas

Vice-président de l'université de Lorraine en charge du numérique, des données et de science ouverte, chargé de mission science ouverte au MESR pour son volet international et élu au conseil des représentants d'ORCID.

Crédit photo Université de Lorraine

Crédit Adobe stock, par Abdul Momin

Pouvez-vous nous présenter la gouvernance d'ORCID ?

ORCID est une organisation à but non lucratif enregistrée dans l'État du Delaware aux États-Unis qui compte une quarantaine d'employés[1] répartis sur la surface de la planète sous la direction d'un *Chief Executive Officer (CEO)*. La gouvernance de l'organisation se fait à travers le conseil des représentants, traduction en français de *board,* selon le site d'ORCID, et qui correspond peu ou prou au conseil d'administration, lequel est assisté par cinq comités, chargés de l'exécution de la stratégie impulsée par le conseil d'administration, des finances, de la stratégie envers les membres, de l'audit des risques pris par l'organisation, et de la nomination des membres du conseil des représentants. Les réunions du conseil des représentants se font en présence et avec l'aide du CEO et des responsables des équipes opérationnelles d'ORCID[2].

Quels sont les grands axes stratégiques qu'elle impulse ?

Les grands axes de la stratégie d'ORCID pour 2022-2025[3] sont l'amélioration des services rendus aux membres, aux chercheuses et chercheurs et l'amélioration de la couverture mondiale, le tout en assurant la continuité de la confiance que l'on peut avoir dans la stabilité et la pérennité de l'infrastructure d'ORCID.

Vous mentionnez l'amélioration de la couverture mondiale. La volonté d'ORCID d'être une infrastructure internationale s'adressant à tous les acteurs du monde de la recherche se reflète-t-elle dans la composition actuelle du *board* ?

L'amélioration du caractère international d'ORCID est l'une des quatre priorités affichées par la stratégie 2022-2024. Toutefois, la composition actuelle du conseil des représentants[4] montre une domination des pays anglo-saxons du nord global. Cet état de fait est reconnu par le conseil ; son ambition est donc d'améliorer la représentativité des communautés internationales. Il se trouve que le renouvellement d'une partie du conseil[5] fin 2023 a attiré un grand nombre de candidatures de qualité de la part de régions sous-représentées, ce qui a permis de proposer une amélioration significative de la composition du conseil.

Quel a été votre propre parcours pour être élu au *board* d'ORCID ?

Je ne peux que faire des hypothèses quant aux raisons précises qui m'ont amené au *board* d'ORCID. L'une d'elles est clairement ma citoyenneté française et européenne, eu égard à la volonté dudit *board* d'améliorer sa représentativité mondiale hors des cercles anglo-saxons. Ma qualité de chercheur encore récemment en activité constitue probablement une autre de ces raisons. Mon implication à la vice-présidence de l'université de Lorraine en charge de la politique numérique a pu être un élément favorable, de même que celle auprès du ministère de l'Enseignement supérieur et de la Recherche et de l'administratrice des données, des algorithmes et des codes.

Vous faites également partie du comité exécutif du consortium ORCID-France et vous connaissez bien le paysage de l'ESR français. Pensez-vous qu'il existe des spécificités françaises qui devraient être mieux prises en compte par ORCID, pour promouvoir ses services, stimuler les adhésions, ou plus globalement améliorer la qualité des données d'ORCID ?

Je n'ai pas aujourd'hui clairement identifié de spécificités françaises qui devraient être prise en compte par ORCID en tant qu'organisation internationale. C'est en réalité heureux car ORCID est une organisation assez petite, constituée de 40 personnes réparties dans le monde, et l'équipe en charge du « produit », qui est le sommet de l'iceberg que nous voyons tous, ne compte que trois personnes, ce qui rend difficile la prise en charge de besoins locaux. Toutefois, le code d'ORCID étant libre, il est possible à tous de proposer des améliorations aux équipes d'ORCID, et ce pourrait être une voie à explorer pour des besoins spécifiques.
Le développement du nombre d'adhésions prend naturellement une grande part dans la stratégie d'ORCID, et l'organisation reste attentive aux stratégies nationales. Le principal enjeu relatif à l'augmentation du nombre d'adhésions à ORCID est, selon moi, la réalité de la plus-value apportée aux membres. De la même façon, les chercheuses et chercheurs porteront une attention accrue à la qualité des données contenues dans leur dossier ORCID quand celles-ci seront utilisées afin de les soulager de tâches administratives. Ces deux enjeux sont corrélés. Dans la mesure où la feuille de route de la politique des données, des algorithmes et des codes sources du MESR[6] inclut spécifiquement l'usage d'ORCID au niveau national, je ne suis pas inquiet : les chercheurs et leurs institutions sont pragmatiques.

Le risque de bascule de la gouvernance aux mains d'intérêts uniquement privés est parfois soulevé. Est-ce un frein à l'adhésion ? Existe-t-il des mécanismes pour prévenir ce risque ?

ORCID est une organisation intrinsèquement à but non lucratif par son statut législatif et ne peut donc faire de bénéfices que s'ils sont réinvestis. Cela limite son intérêt pour les appétits privés. Sa transparence et son ouverture à la diversité de l'écosystème de la recherche sont au cœur de ses valeurs et principes fondateurs[7]. Un conseil des représentants monochrome, représentant les intérêts d'une catégorie unique d'acteurs est non seulement contraire à ces valeurs, mais nuirait probablement à l'adhésion de l'ensemble des communautés. Les risques d'une telle bascule, d'un côté ou de l'autre, sont à mon avis faibles, et la meilleure preuve en est l'évolution actuelle du conseil.

Comment articuler une stratégie nationale, cruciale pour la recherche française, avec la stratégie d'une organisation supranationale telle qu'ORCID, ce qui soulève parfois certaines questions au sein de la communauté de l'ESR, notamment concernant la sécurité des données, ou la pérennité de la structure ?

La recherche est par nature internationale, et la recherche française ne peut fonctionner efficacement qu'en restant connectée au reste du monde. C'est tout l'intérêt d'appuyer une stratégie nationale sur une organisation internationale. L'enjeu est donc de mettre en cohérence la stratégie internationale d'ORCID et la politique nationale. Le fait de participer à la gouvernance d'ORCID y contribue. D'un point de vue plus opérationnel, je pourrais plaider pour une coopération technique entre les équipes d'ORCID et des développeurs nationaux, sur le modèle des communautés du logiciel libre, afin de faire progresser le produit dans les directions que nous souhaitons.
La question de la confidentialité des données se pose peu à ORCID, dont les données sont publiques pour plus de 90 % d'entre elles. La question de la pérennité de la structure est en revanche cruciale. C'est précisément l'objet d'un des quatre axes de la stratégie d'ORCID et l'une des principales préoccupations de l'organisation. Cette question de pérennité, et donc de soutenabilité, concerne ORCID comme toutes les infrastructures nécessaires à l'ouverture de la science : une utilisation responsable d'une telle infrastructure implique une participation raisonnable à son financement. Néanmoins, une telle participation nécessite un degré de confiance qui ne peut être établi que par l'ouverture et la transparence des processus internes et des politiques organisationnelles, financières et technologiques. Tout en conservant une marge de progression, ORCID est, sur ces plans, en très bonne place, et c'est à travers une analyse détaillée de l'ensemble des données de l'organisation, accessibles sur son site, que l'on pourra se forger son propre degré de confiance.
L'enjeu est là.

Propos recueillis par
Benjamin Bober et Véronique Heurtematte

[1] https://info.orcid.org/orcid-team
[2] La gouvernance d'ORCID est présentée en détail ici : https://info.orcid.org/our-governance
[3] https://info.orcid.org/2020-2025-strategic-plans
[4] https://info.orcid.org/orcid-board
[5] https://info.orcid.org/announcing-the-2024-board-slate
[6] https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2021-09/la-feuille-de-route-2021-2024-du-mesri-relative-la-politique-des-donn-es-des-algorithmes-et-des-codes-sources-12965.pdf
[7] https://info.orcid.org/what-is-orcid

ROR : une base d'identifiants rugissante

Base de données collaborative internationale, ROR[1] (*Research Organizations Registry*) œuvre à identifier de manière univoque toutes les entités liées à la recherche. Depuis l'été 2023, ROR a affiné son niveau de description en intégrant une grande partie des laboratoires publics français.

Ces dernières années, le patient travail des chartes de signature, élaboré dans la majorité des établissements, a consisté à tenter de normaliser la manière dont les informations d'affiliation issues des publications vont figurer dans les bases bibliographiques et bibliométriques. Mais même orthographiées de la même manière et scrupuleusement listées dans le même ordre, si ces informations précieuses restent des chaînes de caractères, cela ne suffit pas pour identifier de manière certaine et pérenne les structures qu'elles désignent. Comment boucher les trous de ce tonneau des Danaïdes pour enfin apparier sans erreur les publications aux organisations ?

ROR est l'acronyme du *Research Organizations Registry,* qu'on peut traduire par Répertoire des organisations liées à la recherche. Cette base de données vise à identifier toutes les entités liées à la recherche : universités, entreprises, organismes, hôpitaux, agences publiques de financement, infrastructures matérielles, etc. Le cas d'usage central de ROR est d'expliciter l'affiliation, dans les publications, c'est à dire le lien qui existe entre une personne physique et une organisation pour une production (le si fameux et difficilement traduisible « *research output* »), le plus souvent un article, mais aussi un jeu de données, par exemple.

L'identifiant ROR (ROR ID) est un PID (*Permanent Identifier, identifiant pérenne*) destiné à isoler et reconnaître de manière univoque, sur la scène internationale, une entité collective à laquelle peuvent être liées des personnes qui font de la recherche. Il se présente comme une chaîne non signifiante de 9 caractères alphanumériques à la suite du préfixe du registre ; ainsi l'ID ROR de l'Abes est https://ror.org/027xymc69. Le rôle de cet identifiant est de se diffuser dans tous les systèmes de gestion qui ont trait à la recherche : plateforme de publication, outil d'évaluation, demande de financement…

Les données de ROR sont accessibles de plusieurs manières : par une recherche directe dans le registre[2], par une requête via l'API[3] ou en téléchargeant le jeu de données complet[4]. À noter que les métadonnées visibles dans le moteur de recherche ne reflètent pas la totalité de celles fournies par l'API.

ROR fonctionne également de manière intéressante pour la curation des données : tout un chacun peut suggérer via un formulaire[5] un ajout ou une modification. Le traitement de ces sollicitations est présenté, en toute transparence, sur une instance GitHub[6]. On peut ainsi observer en direct le travail d'aiguillage des demandes et les modifications opérées par la petite équipe dédiée à la curation, composée d'une personne à plein temps, deux en renfort, sous la houlette d'un groupe dénommé « *Curation Advisory Board* » qui juge de la pertinence des requêtes, notamment pour les nouvelles entités[7].

ROR DANS L'ÉCOSYSTÈME DES IDENTIFIANTS PÉRENNES

La collectivité (dirions-nous en bons bibliothécaires) qui possède un ID ROR n'est jamais bien loin des personnes physiques, lesquelles peuvent posséder un identifiant ORCID. ROR, organisation dont la gestion est partagée par tous les acteurs du circuit de l'édition scientifique, des bibliothèques aux éditeurs privés, a nettement marqué dès le départ son intention d'être un identifiant ouvert, dont le registre soit bâti par et pour la communauté.

Les fées qui se sont penchées sur son berceau sont la *California Digital Library*, ainsi que CrossRef et DataCite, deux poids lourds parmi les agences d'attribution de DOI. Le premier état du référentiel a vu le jour en 2019. Il comprenait un bagage de près de 90 000 entrées provenant de GRID, un référentiel ouvert créé en 2015 par la société britannique *Digital Science* pour combler un vide dommageable à ses activités. ROR en a hérité une grande partie de son modèle de données (qu'il a néanmoins fait évoluer en 2022) et récupéré les alignements déjà effectués avec les référentiels généralistes que sont ISNI et Wikidata. Le référentiel s'est ensuite étoffé pour atteindre aujourd'hui près de 107 000 entrées. Les données sont placées sous licence *Creative Commons* CC0 1.0 (le plus proche qu'il soit possible du domaine public) et leur usage est libre et gratuit. Il en va de même pour l'API, gratuite et en *open source*.

COLLABORATION ET OUVERTURE COMME VERTUS CARDINALES

Concernant la gouvernance, ROR est conçu comme un service partagé, et non une entité autonome.

[1] ROR joue sur l'homophonie en anglais entre son nom et le verbe *to roar*, rugir (voir le blog https://ror.org/blog

[2] https://ror.org/search

[3] https://api.ror.org

[4] https://zenodo.org/records/8436953

[5] https://docs.google.com/forms/d/e/1FAIpQLSdJYaMTCwS7muuTa-B_CnAtCSkKzt19IkirAKG4u7umH9Nosg/viewform

[6] https://github.com/ror-community/ror-updates/projects/1

[7] https://ror.org/registry/#curation-advisory-board

[8] https://info.orcid.org/orcid-board

Ses trois fondateurs se sont engagés[9] à conserver un fonctionnement collaboratif, en s'interdisant de monnayer les données ou le service rendu, et de transférer tout ou partie du registre à une entité commerciale. ROR s'engage à souscrire aux POSI – acronyme des *Principles of Open Scholarly Infrastructures*[10] qui articulent gouvernance, pérennité et ouverture. Ceci étant posé, le modèle économique de ROR est toujours fragile car il reste conditionné au soutien volontaire de la communauté et à des subventions, souvent non reconductibles. On peut néanmoins noter par exemple que le Fonds national pour la science ouverte (FNSO) l'a choisi en 2023 comme l'un des quatre bénéficiaires dans le cadre de la campagne d'appel à financement[11], donnant le signe plutôt engageant d'une infrastructure qui commence à devenir incontournable.

Lors de la publication de la note sur les identifiants pérennes du COSO Europe en juin 2019[12], le terrain des organisations n'était pas considéré comme stabilisé. En ce jadis pourtant pas si lointain, GRID existait encore, RingGold était une base crédible (car utilisée par de nombreux éditeurs, bien que payante et fermée) et ISNI pouvait être présenté comme un choix officiel (par exemple par le Jisc, au Royaume-Uni[13]). Cinq ans plus tard, ROR a suffisamment consolidé ses positions pour avaler l'*Open Funder Registry* de CrossRef (référentiel des organismes de financement), devenir l'unique référentiel des organisations pour ORCID au détriment de RingGold[14], et s'imposer. Les implémentations se comptent par dizaines, qu'il s'agisse d'éditeurs, comme pour la prestigieuse revue Science de AAAS, d'archives ouvertes, comme celle de la NASA, voire de CRIS (*Current Research Information System*) comme celui du CERN, ou de bases de données bibliographiques comme OpenAlex.

Les trois fondateurs de ROR se sont engagés à conserver un fonctionnement collaboratif, en s'interdisant de monnayer les données ou le service rendu.

QUID DE LA FRANCE ?

ROR a toujours défendu l'idée de décrire uniquement les « *top-level institutions* ». Il faut entendre par là le refus d'entrer dans la description fine de la composition interne d'un établissement. Néanmoins, la France représente - même si elle n'est pas l'unique – le principal cas de figure où on ne peut pas se contenter de rattacher les productions de la recherche à une seule institution. Si tout le monde sait ce qu'est le CNRS, il est évident que dans un référentiel des structures de recherche digne de ce nom, il va falloir être un peu plus précis. L'échelon auquel la recherche française se fait, se pense, se finance et s'écrit, donc se signe, c'est avant tout le laboratoire. Or, qui dit laboratoire dit, la plupart du temps, multiplicité des tutelles : une ou plusieurs universités, et/ou un ou plusieurs organismes de recherche (ce fameux problème de la *mixité*). La relation d'un laboratoire avec ses tutelles fait qu'il n'est pas contenu ou subordonné ; et la nature de l'établissement peut changer (par exemple, se muer en un EPE) sans que le laboratoire en soit fondamentalement affecté. Nous touchons là à une des limites actuelles de ROR, eu égard au contexte français : le modèle de données ne permet pas de décrire de manière complexe les liens entre structures.

Sans signifier officiellement un changement de braquet, ROR a tout bonnement fait le choix pragmatique d'ajouter dans son registre une grande partie des laboratoires publics français à l'été 2023[15]. Les données issues du RNSR (Répertoire national des structures de recherche, coordonné par le ministère de l'Enseignement supérieur et de la Recherche) étant publiées de manière ouverte[16], après concertation avec le ministère, ROR en a intégré une partie pour enrichir sa base – et mécaniquement renforcer son attractivité pour le public concerné.

Au printemps 2023, l'Abes a mené un chantier d'alignement vers ROR pour tous les établissements habilités à délivrer le doctorat. Ce discret feulement appellera sans nul doute un mouvement plus ample dans les mois à venir. Le rugissement serait-il aussi communicatif que le bâillement ?

Carole Melzac
Service Autorités et référentiels de l'Abes
melzac@abes.fr

Crédit Adobe stock, par byrdyak

[9] https://ror.org/documents/ROR-Memorandum-of-Agreement-2022.pdf

[10] https://openscholarlyinfrastructure.org

[11] https://www.ouvrirlascience.fr/le-fonds-national-pour-la-science-ouverte-inscrit-son-soutien-aux-infrastructures-internationales-dans-la-duree

[12] https://hal-lara.archives-ouvertes.fr/hal-03640303v1

[13] https://researchdata.jiscinvolve.org/wp/2018/09/13/organisation-identifiers-recs

[14] https://info.orcid.org/orcid-support-for-ringgold-organization-ids-ending

[15] https://github.com/ror-community/ror-updates/issues/4930

[16] https://www.data.gouv.fr/en/datasets/repertoire-national-des-structures-de-recherche-rnsr

Stratégie nationale pour les données, algorithmes et codes sources : *un défi à relever collectivement*

Mobilisant tous les acteurs concernés, la feuille de route 2021-2024 des données, algorithmes et codes sources du ministère de l'Enseignement supérieur et de la Recherche a pour objectif de répondre à trois enjeux majeurs : innovation, transparence, simplification.

La structuration, la circulation et l'ouverture des données, initiées depuis plusieurs années, doivent désormais se généraliser à tous les types de données, qu'elles soient d'enseignement, de recherche, de gestion ou de pilotage, pour permettre un gain de valeur scientifique, économique et d'efficacité collective. Ce travail de fond mobilise tous les métiers de l'enseignement supérieur et de la recherche : administratifs, informaticiens, bibliothécaires et documentalistes, juristes, techniciens et ingénieurs en soutien à la recherche, enseignants-chercheurs et chercheurs… ainsi que les étudiants que nous devons former à ces enjeux. La feuille de route 2021-2024 des données, algorithmes et codes sources du ministère de l'Enseignement supérieur et de la Recherche (MESR) fixe ainsi un cadre de travail commun au ministère et à ses opérateurs pour répondre à trois enjeux majeurs :

- Favoriser **l'innovation** en permettant la réutilisation des données et codes sources
- Amplifier **la transparence,** pour davantage de confiance, de l'action publique
- **Simplifier** les processus de travail, notamment en allégeant la charge administrative (mise en œuvre du principe du « dîtes-le nous une fois »), et aider au pilotage à tous niveaux par une bonne circulation des données.

Cette feuille de route[1] comporte 53 actions qui seront, pour la quasi-totalité d'entre elles, achevées en 2024. Une nouvelle version verra le jour en 2024 et prolongera cette dynamique sur les années à venir.

LES BÉNÉFICIAIRES DE CETTE POLITIQUE

Au-delà des profils d'acteurs propres à l'enseignement supérieur et à la recherche, l'ouverture des données, des algorithmes et des codes constitue un vecteur de transparence de notre action et un levier de création de savoirs et de valeurs économique, démocratique, scientifique et politique.

En matière de transparence, on peut citer par exemple l'ouverture du code de Parcoursup. En termes de simplification, la mise en œuvre de dispositifs de circulation de données (via les API attestant du statut étudiant ou de boursier) permet aux étudiants de ne pas devoir justifier de leur statut à chacune de leur démarche administrative, et à l'administration de bénéficier d'une information certifiée à la source, évitant de nouvelles vérifications. Elle facilite la vie des étudiants dans le cadre du recours à de nombreux services au sein et en dehors de l'université (transports, culture, sport…).

Pour la recherche, RechercheDataGouv constitue aujourd'hui une offre souveraine de dépôt et de signalement des données de recherche, associée à une offre d'accompagnement des chercheurs producteurs de données. Le MESR soutient la conservation des codes sources au travers de l'archive universelle *Software Heritage* et coordonne les différents acteurs de manière à ce que ces deux dispositifs constituent, avec l'archive nationale HAL dédiée au dépôt et à la diffusion d'articles scientifiques, un écosystème reliant codes logiciels, données et publications de recherche.

Au travers de la promotion de l'adoption d'ORCID et des identifiants liés aux structures et aux résultats de recherche (publications, données, codes logiciels…), la feuille de route adresse aussi le sujet de l'allègement de la « charge administrative » des chercheurs. L'interconnexion entre les identifiants de personnes, structures et résultats pour la circulation de l'information déjà connue facilite la constitution de CV et de listes de résultats de recherche notamment dans le cadre de candidatures à des appels d'offres, de la participation à des missions d'expertise, de l'évaluation ou de l'accès aux services des infrastructures de recherche, du dépôt en archives ouvertes des publications, des données et codes sources… Ces PIDs (identifiants uniques et pérennes) constituent ainsi la clé de voûte d'une meilleure circulation des informations et réduisent les ressaisies pour les chercheurs.

Dans le domaine de la gestion, les unités mixtes de recherche sont, pour leur pilotage et lors de leur évaluation par l'Hcéres, confrontées à l'exercice complexe et fastidieux de devoir consolider des données issues des systèmes d'informations de leurs différentes tutelles.

Faciliter cette consolidation des données constitue d'ailleurs la recommandation prioritaire en matière de simplification identifiée par la mission Gillet[2]

[1] www.esr.gouv.fr/politique-donnee

[2] https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2023-06/rapport---mission-sur-l-cosyst-me-de-la-recherche-et-de-l-innovation-28193.pdf

[3] https://www.appelsprojetsrecherche.fr

(proposition 7 du rapport). Dans ce même esprit, la feuille de route identifie la nécessité de disposer d'un référentiel national univoque des données décrivant les structures de recherche.

LES IDENTIFIANTS PÉRENNES INTERNATIONAUX : CLÉ DE VOÛTE DE LA CIRCULATION ET DU PARTAGE DES DONNÉES

Dans le domaine de la recherche par exemple, il s'agit d'identifier de manière univoque les « entités clés » d'un système de recherche et d'établir les liens entre les chercheurs, leurs productions (publications, données, logiciels…), les unités de recherche, leurs tutelles, les sources de financement et personnels de l'unité.
Cette représentation s'entend aux niveaux national et international et elle doit être pérenne et tolérante aux transformations des organisations.
C'est pourquoi le ministère porte une stratégie d'emploi d'identifiants pérennes internationaux et souhaite définir et mettre en œuvre un plan d'actions favorisant leur adoption au niveau national.
Directement ou au travers de ses opérateurs, le ministère est déjà impliqué au niveau des gouvernances des instances internationales chargées d'attribuer ces identifiants que sont Crossref (publications), DataCite (données), ORCID (chercheurs) et est en lien étroit avec ROR (structures).

LES ACTIONS IMPORTANTES À MENER ET LE RÔLE DE L'ABES

Au niveau national, l'emploi de ces identifiants pérennes et internationaux est aujourd'hui naturel pour ce qui touche aux productions de la recherche. L'effort doit être porté sur l'identification des chercheurs et des structures de recherche.
Pour l'entité « chercheur », il s'agit d'agir à différents niveaux pour en faire *in fine* un véritable support de simplification et de visibilité :

- Faire en sorte que les différents systèmes qu'utilisent les chercheurs leur permettent d'employer ORCID pour faire transiter, sous leur contrôle, les informations pertinentes de leur profil : le portail des appels à projet[3] le permettra dès 2024 dans le cadre de la réponse à un appel
- Faciliter l'initialisation et l'entretien d'un profil ORCID, au travers de services capables de lier les productions à leurs auteurs, les chercheurs à leurs unités, les unités aux établissements impliqués
- Améliorer le dispositif ORCID lui-même, au niveau international pour qu'il accroisse encore sa couverture au niveau mondial et sa capacité de mise en qualité des informations.

Pour les structures de recherche, un référentiel français univoque aligné avec ROR émergera de façon à consolider l'information de référence et à la propager.

Barques de pêche amarrées au bord de la rivière

Crédit Adobe stock, par Abdul Momin

Un cadre de référence des structures définissant les types de structures (unités, fédérations, tutelles…) et la gouvernance des données associées sera défini en 2024.
Au vu de son expertise sur les identifiants, dont IdRef, et du rôle de coordonnateur du consortium ORCID France qu'elle joue déjà, l'Abes s'est vu confier récemment par le MESR un rôle opérationnel de garant de la qualité et de l'unicité des données de structures dans le respect du cadre de référence qui sera défini.

Hugues Ponchaut
Chargé de mission politique des données, des algorithmes et des codes sources au ministère de l'Enseignement supérieur et de la Recherche
hugues.ponchaut@recherche.gouv.fr

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté Égalité Fraternité

• • • LOI 3 DS : DU « DITES-LE NOUS UNE FOIS » À L'ADMINISTRATION PROACTIVE

La loi du 21 février 2022 dite loi 3 DS (différenciation, décentralisation, déconcentration et simplification) porte des dispositions visant à simplifier le fonctionnement des institutions locales et son article 162 vient renforcer le principe du « dites-le-nous une fois », tel qu'initialement prévu par l'article L.114-8 du code des relations du public avec l'administration[1] : le partage d'informations devient ici obligatoire. La loi autorise également les échanges qui permettent d'informer proactivement les usagers de leurs droits.

LES DÉCRETS :

- Le décret n° 2023-361 du 11 mai 2023[2] fixe les conditions d'application du nouveau principe d'échange d'informations entre administrations.
- Le décret n° 2023-362 du 11 mai 2023[3] détermine, pour chaque type d'informations ou de données, la liste des administrations chargées de les mettre à la disposition d'autres administrations.

[1] https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000045213315
[2] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000047541339
[3] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000047541361

INTERVIEW

« DANS L'ESR FRANÇAIS, C'EST L'ABES QUI, À TRAVERS IDREF, INCARNE LE LIEN ENTRE RÉFÉRENTIELS, DOCUMENTATION ET RECHERCHE »

Entretien avec David Reymond et Henri Bretel, membres du consortium CRISalid

Membres du consortium CRISalid, qui porte les projets « SoVisu+ » et « Idya », David Reymond et Henri Bretel[1] détaillent pour *Arabesques* les liens entre référentiels, documentation et système d'information recherche.

Pourriez-vous, tout d'abord, expliquer ce qu'on entend par « système d'information recherche », ses grands principes, fonctions et objectifs, et sa place dans un système d'information d'établissement de recherche ?

Un système d'information recherche ou SI recherche, ou encore *Current Research Information Systems* (CRIS), est défini par Guillaumet, García et Cuadrón comme un dispositif dédié à la gestion et au suivi des informations liées aux activités de recherche, offrant une variété de fonctionnalités qui s'adaptent aux besoins spécifiques des institutions de recherche.

Ce grand principe se décline en plusieurs fonctionnalités : créer et administrer des profils pour des chercheurs et des structures ; agréger les publications, brevets, thèses, en mesurer l'impact, identifier les tendances ; rassembler les informations sur les projets de recherche (ceci peut inclure les financements, collaborations, documents de suivi administratif) ; faciliter la collaboration entre chercheurs, par exemple par la gestion d'appels à projets ou la génération et la visualisation de réseaux à partir des profils ; enregistrer les accords de licence, embargos et restrictions d'accès, dépôts dans des archives ouvertes, protocoles, approbations éthiques, déclarations de conflit d'intérêts) ; et enfin, relier les informations concernant la recherche avec les autres systèmes institutionnels ou externes pour automatiser les flux de données et réduire la saisie manuelle : bases de données bibliographiques, archives ouvertes et entrepôts de données, systèmes de gestion des subventions etc.

En s'appuyant sur le web sémantique, l'outil que nous construisons permettra aux acteurs de l'écosystème, chercheurs, gestionnaires de la recherche, bibliothécaires, de constituer et valider un « graphe de connaissances institutionnel »[2] de leur institution.

Parmi les fonctionnalités évoquées, certains CRIS, comme le projet « SI Labo », priorisent la gestion administrative. Notre projet met davantage l'accent sur les acteurs et résultats de la recherche.

Le lien entre référentiels, documentation et SI Recherche n'est pas forcément immédiat. Pourriez-vous détailler les interactions qui les lient ?

Le lien est indirect. Les référentiels servent à s'assurer que les entités soient bien reconnues dans les différents contextes où elles sont utilisées : les structures, personnes, productions, concepts ont chacun un identifiant unique, issu d'un référentiel. L'absence d'ambiguïté bénéficie par exemple aux structures de recherche, dont la visibilité et la crédibilité dépendent d'une caractérisation exacte des compétences. Quant aux classements d'institutions, en plus de toutes les précautions qu'ils impliquent, ils nécessitent également des référentiels de production scientifique et des classifications disciplinaires partagées. Dans l'ESR français, c'est l'Abes qui, à travers IdRef, maintient et expose ces référentiels, incarnant le lien entre référentiels, documentation et recherche. Un chantier semblable existe à l'échelle européenne autour du référentiel CERIF.

Au regard de ces deux pôles, référentiel et SI recherche, et de la situation dans d'autres pays, quel est l'état d'avancement de l'ESR français et quelles sont ses spécificités ?

En France, la gestion locale est entravée par l'organisation complexe de la recherche, marquée par l'affiliation par « tutelle », par opposition au standard international de l'affiliation par employeur. Cette situation favorise la multiplication des « silos » de données. Elle nécessite des

outils résilients, ce qui peut devenir un atout, mais elle a aussi des inconvénients que nous voulons pallier. Tout d'abord, elle décourage l'enrichissement des données. SoVisu+ accompagnera la communauté pour améliorer la qualité de l'archivage de la production scientifique, en utilisant au besoin l'IA. Ensuite, elle réduit la lisibilité de la science pour la société. L'*Expert Finder System* permettra de mieux valoriser les forces des institutions. Enfin, elle est un obstacle au partage de données. BiblioLabs adapte aux spécificités françaises, EPE, UMR, la gestion des données bibliographiques et la curation des données.

Les référentiels sont donc des piliers fondateurs. Pouvez-vous exposer les enjeux que véhiculent les données et les exigences qu'elles portent en conséquence ?

Le premier enjeu des données est la qualité, dont dépend le service rendu à tous les acteurs de la donnée, chercheurs, services administratifs, documentalistes. L'enjeu de l'interopérabilité est directement lié aux référentiels. Les données visibles à l'international doivent être normalisées selon les formats CERIF ou VIVO. Mais l'identification des données françaises passe par l'utilisation des référentiels correspondants en France. C'est pourquoi nous travaillons avec les services de l'Abes, mais aussi les référentiels d'AuréHAL, que l'usage de HAL rend incontournables. Enfin la certification par l'instance d'autorité concernée, l'Abes pour les alignements d'identifiants ou une direction de la recherche pour le lien entre unité et tutelles par exemple, est nécessaire pour valider la correspondance entre la réalité et la donnée normalisée.

Pourriez-vous nommer les acteurs, agents et institutions en charge de ces données, expliquer leurs rôles, et la manière dont ils s'articulent ?

Les directions administratives veillent à la mise à jour des référentiels de structures internes, qui ont vocation à être répercutés à l'échelle nationale, par exemple dans le RNSR. Les chercheurs tiennent à jour les données concernant leurs travaux, avec le trio ROR, ORCID, DOI pour l'international, et les informations spécifiques à leur champ. Les documentalistes maintiennent les référentiels bibliographiques et accompagnent la normalisation des données. La création des données est le fait de chaque agent, mais l'enrichissement et la certification sont le fait d'interactions. Le chercheur est seul à même de définir ses champs d'expertise, mais la normalisation passe par l'interaction chercheur-documentaliste ou chercheur-autorité administrative. Les institutions nationales ou internationales interagissent entre elles pour assurer la compatibilité des référentiels.

Telle une course de relais, chaque acteur doit porter un bout de la donnée, la qualifier et la passer enrichie au coureur suivant. Sans froisser ou distribuer de bons points, quelles sont, actuellement, les articulations les plus sûres et celles qui sont « à risque » ?

L'architecture décentralisée s'écarte de l'image de la course de relais car chaque agent du circuit est en même temps producteur et utilisateur de la donnée, interagissant potentiellement avec tous les autres. Dans ce réseau, les maillons d'autorité et de certifications sont critiques car ils ont une responsabilité sur la production de la donnée pour tous les autres acteurs.

Vous formez, autour de SoVisu+, un collectif d'institutions engagées dans le développement d'un outil communautaire apte à doter les établissements français de recherche d'un SI recherche. Pourriez-vous exposer votre démarche, son point de départ et sa cible, et en dresser les clés du succès ?

La communauté s'est réunie autour de trois projets initiaux, SoVisu de l'université de Toulon, l'*Expert Finding System* de l'université Paris 1 Panthéon-Sorbonne, et BiblioLabs de l'université Paris-Saclay, avec pour objectif d'urbaniser et industrialiser ces éléments sous forme de modules logiciels libres et techniquement à jour. Elle demeure ouverte aux nouvelles participations. Elle est dirigée par un comité de pilotage et répartit le travail dans des groupes d'expertises et de conception. L'objectif est de développer des briques applicatives indépendantes, conformes à l'état de l'art technologique, favorisant la durabilité, l'écoresponsabilité et la réduction des coûts. Le projet s'inscrit dans les préconisations de rapports internationaux qui encouragent l'investissement dans la gestion institutionnelle des données, le soutien des identifiants pérennes, la collaboration entre fonctions, l'investissement dans du personnel dédié, et l'intégration des informations de recherche dans la gouvernance. La démarche implique de définir une architecture décentralisée, de développer les services de manière agile, d'arrêter les choix de formats et de technologies. La communauté communique vers l'extérieur (FU, Esup Days, webinaire du GT Recherche ESR), et consulte des experts nationaux ou internationaux dont les communautés SemApps et VIVO. Le déploiement des installations pilotes est envisagé pour la rentrée prochaine. Les travaux visent à homogénéiser les attentes, à anticiper les impacts sur les activités existantes, et à identifier les points techniques critiques et les besoins de formation. Des moyens seront sollicités pour orchestrer et animer la communauté, documenter les réalisations, et développer ou intégrer des modules répondant aux exigences de qualité et d'interopérabilité.

Propos recueillis par
François Mistral et Véronique Heurtematte

[1] David Reymond est maître de conférences à l'université de Toulon, et Henri Bretel, chargé de bibliométrie à l'université Paris-Saclay.

[2] Les graphes de connaissances représentent des concepts (des personnes, des lieux, des événements) et leurs relations sémantiques. Ce sont des structures de données utiles à des fins d'exploration, de cartographie et de visualisation. Les données ouvertes concernant la recherche forment désormais un réseau mondial de connaissances interconnecté et décentralisé, qui peut être modifié et enrichi par la communauté.

Rameau et l'automate : *que vaut l'indexation générée par une intelligence artificielle ?*

Les expérimentations menées par le Labo de l'Abes révèlent le potentiel de l'Intelligence artificielle comme outil d'indexation. Prochaine étape : tester en situation réelle.

D'un point de vue informatique, l'indexation automatique avec un vocabulaire tel que Rameau est une tâche complexe et ambitieuse. En termes techniques, on parlera d'une classification *multilabel* extrême. En effet, il s'agit bien de *classer,* c'est-à-dire de ranger les documents dans des cases prédéfinies, et non pas de regrouper les documents semblables (ce qu'on appelle « *clustering* »). En second lieu, cette classification est dite « multilabel » car un même document peut appartenir à différentes classes, c'est-à-dire être indexé par plusieurs concepts Rameau à la fois. Enfin, cette classification multilabel est dite *extrême* car le nombre de classes est très important, en l'occurrence autour de 100 000, ce qui complique considérablement l'affaire. Les caractéristiques de cette tâche en font un réel défi pour une machine, mais aussi pour les humains. Cette symétrie inhabituelle est un point important, que nous retrouverons plus loin.

ENTRAÎNER

Selon l'approche *machine learning,* nous essayons d'apprendre à la machine à indexer avec Rameau en lui soumettant de nombreux exemples. En l'occurrence, nous avons extrait du Sudoc un corpus d'environ 150 000 notices d'*ebooks* ayant une indexation Rameau et un résumé en français. Nous avons entraîné ces données avec deux techniques : le logiciel ANNIF (qui intègre différents algorithmes) et le calcul d'*embeddings* (vectorisation sémantique du texte). Avant de lancer le programme d'entraînement, nous avons mis de côté 30 % des notices extraites pour en faire un corpus de test. Après entraînement, nous demandons au programme de proposer une indexation Rameau pour ces notices, puis nous comparons ces propositions à l'indexation réelle (celle du Sudoc), pour savoir si la machine a « bien » travaillé ou pas. Toute la question est de savoir ce que signifie « bien » pour ce type de tâche.

ÉVALUER

En principe, le travail de la machine sera parfait si elle parvient à indexer ces notices de test exactement comme les catalogueurs du Sudoc :

- Tous les concepts du Sudoc ont été proposés par la machine (rappel = 1)
- Tous les concepts proposés par la machine sont présents dans les notices Sudoc (précision = 1).

Si cette approche de l'évaluation est pertinente pour des tâches de classification simple (binaire ou multiclasse), elle semble inadéquate pour notre classification multilabel, pour les raisons suivantes :

1. **Une indexation peut être en partie correcte.**
2. **Il n'y a pas qu'une seule manière de bien indexer une notice.**

1. Une indexation peut être en partie correcte

Si la machine propose un tiers des concepts Sudoc (rappel = 0,33) et qu'un tiers de ces propositions sont dans le Sudoc (précision = 0,33), on est loin d'un échec complet.
En l'occurrence, voici les scores obtenus avec l'un des algorithmes proposés par le logiciel ANNIF :

- Rappel = 0,25 (25 % des concepts Sudoc sont trouvés)
- Précision@5 = 0,35 (35 % des 5 premiers concepts proposés sont dans le Sudoc)

À ce stade, comme on le verra, il serait imprudent de conclure que ce résultat est décevant ou satisfaisant.

2. Il n'y a pas qu'une seule manière de bien indexer une notice

Qu'en est-il des concepts proposés par la machine et absents du Sudoc ? Sont-ils pour autant incorrects ? Il se peut que ces propositions originales couvrent une notion que l'indexation Sudoc couvre au moyen d'un autre concept Rameau proche, voire que la proposition machine exprime une notion négligée par l'indexation Rameau (incomplète dans ce cas). Bref, que l'indexation automatique soit meilleure que le laisse présager les chiffres ci-dessus.
Cette hypothèse est d'autant plus pertinente qu'on sait bien qu'il n'existe pas une seule manière d'indexer correctement un document. C'est vrai pour les humains. Il faut donc tenir compte de cette pluralité pour évaluer la machine.
Partant de cette intuition bien connue et documentée, nous avons décidé de « réindexer » une centaine de notices Sudoc. Six bibliothécaires de l'Abes, aux profils variés, ont été invités à proposer au maximum trois concepts (ou chaînes) Rameau, en ne disposant que du titre et du résumé.
À l'issue de ce travail d'autant plus fastidieux qu'il était contraint par les artifices inhérents à tout dispositif expérimental, nous disposions d'une pluralité d'indexations humaines et machine pour une centaine de notices permettant d'effectuer différentes comparaisons, plus riches et réalistes que la seule comparaison Sudoc/machine, et notamment de comparer chaque réindexation avec celle du Sudoc, avec l'algorithme machine, ou de comparer les réindexations entre elles.

Le schéma illustre certaines de ces comparaisons :

Même si la pluralité des méthodes de mesure complique la lecture, on peut voir que, pour les humains comme pour la machine, ni la précision ni le rappel ne dépassent 0.6. D'autre part, on constate que l'algorithme Omikuji d'ANNIF fait en général mieux que le réindexeur Lolo (le prénom a été modifié).

Compte tenu de la divergence entre les indexations produites par des bibliothécaires (Sudoc compris) à propos du même document, l'évaluation des propositions de la machine ne peut se limiter à mesurer la distance *absolue* entre celles-ci et les indexations Sudoc. Il est plus judicieux de se demander si la machine est beaucoup plus loin du Sudoc que les indexations humaines. Mais dans ce cas, pourquoi prendre le Sudoc comme point de référence ? Il est aussi légitime de privilégier n'importe quelle indexation humaine, voire une agrégation des indexations humaines.

Union ? Intersection ?

Le tableau ci-contre permet de comparer les propositions de deux algorithmes à l'union et l'intersection des réindexations pour l'ouvrage *Éthologie animale et humaine : communication et comportement* de Jacques Goldberg[1].

SUDOC + UNION DES 6 RÉINDEXEURS	INTERSECTION DES 6 RÉINDEXEURS	ANNIF	EMBEDDINGS V2
Sudoc	Éthologie comparée (4 fois)	Éthologie	***Comportement animal***
Comportement animal		Comportement animal	Manuels d'enseignement supérieur
Éthologie comparée	Comportement humain (2)	Relations homme-animal	Éthologie – Manuels d'enseignement supérieur
Éthologie humaine		Comportement humain	Neuroéthologie
Union des réindexations	Éthologie humaine (2)	Communication	
Éthologie comparée		Communication interpersonnelle	***Éthologie humaine***
Éthologie	Comportement animal (2)	Communication non verbale	Éthologie
Comportement animal		Éthologie humaine	Comportement social des animaux
Comportement humain	***Subdivisions***	Communication dans les organisations	
Éthologie humaine	Modèles animaux	Animaux	
Comportement humain Modèles animaux			
Communication Comportement social des animaux			
Sciences du comportement			

D'une manière générale, les indexations propres à la machine sont sans doute moins pertinentes que celles propres à tel ou tel bibliothécaire. Sans pour autant être systématiquement incorrectes ou incongrues. Comment le savoir ?

NOTER LES INDEXATIONS

Nous avons donc choisi d'ajouter une strate d'évaluation consistant à porter un jugement qualitatif sur les indexations proposées, celles du Sudoc, celles des « réindexeurs » et celles de la machine (plusieurs algorithmes). Pour autant, cette notation n'est pas arbitraire : elle suit un barème que nous avons défini pour noter à la fois la qualité de chaque proposition de concept et la qualité du groupe de concepts choisi par tel indexeur pour tel document.

Selon notre grille,

une proposition de concept est :

- exacte ou non (0 ou 1)
- plus ou moins précise (0 ou 1 ou 2)

Un groupe de concepts est :

- plus ou moins complet (0 ou 1 ou 2)
- redondant ou non (0 ou 1)

Le notateur s'appuie sur la lecture du titre et du résumé (comme la machine) pour donner une note sur chaque composante du barème. Il reste ensuite à effectuer la moyenne sur chaque composante, puis la moyenne globale.

	COMPLÉTUDE (0 OU 1 OU 2)	REDONDANCE (0 OU 1)
Sudoc	1,5	0,98
Union des réindexeurs	1,6	0,98
ANNIF (omikuji)	1,5	0,26
Embeddings v1	1,3	0,40

	EXACTITUDE (0 OU 1)	PRÉCISION (0 OU 1 OU 2)
Sudoc	0,99	1,8
Union des réindexeurs	0,98	1,9
ANNIF (omikuji)	0,60	1,5
ANNIF (omikuji). 2 premières propositions	0,86	1,6
Embeddings v1	0,66	1,6
Embeddings v1. 2 premières propositions	0,79	1,6

Même si des marges et des pistes d'amélioration existent, ces résultats nous semblent suffisamment bons pour envisager d'aller plus loin, à savoir expérimenter en situation, en intégrant un service de proposition d'indexation Rameau dans l'environnement de catalogage du Sudoc en tant qu'aide à la décision. Les modalités de cette expérimentation sont encore à préciser.

YANN NICOLAS
Responsable du Labo de l'Abes
nicolas@abes.fr

[1] https://www.sudoc.fr/147294509

Dans le monde entier, la transition bibliographique fait évoluer les règles de description des documents auxquels les bibliothèques donnent accès mais aussi les données et leur circulation. Une transformation en profondeur à laquelle le référentiel Rameau n'échappe pas.

Réforme Rameau : *vers de nouveaux référentiels pour l'indexation sujet*

Élaborer des données bibliographiques réellement partagées, au-delà de leur rattachement institutionnel, et accessibles de façon équitable dans le monde entier : c'est le rêve qui se trouve derrière des réalisations conceptuelles comme les modèles IFLA LRM ou BIBFRAME, actuellement en cours d'implémentation dans des systèmes de nouvelle génération. En mettant en valeur les données les plus utiles parmi celles que produisent et gèrent les bibliothèques, ces nouveaux SGB et réservoirs de données ouvrent des horizons en matière de partage de données entre bibliothèques, et de diffusion des connaissances sur le web.

Au lieu d'être constitués de données en MARC structurées autour de la notion de notice, ces systèmes ont pour caractéristiques d'être des bases d'entités et de relations, intégrant de nombreux liens internes et externes. Autrement dit, pour décrire une ressource documentaire dans ce type de système, on puisera dans des référentiels d'œuvres, d'expressions, qui auront potentiellement été produits par d'autres bibliothèques ou réseaux. On ne décrira que ce qui n'aura pas déjà été décrit dans notre système ou dans les référentiels liés. Ces entités bibliographiques (Œuvre, Expression, Manifestation, Item) seront en relation avec des descriptions de personnes, collectivités, concepts, lieux, ou encore avec des repères chronologiques. Une grande partie des attributs permettant de décrire ces entités seront, plus qu'aujourd'hui encore, issus de référentiels (par exemple les catégories d'œuvre, les types de support…), de même que les relations permettant de les associer (relations d'adaptation entre œuvres, de création entre une œuvre et une personne…).

Le recours à ces données et à leurs identifiants au sein de référentiels deviendra donc le principe de base de notre gestion de l'information bibliographique, celui qui permettra à nos données d'être pleinement utiles et utilisées par d'autres bibliothèques et sur le web, y compris dans des technologies non spécifiques à notre domaine. La réforme Rameau s'inscrit dans ce mouvement d'ouverture vers le web de données.

Né il y a 40 ans avec les débuts de l'informatisation et l'ambition d'établir des catalogues collectifs, Rameau est devenu aujourd'hui le langage d'indexation le plus largement utilisé par les bibliothèques françaises.

QU'EST-CE QUE LA RÉFORME RAMEAU ?

Il y a 40 ans, avec les débuts de l'informatisation des bibliothèques et l'ambition d'établir des catalogues collectifs, est né le Répertoire d'autorité-matière encyclopédique et alphabétique unifié, Rameau, devenu aujourd'hui le langage d'indexation le plus largement utilisé par les bibliothèques françaises. Rameau répondait alors à des contraintes techniques et usages de l'époque (recherche par index tout particulièrement), aujourd'hui révolus. Jugé trop complexe et inextricable dans sa syntaxe, assigné à résidence dans les seuls catalogues des bibliothèques à l'heure où le principe de LinkedData s'impose sur le web, Rameau devait donc être transformé et modernisé, pour être simplifié, plus visible et mieux exploitable[1].

Un rapport d'avril 2017 présenté au Comité opérationnel Rameau a défini les axes forts de la réforme[2] :

- faire évoluer Rameau pour le rendre compatible avec l'ensemble des évolutions actuelles
- rendre Rameau plus simple et plus intuitif, pour la recherche d'information comme pour le catalogage
- dépasser la syntaxe pour libérer la richesse sémantique et terminologique de Rameau.

L'objectif principal est de recentrer Rameau autour de l'entité Concept, et de transformer en entités dans des référentiels séparés les autorités propres au Genre et à la Forme de la ressource ainsi que les subdivisions de Lieu et Temps initialement intégrées au vocabulaire contrôlé.

Le Centre national Rameau (CNR), instance opérationnelle rattachée à la Bibliothèque nationale de France qui assure l'évolution du référentiel, a choisi pour ce faire de procéder par étapes successives.

COMMENT APPLIQUE-T-ON LA RÉFORME RAMEAU À L'ABES ?

Entre 2019 et 2021, les premières grandes étapes de la réforme mises en œuvre par le CNR ont été répercutées dans le Sudoc, accompagnées d'actions de communication et de formation, pour les bibliothécaires et les administrateurs de SGB.

[1] Michel Mingam, responsable du centre Rameau entre 2004 et 2015, avait déjà fait ce constat dans son article « Rameau : bilan, perspectives », in *Bulletin des Bibliothèques de France*, 2005, 5, http://bbf.enssib.fr/consulter/bbf-2005-05-0038-001)

[2] Rapport final du Groupe de travail national sur la syntaxe de Rameau : https://rameau.bnf.fr/sites/default/files/chantier_syntaxe/pdf/rapport_final_syntaxe_rameau.pdf

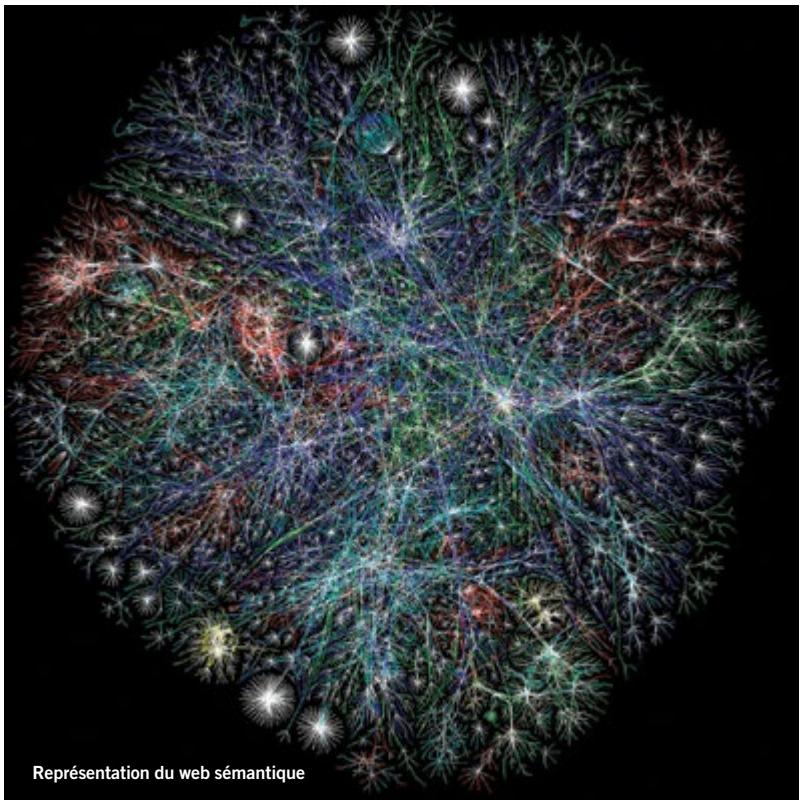
1 - Vers un ordre unique « concept – lieu – temps »
Axe majeur de la réforme : la simplification de sa syntaxe, jugée bien trop limitative et complexe, via l'instauration d'un ordre unique de construction de chaîne d'indexation « concept - lieu – temps ». Il acte la disparition des subdivisions par domaine et des règles spécifiques d'emploi d'un certain nombre d'autorités. 72 500 autorités ont été ainsi transformées à cette occasion et plus de 700 0000 notices bibliographiques modifiées dans le Sudoc.

2 - Création du référentiel Genre/Forme
Autre modification d'importance en 2020, dans une logique de modélisation entités-relations conforme au modèle LRM : la création d'un nouveau référentiel Genre/Forme.
Après création des nouvelles zones UNIMARC, quelque 17 000 autorités anciennement considérées comme des concepts ont été identifiées par le CNR et sont devenues des entités Genre/Forme. S'en est suivi le début des chantiers rétrospectifs dans les notices bibliographiques du Sudoc utilisant ces autorités (3 millions concernées, entre autres celles utilisant « Thèses et écrits académiques », « Actes de congrès »…).

3 - Vers un ensemble unique et cohérent de descripteurs
En complément, deux vagues d'opérations de « lissage terminologique » ont été menées par le CNR. Il s'agit principalement de créations, suppressions, scissions et fusions d'autorités, mais aussi de désambiguïsation des concepts et modifications de type d'autorités. Par exemple, de nombreuses autorités préconstruites (rassemblant plusieurs termes) ont été scindées, les autorités singulier/pluriel fusionnées, la fameuse et très usitée « Histoire et Critique » supprimée… donnant lieu à près de 1 000 autorités concernées pour une reprise de données dans plus de 500 000 notices bibliographiques du Sudoc.
Aujourd'hui les opérations de lissage terminologique se poursuivent au CNR au fil de l'eau et sont, dans la mesure du possible, répercutées dans le Sudoc. L'Abes poursuit de son côté les chantiers de correction rétrospective, loin d'être achevés. Les établissements membres du réseau récupèrent les notices bibliographiques modifiées au gré de leurs besoins.

VERS UN NOUVEL ÉCOSYSTÈME

Rameau restera un vocabulaire encyclopédique, relativement générique. Une réflexion reste à mener sur son articulation avec d'autres référentiels Sujet spécialisés ou des référentiels internationaux. À noter toutefois que les standards internationaux tels que RDA ne décrivent pas finement aujourd'hui l'entité Concept ni la subtilité des relations possibles entre différentes instances (combinaison, juxtaposition, influence…). Le CNR y travaille actuellement avec un groupe de normalisation RDA-FR.


crédit photo flickr / Patrick Barry
Représentation du web sémantique

En termes d'usage, selon IFLA LRM, la relation de sujet qui correspond au concept actuel d'indexation sera portée par l'entité Œuvre, de même que la relation vers les entités Genre/Forme. Les référentiels Lieu et Temps pourront quant à eux être utilisés en lien avec toutes les entités du modèle : date de création d'une œuvre, période ou lieu comme sujet d'une œuvre, date et lieu de publication d'une manifestation, d'organisation d'un congrès, ou encore lieu où se trouve une collectivité ou année de naissance d'une personne. Les possibilités de liens riches et utiles sont très nombreuses.
Pour autant, l'Abes ne pourra exploiter le plein potentiel de ces nouveaux référentiels, dont Rameau, que lorsqu'elle se sera dotée d'un système de nouvelle génération basé sur un modèle entités/relations. Le Projet d'établissement 2024-2028 de l'Abes devrait offrir l'opportunité d'aller vers ces nouvelles modalités de gestion de l'information bibliographique.

Aurélie Faivre
Gestionnaire métadonnées – transition bibliographique à l'Abes
faivre@abes.fr

Héloïse Lecomte
Chargée de mission coordination des données
lecomte@abes.fr

BnF RAMEAU

Alliant précision scientifique et démarche collaborative, la base de données géographiques mondiale GeoNames s'est imposée comme une ressource indispensable pour les chercheurs comme pour les bibliothécaires.

GeoNames : *pierre angulaire des données de référence géographiques*

Geonames.org[1] est un répertoire mondial de données géographiques compilant 12 millions de caractéristiques et 25 millions de noms. Régi par une licence d'attribution ouverte, ce dépôt non seulement facilite le téléchargement et l'édition des données, mais contribue également à la création collaborative de connaissances à l'échelle mondiale. Geonames.org se distingue par le fait qu'il s'appuie sur des sources officielles, notamment les agences nationales de cartographie. Par sa rigueur et son exactitude scientifiques, ses solides mesures pour garantir la qualité des données, et la description complète des champs, GeoNames constitue un outil inestimable et une ressource essentielle pour les chercheurs et les bibliothécaires à la recherche de données fiables.

Attributs des entités du répertoire : au cœur de Geonames.org se trouve une compilation minutieuse d'attributs. Ceux-ci comprennent des éléments fondamentaux tels que le nom, la classe d'entités, le code d'entités, la latitude, la longitude, le pays, la hiérarchie administrative (région, département, arrondissement, commune), la population, l'altitude et le fuseau horaire. La représentation de l'empreinte géographique sous la forme d'un point, qui désigne souvent l'église ou le siège du gouvernement pour les lieux peuplés, ajoute de la précision à l'information.

Modèle d'attribution : Geonames.org adopte un modèle basé sur l'attribution, s'éloignant de l'approche strictement collaborative. Ce modèle permet une utilisation plus libre des données, facilitant une intégration transparente avec les ensembles de données des utilisateurs. L'accent mis sur l'attribution encourage un environnement collaboratif dans lequel le référentiel devient un espace partagé pour l'amélioration des connaissances.

Crédit Adobe stock, par Keitma

Identifiant végétal, généré à l'aide de l'IA

Sources officielles et collaboration : dans un souci d'exactitude, Geonames.org s'appuie largement sur des sources officielles, notamment les agences nationales de cartographie et les bureaux de statistiques. La disponibilité croissante des données officielles sous licence ouverte ouvre la voie à un processus d'intégration continue, s'alignant sur l'objectif primordial de la plateforme qui consiste à utiliser des sources faisant autorité. Alors que les États-Unis ont lancé le mouvement de diffusion des données par les sources gouvernementales sous licences libres, la plupart des autres pays suivent et diffusent leurs données. La situation est plus difficile dans les pays en développement où des organisations d'aide et de secours collectent et publient des données qui sont souvent absentes des sources gouvernementales officielles.

Identifiants pérennes et variantes de noms : Geonames.org utilise un identifiant pérenne, GeonameId, qui assure la stabilité des références. Le nom principal est présenté en anglais ou dans des variantes internationales, accompagné de noms alternatifs liés. Cette caractéristique enrichit l'ensemble des données, en tenant compte des diverses particularités linguistiques et de la nomenclature historique.

Prise en charge des noms alternatifs : le répertoire adopte une approche flexible en prenant en charge les noms alternatifs. Ces noms alternatifs sont assortis d'attributs tels que le code de langue ISO, le statut de préférence, les variations de longueur et la signification historique ou vernaculaire. Pour les noms historiques, l'inclusion des champs « depuis » et « jusqu'à » permet de suivre l'évolution temporelle de la nomenclature. Pour l'affichage aux utilisateurs finaux, il est recommandé d'utiliser le nom alternatif marqué comme *isShort* dans la langue concernée. Le champ du code de langue pour les noms alternatifs accueille également des pseudo-codes comme *link* pour une url vers un site web (le plus souvent l'article Wikipedia correspondant). D'autres pseudo-codes sont utilisés, comme *post* pour les codes postaux, *icao* et *iata* pour les codes d'aéroport.

[1] https://www.geonames.org

Garantie de qualité et processus de confirmation : Geonames.org accorde une grande importance au maintien de l'intégrité des données en les soumettant à un processus de garantie qualité approfondi, comprenant plus de 1 700 vérifications. Ces vérifications systématiques permettent d'identifier rapidement toute incohérence. En outre, les modifications pertinentes sont soumises à un processus de confirmation.

Modèle premium et mises à jour en temps voulu : la plateforme comprend un modèle premium proposant des mises à jour par téléchargements quotidiens gratuits. Les abonnés Premium bénéficient d'un abonnement mensuel leur donnant accès à un ensemble de données améliorées soumises aux processus de garantie de qualité, alors que les extractions quotidiennes gratuites sont des « *work in progress* ». Cette approche à deux niveaux permet de trouver un équilibre entre l'accessibilité et la fiabilité des données.

Divisions administratives hiérarchiques : Geonames.org fournit une structure hiérarchique pour les divisions administratives, enrichissant automatiquement les entrées des utilisateurs avec les relations administratives des frontières. Cette structure s'étend aux champs relatifs aux pays, y compris un attribut spécifique (cc2) pour les pays alternatifs, ce qui est particulièrement pertinent pour les éléments tels que les montagnes ou les lacs bordant plusieurs pays. Les frontières des pays et des divisions administratives jouent un rôle essentiel dans le processus d'enrichissement en améliorant la compréhension contextuelle des caractéristiques géographiques.

Relations « parents-enfants » et caractéristiques créées par les utilisateurs : une caractéristique distinctive de Geonames.org est la définition d'une relation « parent-enfant » entre les caractéristiques, établissant une hiérarchie secondaire. Pour les entités créées par les utilisateurs, la plate-forme complète le fuseau horaire et l'élévation approximative à partir des modèles DEM (*Digital Elevation Models,* ou coordonnées en trois dimensions) garantissant ainsi l'exhaustivité, même en l'absence de données d'élévation explicites.

Par essence, Geonames.org apparaît non seulement comme un dépôt de données géographiques mais aussi comme un outil polyvalent, méticuleusement conçu pour répondre aux besoins des bibliothécaires. Grâce à un mélange judicieux de précision scientifique, de collaboration et de fonctionnalités centrées sur l'utilisateur, il s'impose comme une ressource indispensable dans le paysage dynamique des études géographiques.

MARC WICK
Fondateur de GeoNames
marc@geonames.org

• • • PLEIADES : UNE BASE DE DONNÉES GÉOGRAPHIQUES POUR LES MONDES ANCIENS[1]

Développée dans le cadre d'un projet conjoint de l'Institute for the Study of the Ancient World de l'Université de New York et du Ancient World Mapping Center de l'Université de Caroline du Nord (Institut pour l'étude du monde antique et Centre de cartographie du monde antique), Pleiades est, depuis sa mise en ligne en 2007, la plus importante base de données géographiques des mondes anciens.

SES FONCTIONNALITÉS

Pleiades opère une distinction entre les lieux, les toponymes et les localisations. Le lieu est l'unité de base correspondant à une entrée dans la base de données et peut posséder plusieurs noms et localisations.

La base de données est riche de plus de 40 000 lieux, 37 000 noms et 43 000 localisations. Chaque lieu est décrit dans une page de présentation possédant un URI et indiquant sa localisation sous la forme de degrés décimaux, les noms associés dans différentes langues et à différentes époques, les autres lieux en relation – c'est-à-dire incluant ou faisant partie du lieu décrit – des mots-clés, des références bibliographiques distinguant les sources anciennes et contemporaines, des liens vers d'autres bases de données comme GeoNames, des marqueurs permettant de lier des photographies déposées sur Flickr et des liens de téléchargement. Ces présentations font l'objet d'une évaluation par les pairs.

Pleiades permet de télécharger les données relatives à un lieu aux formats Atom, JSON, KML, RDF+XML et Turtle depuis sa page de présentation ou l'ensemble des données aux formats CSV, JSON, KML et RDF depuis la page *Downloads*.

SES LIMITES

Produit par la collaboration de différents projets œuvrant notamment dans le domaine des études classiques, Pleiades donne accès à de nombreuses informations géographiques, mais principalement concentrées sur le monde méditerranéen gréco-romain. Néanmoins, le projet est toujours en cours et permet à chacun de participer à l'enrichissement des données y compris dans un périmètre géographique plus large.

NICOLAS SOUCHON[2]
Assistant-égyptologue à l'Institut français d'archéologie orientale
nsouchon@ifao.egnet.net

[1] https://pleiades.stoa.org
[2] EPHE, PSL, AOROC UMR 8546, Paris/IFAO, Le Caire

IDREF, BRIQUE ESSENTIELLE DANS L'ÉCOSYSTÈME DE L'ESR

Quatre témoignages, quatre cas d'usage pour illustrer le potentiel du référentiel IdRef dans la construction d'outils de référencement au service de la valorisation de la recherche.

FRANTIQ : UN PARTENARIAT AUTOUR DES RÉFÉRENTIELS DES DISCIPLINES ARCHÉOLOGIQUES

Née en 1984 à l'initiative d'archéologues spécialistes de l'Antiquité, la Fédération et Ressources sur l'Antiquité (Frantiq)[1] du CNRS rassemble une quarantaine de centres documentaires aux statuts variés (unités de recherche du CNRS, services du ministère de la Culture, collectivités territoriales, musées nationaux...).

Depuis les années 1980, Frantiq a travaillé à l'amélioration de l'accès centralisé aux ressources documentaires, au travers du catalogue collectif indexé (CCI)[2] en catalogage partagé sous Koha (plus d'un million de références). En 2021, Frantiq a signé une convention avec l'Abes en faveur de l'ouverture d'IdRef aux producteurs d'autorités et a développé un *plug-in* IdRef pour Koha.

Ce changement a été opéré de 2022 à 2023 (suivi du projet, formation des correspondants auteurs puis des catalogueurs). Les catalogueurs Frantiq peuvent désormais utiliser les autorités IdRef existantes mais également en produire, enrichissant ce référentiel de données issues de différentes institutions archéologiques. Frantiq n'a eu de cesse d'ouvrir son catalogue et de le « connecter » à différents référentiels utilisés comme des briques techniques afin de faire rebondir les données entre elles à partir des données d'autorité mais aussi des données sujets au travers du thésaurus Pactols[3], riche de 12 000 concepts sujets et 50 000 termes lieux, publié sur le web via le logiciel open source OpenTheso.

Une meilleure communication entre les corpus scientifiques, les catalogues et les autres référentiels des disciplines archéologiques est en effet l'un des objectifs prioritaires du réseau Frantiq, qui veille à l'interopérabilité technique et sémantique des outils qu'il déploie (catalogue, développement et alignement du vocabulaire du thesaurus Pactols).

Véronique Humbert
Directrice de Frantiq
veronique.humbert@yahoo.fr

[1] https://www.frantiq.fr
[2] https://catalogue.frantiq.fr
[3] https://pactols.frantiq.fr/opentheso

ELSEVIER : IDREF, UN RÉFÉRENTIEL NATIONAL PIVOT POUR LES ÉDITEURS

De nombreuses initiatives ont vu le jour pour identifier les acteurs de la recherche et accéder à leurs travaux grâce à des référentiels d'identifiants permanents. Toutefois, la coexistence des différents systèmes d'autorités des personnes et des structures ne facilite pas l'entreprise. Aussi, le rôle d'IdRef, par sa légitimité nationale, s'avère essentiel comme référentiel pivot afin d'optimiser la qualité des données de chaque acteur et permettre également l'alignement des multiples identifiants permanents qui coexistent.

C'est dans cette perspective qu'un partenariat a été initié avec Scopus pour la livraison de métadonnées relatives aux auteurs français depuis 2011 pour contribuer tout d'abord à l'optimisation des travaux de désambiguïsation menés par l'Abes sur les profils d'auteurs. Pour mémoire, l'algorithme Scopus enrichit quotidiennement les profils d'auteurs et d'affiliations, qui font l'objet d'une curation permanente. Ce sont ainsi plus de 3 749 profils d'affiliations qui ont été créés pour refléter l'articulation des structures de recherche française.

Les bénéfices attendus sont les suivants :

- Une amélioration de la qualité des métadonnées avec une plus grande précision de description, des données plus riches, plus fiables, à jour et interopérables
- Une économie de temps en automatisant encore plus les opérations de dédoublonnage et de désambiguïsation
- Une interopérabilité accrue des données de toutes natures, telles que les données de financements, données de la recherche, *preprints*, publications, thèses, brevets, documents de politique publique, et tout autre objet susceptible d'enrichir la compréhension des activités des chercheurs et de leur impact dans toutes ses dimensions

Nous sommes ainsi convaincus que la convergence des expertises techniques développées par les différents acteurs est un facteur décisif pour faciliter l'exploration et la réutilisation des travaux des chercheurs et accroître la visibilité de la recherche française.

Anne-Catherine Rota
Research Intelligence chez Elsevier
a.rota@elsevier.com

Crédit Adobe stock, par Abdul Momin Séchage de tissus au soleil

UNIV-DROIT.FR : IDREF, GARANT DE LA FIABILITÉ DES DONNÉES

Lorsque la Conférence des doyens des facultés de droit a souhaité prolonger, sur son portail univ-droit.fr, la présentation des facultés par un annuaire exhaustif des professeurs et maîtres de conférences en droit et science politique, la question de l'affichage de données externes, sur chaque page des notices ainsi constituées, s'est très rapidement posée. Les juristes étant, à l'époque, pratiquement absents de HAL, c'est assez naturellement que la collecte des données présentes dans le Sudoc s'est imposée, ce qui supposait un alignement entre les notices individuelles univ-droit et le référentiel IdRef. Grâce à l'aide et au soutien des équipes de l'Abes, l'alignement initial a été réalisé très rapidement, et a permis d'afficher tous les ouvrages des universitaires, puis leurs thèses, sans qu'il leur soit nécessaire de compléter manuellement leur notice.

Par la suite, la collecte d'autres données non alors présentées sur IdRef (notices d'articles de revues ou de chapitres d'ouvrages) a rencontré plus de difficultés, notamment en raison du défaut d'identifiant de personne unique dans HAL, seul un tout petit nombre de juristes s'étant créé un IdHal. Là encore, la recherche de fiabilité des données exposées sur univ-droit a poussé à rechercher un alignement avec les identifiants disponibles, ce qui a été rendu partiellement possible dans le cadre d'un projet CollEx (Droit2Hal) : les plus de 60 000 notices d'articles de revues d'un éditeur (Dalloz) déposées dans HAL comportent l'ajout de l'identifiant IdRef d'universitaires qui n'étaient jusqu'alors pas présents dans HAL.

Prolongeant cette démarche, une importante refonte en cours exploite, toujours au moyen de l'IdRef, d'autres sources de données (Cairn, Isidore), via une interrogation de data.idref.fr et SciencePlus.

Enfin, pour permettre un meilleur référencement, une redirection de type univ-droit.fr/IdRef pointe vers chaque notice d'universitaire.

Gilles Dumont
Professeur de droit public à l'université Paris Cité
gilles.dumont@u-paris.fr

OPENEDITION : LE CHOIX D'IDREF POUR LE PROJET QUAMÉO

Le projet QUalité des Métadonnées d'OpenEdition (QUAMÉO), déposé par l'infrastructure nationale de recherche OpenEdition et l'Abes, est l'un des lauréats du troisième appel à projets du Fonds national pour la science ouverte (FNSO). Son objectif principal est d'améliorer la qualité des métadonnées associées aux publications numériques disponibles sur les plateformes d'OpenEdition, favorisant ainsi leur visibilité et leur signalement.

Le premier volet de ce projet se concentre sur l'enrichissement des métadonnées envoyées à Crossref pour l'attribution de DOI sur les plateformes Books et Journals.

Le deuxième axe du projet consiste à mettre en place une base d'autorités auteurs adossée à IdRef afin de couvrir les quelque 210 000 auteurs présents sur les plateformes d'OpenEdition. L'objectif est d'identifier les auteurs de manière unique en récupérant l'identifiant IdRef qui leur correspond et d'autres identifiants s'ils existent (ORCID, ISNI, idHAL...).

Les bénéfices seront multiples :

- Mise à disposition d'un index auteurs global pour les lecteurs
- Ajout des identifiants récupérés dans l'entrepôt OAI-PMH d'OpenEdition ainsi qu'aux métadonnées transmises à Crossref
- Partage des connaissances et des bonnes pratiques liées à l'identification des auteurs via la sensibilisation et la formation des auteurs et des équipes éditoriales

Les développements informatiques réalisés dans le cadre du projet seront rendus accessibles à la communauté, notamment aux pépinières de revues qui utilisent Lodel.

Ce projet consolidera la position d'OpenEdition dans le paysage de l'accès ouvert et de l'édition scientifique en lui permettant d'intégrer de nouvelles fonctionnalités au service de ses utilisateurs.

Émilie Cornillaux
OpenEdition - Service Données
emilie.cornillaux@openedition.org

ALIGNER LES DONNÉES DES CHERCHEURS DE MON ÉTABLISSEMENT SUR IDREF

DEUX MÉTHODES COMPLÉMENTAIRES

Identifiant pivot et agrégateur de ressources bibliographiques, IdRef permet d'identifier et de valoriser la production des chercheurs. La preuve par l'exemple avec le cas de Nantes Université.

Qui est affilié à mon établissement ? Qui a quels identifiants auteur ? Qui a publié quoi dans HAL ? Pour valoriser la production de ses chercheurs dans HAL ou mesurer l'adoption d'ORCID, chaque établissement se pose ces questions. L'Abes tente d'y répondre en s'appuyant sur IdRef en tant qu'identifiant pivot et agrégateur de ressources bibliographiques, et ce même quand une base tierce n'exploite pas nativement et systématiquement IdRef, ce qui est le cas de HAL.

PRENONS L'EXEMPLE DE L'UNIVERSITÉ DE NANTES[1] ET SUIVONS 5 CHERCHEURS ET 4 PUBLICATIONS PRÉSENTES DANS LE PORTAIL HAL NANTES UNIVERSITÉ CAR ELLES SONT RATTACHÉES À DES UNITÉS DE RECHERCHE NANTAISES.

V — Chercheur junior, travaille à l'université de Nantes depuis 2022 dans le même laboratoire que chercheuse W

W — Chercheuse W, travaille à l'université de Rennes dans le même laboratoire que chercheur V

X — Chercheur X, travaille à l'université de Nantes depuis 2000

Y — Chercheur Y, a travaillé à l'université de Nantes de 2000 à 2020 depuis il est en poste à Sorbonne

Z — Chercheuse Z exerce à Lille

PUBLICATION 1
Chercheur **V** Univ Nantes - Labo 1
Chercheuse **W** Univ Rennes - Labo 1

PUBLICATION 2
Chercheur **V** Univ Nantes - Labo 1
Chercheur **Y** Sorbonne Univ - Labo 2

PUBLICATION 3
Chercheur **Y** Univ Nantes - Labo 1
Chercheuse **Z** Univ Lille - Labo 3

PUBLICATION 4
Chercheur **X** Univ Nantes - Labo 4
Chercheuse **Z** Univ Lille - Labo 3

[1] Comme d'autres, Nantes a connu une évolution institutionnelle en 2022 passant de Université de Nantes (1962-2021) www.idref.fr/026403447/id à Nantes Université (2022-....) www.idref.fr/258086599/id Les travaux d'alignements évoqués dans cet article concerne les effectifs et publications de l'université de Nantes au sein de l'établissement expérimental Nantes Université.

LA MÉTHODE « ANNUAIRE »

Cette méthode exploite IdRef pour répondre à la question : « Qui [parmi les personnes de l'établissement] a quels identifiants auteur ?»

Données initiales fournies par l'établissement : un fichier csv avec une ligne par personne appartenant à l'établissement. A minima : nom | prénom | date de naissance[2] | laboratoire. Les chercheurs V et X sont dans ce fichier ; W et Y (et évidemment Z) n'y sont pas.

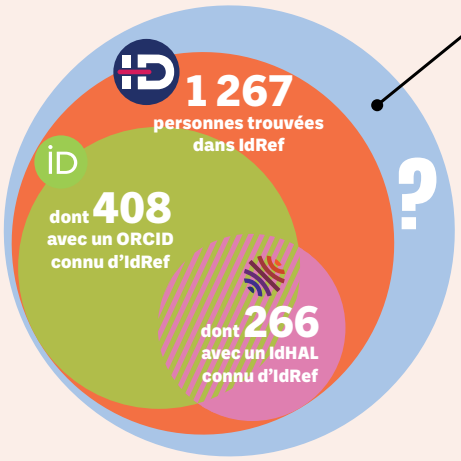
Source des données : extraction du SI RH, LDAP, collecte opérée en amont par le SCD et/ou la direction de la recherche, etc.

Volumétrie dans l'exemple de Nantes : 1 157 maitres de conférences et professeurs, 86 doctorants, 54 post-doctorants, et 345 personnes aux statuts variés. Le SCD[3] a choisi une liste cumulative (les personnels de l'université de Nantes hier et aujourd'hui) avec une interprétation très large de la notion de chercheur.

Résultat du travail de l'Abes : le même fichier csv enrichi des colonnes suivantes : candidat IdRef | ORCID connu dans IdRef / IdHAL connu dans IdRef.

Quand et à quelle fréquence utiliser cette méthode ? Chaque année pour couvrir les départs et arrivées des personnels ! Après une première fois, le suivi annuel est une routine légère pour l'établissement.

1 642
personnes dans l'annuaire U Nantes



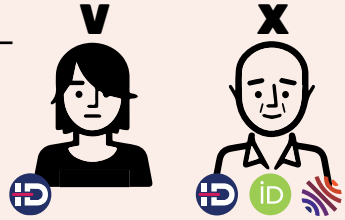
LE TRAVAIL À FAIRE POUR LE CORRESPONDANT AUTORITÉS :

> 367 personnes pour lesquelles il faut investiguer
- soit mettre en attente (les doctorants)
- soit créer un IdRef

> 8 personnes à desambiguiser

RÉSULTAT

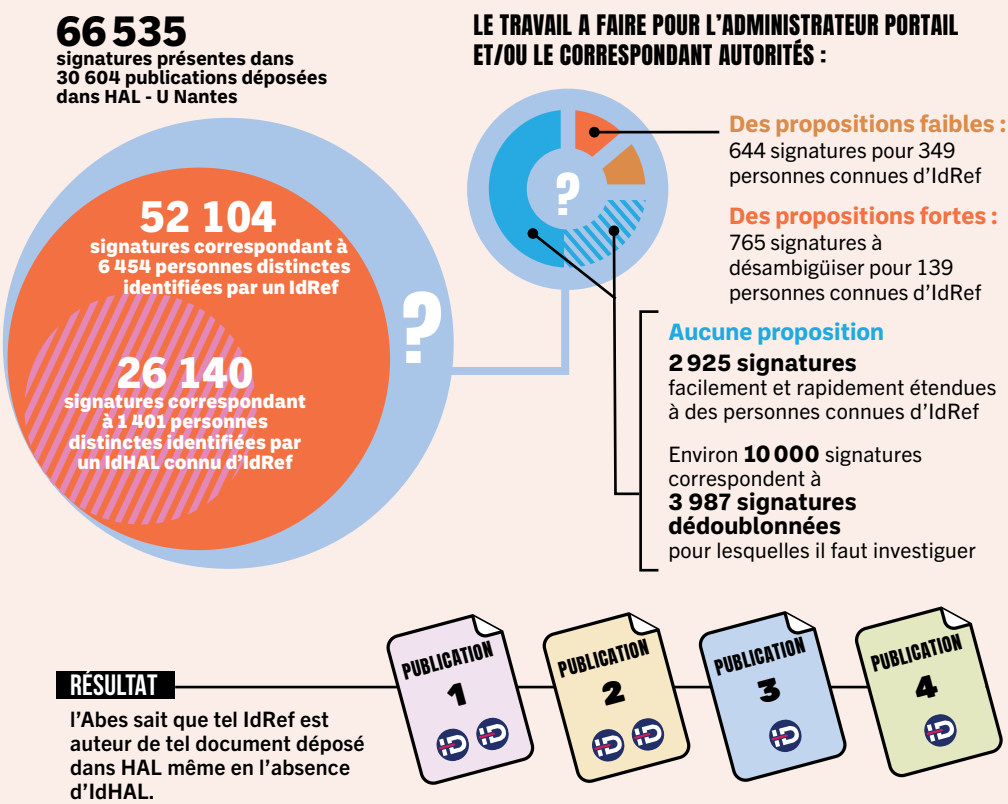
Chaque personne de l'établissement supposée avoir un IdRef en dispose. L'établissement ne gère plus un annuaire avec des chaines de caractères mais un annuaire d'identifiants.



[2] Bien que la loi 3 DS du 21 février 2022 et ses décrets d'application posent le principe de l'échange des données entre administrations, l'Abes a constaté des réticences de certains délégués à la protection des données concernant la date de naissance. Nous pouvons vous aider à les lever si besoin.
[3] Notamment Guillaume Godet et Natacha Martin : merci à eux !

LA MÉTHODE « PORTAIL HAL »

Cette méthode exploite IdRef et HAL pour répondre à la question : « Qui [qualifié par son IdRef] a publié quoi dans HAL ? »



Données initiales et source : l'Abes est autonome pour récupérer les données HAL une fois confirmation par l'établissement de la requête à passer à l'API de HAL pour cibler son périmètre institutionnel.

Volumétrie dans l'exemple de Nantes : le portail HAL Nantes Université restreint à la composante de l'université initiale i.e. les métadonnées des 30 604 publications déposées depuis l'origine soit 66 535 signatures.

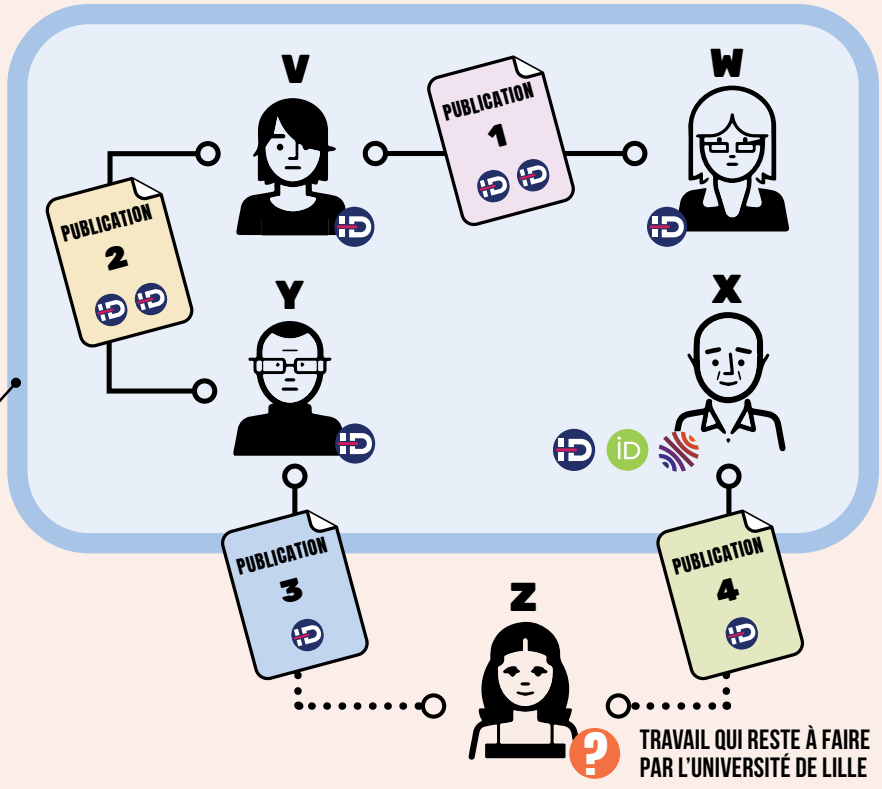
Résultat du travail de l'Abes : le programme d'alignement exploite le matériau bibliographique brut et privilégie certains critères (coauteurs, titres, discipline…).
Pour HAL Nantes Université, 6 454 personnes distinctes ont été repérées. L'exploitation de l'IdHAL ne permet d'identifier que 1 401 personnes. En faisant un détour par le document, l'Abes brasse plus de données et peut déduire des documents une meilleure identification des auteurs.

Quand et à quelle fréquence utiliser cette méthode ?
une seule fois car l'Abes traite depuis le printemps 2023 tous les nouveaux dépôts dans HAL.

COMPLÉMENTARITÉ DES DEUX MÉTHODES

Lorsque les deux méthodes sont combinées, un graphe de connaissances très complet est généré et exploité dans IdRef et dans data.idref.fr.

Dans notre exemple, le travail mené par les collègues de Nantes a permis d'inclure dans data.idref.fr les triplets qui concernent les chercheurs V, W, X, Y et leurs publications respectives. Le chercheur Z, qui travaille à l'université de Lille, n'a évidemment pas été traité par la méthode« annuaire ». Ses publications ont en revanche été traitées par la méthode « portail HAL» : un candidat IdRef a peut-être été détecté par le programme de l'Abes mais les collègues nantais n'ont pas vérifié, car ils avaient suffisamment à faire : ce serait aux collègues lillois de prendre le relais.



RÉSULTAT
Ensemble des triplets désormais dans data.idref.fr grâce à l'utilisation des deux méthodes par l'Université de Nantes

SI VOTRE ÉTABLISSEMENT N'A PAS EU RECOURS À CE SERVICE, N'HÉSITEZ PAS À PRENDRE CONTACT VIA LE GUICHET ABESSTP› IDREF › DONNÉES !

article établi par Benjamin Bober, Isabelle Mauger Perez, Carole Melzac, François Mistral
idref@abes.fr
Infographie : Anne Ladevie

Données liées ouvertes et référentiels public : *un changement de paradigme pour la recherche en sciences humaines et sociales*

Le partenariat entamé en 2019 entre le laboratoire LARHRA et l'Abes marque une étape décisive dans la collaboration entre bibliothécaires et chercheur·euse·s, et démontre la nécessité d'associer référentiels et ontologies pour la compréhension et la réutilisation des données issues de la recherche.

Un article publié dans *Arabesques* en 2017 faisait état d'un premier alignement avec IdRef de personnes recensées dans la plateforme symogih.org, un environnement virtuel de recherche (EVR) mis en place au Laboratoire de recherche historique Rhône-Alpes (LARHRA) en 2008 : « l'intégration des autorités SyMoGIH avec les IdRef doit faciliter l'ouverture de notre entrepôt vers d'autres réservoirs de qualité, tout en enrichissant les IdRef »[1]. Sept ans après, ce projet a connu des développements importants qui s'inscrivent dans une collaboration entre le laboratoire LARHRA et l'Abes formalisée en 2019 par une convention de coopération scientifique.

ENCOURAGER LA RÉUTILISATION DES DONNÉES DE LA RECHERCHE

Deux éléments principaux sont au cœur de cette démarche : d'une part, la publication avec les technologies sémantiques de données de la recherche afin de faciliter leur réutilisation ; d'autre part, l'enrichissement du référentiel IdRef avec les informations issues de la recherche. La finalité de cette opération est d'encourager la réutilisation des données pour de nouvelles recherches en sciences humaines et sociales (SHS), en application des principes FAIR (*Findable, Accessible, Interoperable, Reusable*). Pourquoi est-il essentiel, dans ce contexte, de pouvoir se référer à des autorités telles celles d'IdRef ?

Selon une intuition qui était à l'origine du projet symogih.org, il est indispensable en vue de la réutilisation des données de distinguer entre les questions de recherche d'un projet et l'information collectée pour y répondre[2]. Si, en effet, le savoir issu de la démarche scientifique peut être défini comme une interprétation du monde, un modèle qui répond aux questions des chercheur·euse·s, l'information collectée pour produire ce savoir doit viser une représentation la plus factuelle possible du monde étudié, c'est-à-dire des objets qui le composent, de leurs propriétés et de leurs relations[3]. Cette distinction permet de produire des données qu'on pourra réutiliser pour répondre à de nouveaux questionnements.

Grâce au web sémantique, il devient possible de créer un graphe géant de relations entre objets du discours scientifique, relations sémantiquement explicites, et de capitaliser ainsi l'information produite par chaque projet en permettant sa réutilisation pour de nouvelles recherches. La condition est l'identification précise des objets grâce aux référentiels. Si Google a su réaliser un *Giant Knowledge Graph* comportant, en mars 2023, 8 milliards d'objets identifiés et 800 milliards de « faits » (source : Wikipedia), pourquoi les SHS n'en feraient pas autant, notamment en utilisant IdRef ?

IDREF, PIVOT DE L'IDENTIFICATION DES OBJETS DU DISCOURS SCIENTIFIQUE

Pour que ce projet scientifique et technologique aboutisse, trois composantes sont indispensables : un référentiel partagé permettant d'identifier clairement les objets du monde (personnes, organisations, concepts, etc.) ; une méthode de modélisation des relations entre objets capable d'intégrer les approches de différentes disciplines ; une infrastructure distribuée durable (cf. l'illustration), permettant de soutenir la démarche de recherche et l'interconnexion des données existantes.

Le référentiel IdRef se prête bien à cette fin car il est connecté avec la bibliographie du Sudoc, ainsi qu'avec la plateforme Persée, les archives dans Calames ou encore l'entrepôt de publications *SciencePlus.abes.fr*[4]. Il peut servir comme l'un des pivots de l'identification des objets du discours scientifique : non seulement il fait le lien vers d'autres référentiels tel celui de la Bibliothèque nationale de France ou Wikidata, mais il admet un enrichissement par les chercheur·euse·s (soumis à un contrôle de qualité) et, en retour, il tire profit d'un processus de désambiguïsation collectif.

UNE APPLICATION DE GESTION COLLABORATIVE D'ONTOLOGIES

Il faut ensuite disposer d'une ontologie, c'est-à-dire d'un modèle conceptuel formalisé et partagé, modulaire et ouvert aux différentes disciplines scientifiques. Pour répondre à ce défi, le LARHRA a travaillé, sur le plan pratique, à la mise en ligne d'une application de gestion collaborative d'ontologies, OntoME[5]. Cette plateforme permet d'étendre les standards, tel le CIDOC CRM, afin de disposer de classes et propriétés qui correspondent aux besoins des différentes disciplines

[1] Pierre Vernus, « SyMoGIH, de l'UMR 5190 – Larhra, et les 'objets historiques' », *Arabesques,* 85 | 2017, 14.

[2] Francesco Beretta and Pierre Vernus, « Le projet SyMoGIH et la modélisation de l'information : une opération scientifique au service de l'histoire », *Les Carnets du LARHRA,* 1, 2012, 81–107.

[3] Tom Gruber, « Ontology », in Liu, Ling, and M. Tamer Özsu, éds., *Encyclopedia of Database Systems,* Second Edition (Springer, 2018), 2574–76 https://doi.org/10.1007/978-1-4614-8265-9

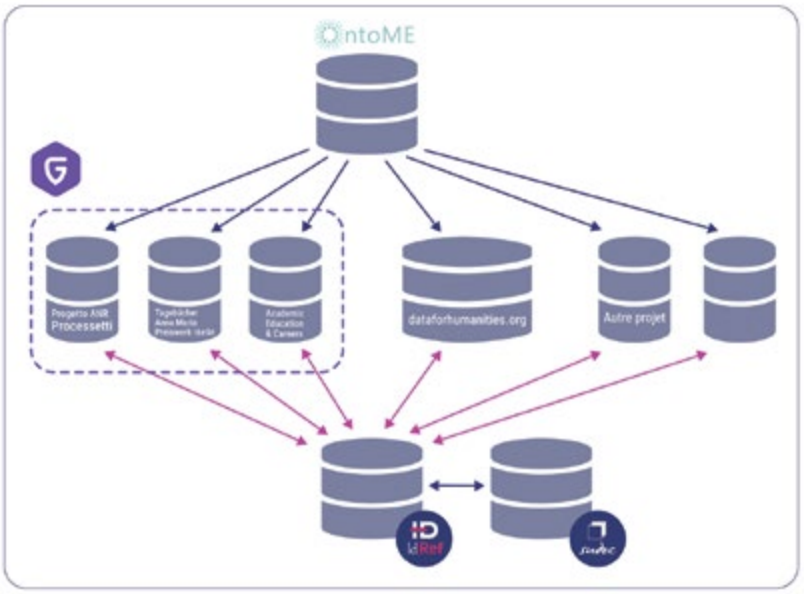
[4] Yann Nicolas, « Scienceplus.abes.fr : une nouvelle base de données au service de la science ouverte », *Arabesques,* 103 | 2021, 22.

[5] http://ontome.net

SHS, et de gérer des profils applicatifs qui facilitent l'appropriation du modèle par les chercheur-ses[6].
Sur le plan scientifique, l'utilisation de méthodologies de développement d'ontologies telle OntoClean, ainsi que l'analyse fondationnelle à l'aide de DOLCE, a permis de mettre en place un écosystème d'extensions du CIDOC CRM dans le projet *Semantic Data for Humanities and Social Sciences* (SDHSS)[7]. Cette méthodologie facilite également l'intégration d'autres standards, tels *Records in Contexts* (RiC) ou le *IFLA Library Reference Model* (LRM). À noter que l'écosystème d'ontologies SDHSS se limite à proposer un ensemble cohérent de classes et propriétés, afin de disposer d'un langage commun pour décrire les éléments essentiels de la vie sociale (le fait d'être propriétaire d'un objet, ou d'avoir un rôle dans une organisation, etc.), tandis que la gestion de vocabulaires contrôlés de types d'objets, ou de rôles sociaux, sont librement gérés par les chercheur-ses dans leurs projets respectifs, si possible en lien avec un référentiel comme IdRef.
Au niveau de l'infrastructure, un contrat de transfert de savoir-faire entre le CNRS et l'entreprise KleioLab a permis de créer un nouvel EVR, *geovistory.org,* qui remplace celui du projet *symogih.org* et intègre la plateforme *ontome.net.* Depuis cette année, le projet *LOD4HSS*[8], piloté par Tobias Hodel (professeur d'humanités numériques à l'université de Berne), vise à promouvoir la pérennisation de cette infrastructure, qui sera portée par un consortium international d'organismes publics, et à développer de nouvelles fonctionnalités, telle l'intégration avec les graphes sémantiques de documents au format XML, encodés selon les standards TEI ou EAD. IdRef s'inscrit dans cette vision d'avenir, notamment via l'enrichissement des notices d'autorité avec des informations issues de la plateforme *geovistory.org.*
Pour les chercheur·euse·s, l'utilisation de cet EVR permettrait d'éviter deux écueils majeurs. D'une part, le fonctionnement en silos, selon le principe « nouveau projet = nouvelle base de données », qui est problématique en raison du caractère temporaire des projets et qui conduit souvent à la disparition des plateformes, et des données, une fois les financements terminés. D'autre part, l'absence d'une sémantique commune rend la réutilisation des données difficile voire impossible. Même en se servant du même outil (que ce soit Heurist, NodeGoat ou Wikibase) les données restent « prisonnières » de dépôts étanches les uns aux autres et leur interopérabilité est mise à mal par des choix de modèles conceptuels divergents ou contradictoires[9].

VERS UNE NOUVELLE MANIÈRE DE PRODUIRE LE SAVOIR EN SHS

Certes, des méthodologies existent pour transformer ces données et les aligner avec les référentiels et une ontologie partagée. Un projet pilote a été mené dans le cadre de la collaboration entre l'Abes et le LARHRA, dans le contexte de l'ANR HisArc-RDF, qui a permis de créer un prototype de processus de transformation et publication de données sous forme de données liées ouvertes (Linked Open Data, LOD)[10] : après alignement avec les IdRef et en utilisant le standard FRBRoo de l'IFLA, une partie des données du projet PRELIB, consacré au monde littéraire breton, est désormais accessibles sur le serveur SPARQL du projet *dataforhumanities.org*[11]. Reste que cette démarche comporte des coûts supplémentaires, rarement prévus dans le budget des projets.
L'évolution vers la publication de données de la recherche sous forme de LOD alignés avec les référentiels (si possible produits dès l'origine comme tels) permet d'envisager un renouvellement important des SHS grâce à un changement d'échelle du volume d'information disponible, virtuellement infini et de bonne qualité, facilement réutilisable grâce aux technologies du web sémantique. Le potentiel est tel qu'on peut prévoir un changement de paradigme dans ces disciplines, une transformation de leur manière de produire le savoir et de former les nouvelles générations de chercheur·euse·s[12]. Pour ce faire, une infrastructure collaborative et ouverte telle *geovistory.org,* capable d'accueillir grâce aux méthodologies sémantiques une grande variété de projets en SHS, par exemple de type Collex-Persée, est indispensable. De même en va-t-il de l'intégration des compétences liées aux LOD dans les métiers des bibliothèques, de l'information et du patrimoine, afin d'accompagner les chercheur·euse·s, et le public, dans la transition numérique.



Francesco Beretta
Historien, spécialiste en systèmes d'information pour les sciences humaines et sociales, chargé de recherche au CNRS, UMR 5190 LARHRA, Lyon
francesco.beretta@cnrs.fr

[6] Francesco Beretta, « *A Challenge for Historical Research: Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME)* », Semantic Web, 12.2 (2021), 279–94, https://doi.org/10.3233/SW-200416

[7] Id., « Interopérabilité des données de la recherche et ontologies fondationnelles : un écosystème d'extensions du CIDOC CRM pour les sciences humaines et sociales », in Nicolas Lasolle, Olivier Bruneau, and Jean Lieber, éds, *Actes des journées Humanités Numériques et Web sémantique,* (Nancy, France, 2022), pp. 2–22 https://doi.org/10.5281/zenodo.7014341

[8] https://www.geovistory.org/lod4hss

[9] https://www.mediawiki.org/wiki/Wikibase/FAQ : "*Wikibase users can design their own data model. Are there downsides to this?*"

[10] https://dataforhumanities.org/sparql-endpoint/prelib-v1

[11] François Mistral, « Des catalogues de bibliothèques aux projets en humanités numériques : les autorités IdRef font le lien », *Arabesques,* 105 | 2022, 16-17.

[12] Francesco Beretta, « Données ouvertes liées et recherche historique : un changement de paradigme », *Humanités numériques*, 7, 2023, https://doi.org/10.4000/revuehn.3349

Rouverte en mai 2023 après 3 ans de travaux, la bibliothèque universitaire Lettres de Nantes offre des espaces modernisés et de nouveaux services lui permettant de jouer pleinement son rôle de cœur de campus.

BU Lettres de Nantes : *une rénovation alliant esprit vintage et modernité*

Construite en 1967, la bibliothèque universitaire Lettres de Nantes avait besoin d'une rénovation d'ampleur afin d'adapter ses bâtiments vétustes aux nouveaux usages des bibliothèques universitaires.

Après une phase d'étude et de programmation confiée au cabinet Aubry et Guiguet, à laquelle ont été associés le personnel de la bibliothèque et les délégués étudiants, le chantier a été lancé, doté d'un montant de sept millions d'euros (dont cinq millions financés par l'État dans le cadre du contrat de plan État Région 2015-2020 (CPER), un million par la Région Pays de la Loire et un million par Nantes Métropole).

L'équipe retenue pour la maîtrise d'ouvrage offrait la particularité d'associer un cabinet d'architectes (TICA architectes & urbanistes), des designers (Barreau et Charbonnet), une agence de graphistes composés de jeunes artistes nantais (Appelle-moi papa) et un bureau d'études (CETRAC).

Le programme préconisait de rester fidèle au caractère années 1960 du bâtiment et de tirer parti d'un environnement verdoyant et de volumes architecturaux de qualité. Pour autant, les espaces devaient être diversifiés, adaptés aux usages et besoins contemporains et permettre de proposer de nouveaux services.

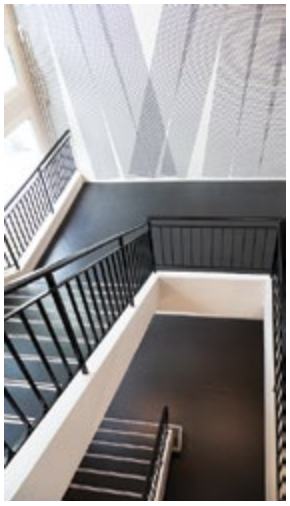
Programmés de janvier 2021 à mai 2023, les travaux se sont déroulés dans de fortes contraintes : travaux en site occupé, accès aux services et aux collections garanti pour les lecteurs, pas d'externalisation possible pour le stockage malgré la mobilisation d'espaces dans la BU Droit adjacente. À cela se sont ajoutées les différentes difficultés induites par l'épidémie de Covid-19 et le lancement d'un chantier dans le chantier, celui du clos couvert de la BU, pour un million d'euros.

L'ESPRIT ANNÉES 1960 RÉINVENTÉ

Le traitement des espaces a été fidèle à la demande de respecter l'esprit chaleureux du lieu tout en le réinventant : des portes ont été supprimées et des cloisons déplacées afin de créer des perspectives et d'ouvrir de nouvelles vues sur les bois environnants. Certains murs ont été traités en medium de couleurs sourdes, jaune profond et rouge mat. Les sols ont reçu un traitement différencié avec un sol souple de type linoléum dans la cafétéria et la circulation du rez-de-chaussée, traitée comme une « rue » ouvrant sur les différents espaces, et de la moquette partout ailleurs.

Tout en maintenant le caractère originel du lieu, l'équipe s'est attachée à créer autant d'atmosphères différentes qu'il y a d'espaces. Les grandes salles de lecture offrent chacune une ambiance propre, avec de grandes tables centrales équipées de liseuses, des places individuelles à proximité des fenêtres ou des îlots de travail entre les rayonnages ; les mezzanines – autrefois dédiées au stockage de collections – ont été ouvertes au public, transformées en lieux propices au travail studieux. Des niches ont également été vidées de leurs collections pour accueillir des places de travail protégées, voire intimes. Ailleurs, des canapés et des fauteuils installés dans des circulations ou des renfoncements ont immédiatement été adoptés par les lecteurs, au point qu'un complément de commande a dû être rapidement mis en œuvre.

L'envie de garder l'esprit des années 1960 a également conduit à réutiliser au maximum le mobilier. La plupart des tables ont été conservées mais leur plateaux refaits et complétés de blocs prises. Des rayonnages de magasin ont trouvé une nouvelle vie en salle, associés à des rayonnages neuf. Des chutes de chantier (bois ou métal) ont été utilisées pour créer de nouveaux éléments mobiliers. Chaque salle de lecture comporte ainsi au moins trois modèles de rayonnages différents, unifiés par une signalétique harmonieuse et visible. Le peu de mobilier acheté notamment auprès de « J et J » à Bruxelles est parfaitement en phase avec l'esprit vintage recherché. Sans nier un souci de maîtrise des coûts, le choix de réutiliser la grande majorité des équipements d'origine découlait également du souhait de l'univer-



sité d'être cohérente avec ses objectifs de développement durable.

NOUVEAUX ESPACES, NOUVEAUX SERVICES

Autant d'espaces, autant de services. À l'origine du projet, les quatre demandes principales concernaient la création d'espaces de travail en groupe, d'un lieu de restauration, de salles d'innovation pédagogique (20 et 30 places) et le déploiement de prises électriques en nombre suffisant. À la livraison des travaux, les engagements ont été tenus. Seize salles de travail en groupe ont été créées, pouvant accueillir de quatre à huit personnes. À rebours, une salle « super silence » accueille des places de travail individuelles, séparées par des écrans acoustiques. Des *phones rooms* et des casiers connectés ont également été installés.

La réflexion sur la rénovation a coïncidé avec le départ à la retraite d'un personnel logé. Le logement a pu être utilisé pour créer un lieu de restauration sur deux niveaux et donnant sur la « rue » principale du rez-de-chaussée, sans pour autant diminuer la surface des lieux dédiés aux services ou aux collections. À l'issue d'une procédure d'autorisation d'occupation temporaire (AOT), une entreprise de l'économie sociale et solidaire a été retenue avec une offre complémentaire de celle du Crous : localisme, production bio ou raisonnée, démarche écoresponsable, etc. Elle propose une petite restauration chaude entre 11 h 30 et 14 h, complétée par des distributeurs de boissons et de *snackings* sur le reste des horaires d'ouverture.

Suite aux échanges avec les deux vice-présidents étudiants de Nantes Université, une salle polyvalente en mode « salle sur demande » sur 140 m² a été imaginée pour répondre aux différents besoins des organisations et associations étudiantes : conférences, projections, *workshop,* etc. Équipée d'un mobilier sur roulettes, ce plateau est accessible sur réservation. Quand elles ne sont pas réservées, ces salles sont librement utilisables par les lecteurs.

Trois ans ou presque de travaux ont exigé de l'équipe endurance et souplesse, mais ont également donné lieu à de nombreuses réflexions autour de ses pratiques. Les différents points d'accueil étaient spécialisés : accueil principal avec du prêt-retour et du renseignement de premier niveau (auquel l'ensemble du personnel, toutes catégories confondues, participait), une banque de communication pour les collections en magasin, un bureau de renseignements pour un renseignement de second niveau et les inscriptions. Au fil des travaux et de la neutralisation de certaines zones, les fonctions des différents accueils ont évolué. Rebaptisés « accueils principaux » ou « points d'accueil », leurs fonctions ont été mutualisées. À l'exception des inscriptions, toutes les opérations se font indistinctement à tous les points d'accueil, avec la participation de toutes les catégories d'agents. Le bénéfice est évident pour les lecteurs mais également pour le personnel, qui s'est acculturé à l'ensemble des missions d'accueil.

UNE BIBLIOTHÈQUE CŒUR DE CAMPUS

Les usagers sont de plus en plus nombreux à fréquenter la bibliothèque. D'environ 1 200 entrées par jour en moyenne en 2019, la fréquentation a atteint les 2 000 entrées par jour dès fin septembre et dépasse désormais les 3 000. Tous les espaces ont immédiatement trouvé leurs occupants, des plus studieux – la salle « super silence » ne désemplit pas et les salles de travail en groupe sont prises d'assaut – aux plus décontractés : beaucoup de siestes sur les nouveaux canapés mais également directement sur la moquette ou sur les poufs. Si la « rue » du rez-de-chaussée est très animée, l'atmosphère des espaces de travail et de l'étage, y compris ses salons de détente, reste très calme, voire feutrée, grâce à l'aménagement intérieur et aux matériaux choisis. Grâce à ses nouveaux espaces très accueillants, sa cafétéria et ses salles ouvertes à la réservation, la bibliothèque joue pleinement son rôle de cœur de campus.

D'autres évolutions sont d'ores et déjà programmées : rénovation des façades et changement des huisseries en 2025, mise en œuvre de la RFID ou encore coconstruction de nouveaux services dans le cadre du schéma directeur Vie de campus, notamment un guichet unique de renseignement en présentiel et en ligne.

LE PROJET EN CHIFFRES

21 CORPS DE MÉTIER : architectes, designers, graphistes, acousticien, chef de chantier, ingénieurs, contrôleur technique, coordonnateur de sécurité et de protection de la santé, plaquistes, menuisiers, charpentiers, fabricants de mobilier, plombiers, désamianteurs, électriciens, peintres, ascensoristes, soliers, fabricants de signalétique, démolisseurs, maçons, déménageurs, etc.

355 000 VOLUMES (soit près de 10 km) de livres et de revues déplacés, certains plusieurs fois.

Une fréquentation qui a doublé avec des pointes régulières à près de 3 500 entrées par jours.

LAURE TEULADE
Responsable de la bibliothèque universitaire de Lettres de Nantes Université
laure.teulade@univ-nantes.fr

Contact presse :
Émilie Le Cour
Responsable Mission communication, bibliothèques universitaires de Nantes Université
emilie.lecour@univ-nantes.fr

Nantes Université

data.idref.fr : un référentiel d'autorités dans le web sémantique pour l'ESR et au-delà

Lancé en 2018, le triplestore data.idref.fr a progressivement étendu son périmètre au-delà des seules données de l'Abes et compte aujourd'hui plus de 6 millions de données.

Approfondissant la stratégie d'exposition des données sur le web sémantique, le triplestore data.idref.fr a été lancé en 2018, permettant d'interroger en SPARQL et en RDF les données d'IdRef, alimentées par les réseaux d'utilisateurs des applications de l'Abes. Dès le départ data.idref.fr a été conçu comme un miroir de la base IdRef elle-même, et non pas comme un *dump* mis à jour périodiquement comme c'est fréquemment le cas : chaque modification d'IdRef y est répercutée en temps réel. Cinq ans plus tard, il est temps de faire un bilan du chemin parcouru par ce service.

Une base d'entités en plein essor…

Un nombre donne une première idée de l'évolution de la base : alors qu'à son ouverture elle contenait 110 millions de « triplets », c'est-à-dire de données, elle en compte désormais, fin 2023, près de 275 millions. Dans le détail, les entités sont passées de 3,5 millions à plus de 6 millions en 2023, dont près de 4 millions pour les seules personnes. Cette progression reflète pour partie les créations d'autorités dans IdRef. Mais elle est aussi le résultat d'enrichissements successifs de la modélisation. Par petites touches, toujours plus d'informations de l'Unimarc natif d'IdRef ont été extraites : libellés, genre pour les personnes, notes, identifiants externes, relations entre entités (pour les organisations notamment)[1].

… qui est aussi une base bibliographique

D'autre part, data.idref n'est pas seulement un pur référentiel d'autorités mais aussi un réservoir de références bibliographiques. À l'origine, on n'y trouvait que celles issues du Sudoc, ainsi que les thèses. Leur description, tout en restant succincte, a également été enrichie pour faciliter les recherches : précision des types de documents, dates de publication, nombre de localisations dans le Sudoc (holdings). IdRef étant devenu au fil du temps le pivot des applications de l'Abes, data.idref.fr se devait de refléter cette centralité. Cela a été fait en intégrant les références issues d'autres applications de l'Abes : Calames, le catalogue des archives et manuscrits de l'enseignement supérieur, et SciencePlus, autre triplestore hébergeant une sélection des références d'articles et chapitres de documentation électronique provenant d'éditeurs ou diffuseurs, et où les auteurs sont identifiés à chaque fois que possible à des entités IdRef.

data.idref.fr dans IdRef !

Plusieurs informations proposées dans IdRef illustrent le type de service que peut apporter un triplestore : pour chaque auteur les listes de ses coauteurs, des collectivités associées et de ses champs disciplinaires, ou pour une organisation les autres collectivités liées, sont fournies par des requêtes SPARQL envoyées de manière dynamique à data.idref.fr, utilisées donc comme des webservices.

Une intégration dans un écosystème plus large

Parallèlement, IdRef et data.idref ont progressivement étendu leur périmètre au-delà des seules données de l'Abes, par des alignements (réalisés par l'Abes ou produits par les professionnels) avec des gisements documentaires extérieurs : BnF, HAL, Cairn, OpenEdition, Erudit… Ces alignements permettent d'intégrer dans le triplestore les références bibliographiques à partir de toutes ces sources, réunies autour des identifiants IdRef, et donc de les interroger en une seule requête, pour les réutiliser[2].
Les documents signalés dans data.idref sont ainsi passés de 11 à 16,2 millions depuis 2018, et sont liés aux entités IdRef par 57 millions de liens distincts, dont plus de 28 millions de relations de contributions.

Au-delà de ces sources documentaires, d'autres alignements d'entités intègrent toujours davantage data.idref.fr dans un écosystème plus vaste : ISNI, VIAF, Wikidata, ORCID, ROR, permettant de rebondir vers d'autres environnements, notamment via le web sémantique[3].

Michaël Jeulin
Gestionnaire de métadonnées, Service Outils et Méthodes de l'Abes
jeulin@abes.fr

[1] Documentation du modèle de données : **https://documentation.abes.fr/aideidrefdata**

[2] Voir l'article consacré à SoviSu+ pages 14 et 15.

[3] Une sélection de triplestores « amis » est proposée en page d'accueil de data.idref.fr, et parmi les exemples de requêtes proposés : **https://data.idref.fr/yasgui.html,** des requêtes « fédérées » avec data.bnf.fr, data.persee.bnf.fr, **https://query.wikidata.org**

LES CATALOGUES ET GISEMENTS DOCUMENTAIRES DANS DATA.IDREF.FR,
en nombre de liens document - contributeur

Sudoc : 24,4 millions
BnF : 6 millions
HAL : 3,6 millions
theses.fr : 2,4 millions
SciencePlus.abes.fr : 580 000
Cairn.Info : 208 000
OpenEdition Journals : 120 000
Calames : 21 000*
Erudit : 2 106*

* *pour Calames et Erudit, des chargements du rétrospectif sont prévus.*

Retour sur le congrès **ELECTRONIC THESIS AND DISSERTATIONS 2023**

Organisé en Inde du 26 au 28 octobre, le congrès ETD 2023 a mis à l'honneur les dispositifs du pays d'accueil et a abordé plusieurs grandes problématiques dont l'importance de la qualité des métadonnées.

Le congrès international *Electronic Thesis and Dissertations* (ETD) 2023 a eu lieu du 26 au 28 octobre dans la ville de Gandhinagar, en Inde, dans l'État du Gujarat. Organisé par INFLIBNET (*Information and Library Network*), en association avec l'organisation NDLTD (*Networked Digital Library of Theses and Dissertations*), il a accueilli 260 participants, pour une quinzaine de nationalités différentes.

L'Inde et son portail national des thèses Shodganga (Shodh = rechercher et trouver ; Ganga = le Gange) ont été mis à l'honneur. Shodhganga a été créé en 2009, à l'instigation de la Commission des subventions aux universités, l'organisme qui accrédite et finance les universités indiennes, et sa gestion a été confiée à INFLIBNET. Basé sur DSpace, il s'agit à la fois d'un catalogue et d'une plateforme de diffusion. Le format de métadonnées utilisé est une adaptation du format ETD-MS, fourni par NDLTD. Le portail référence aujourd'hui 490 000 thèses soutenues accessibles en ligne, avec un accroissement annuel d'environ 53 000 thèses. Le dépôt électronique a été rendu obligatoire en Inde en 2017 par une réglementation nationale. 840 institutions alimentent aujourd'hui le portail. Si ce chiffre est important, il ne représente néanmoins pas toutes les universités, un certain nombre d'entre elles restant réfractaires au projet, et ce malgré le soutien affiché du ministère de l'Éducation.

Les autres thématiques abordées lors du congrès étaient :

L'importance de la qualité des métadonnées descriptives des thèses, indispensable pour que les métadonnées puissent être exploitées et réutilisées, mais aussi pour augmenter la visibilité, au niveau international, de la recherche académique locale.

La prise en compte de l'expérience utilisateur, de l'ergonomie, des performances et de l'accessibilité dans la maintenance et l'amélioration des outils. L'évaluation des usages se fait encore essentiellement à travers des enquêtes de satisfaction, mais les institutions s'appuient également sur des outils d'analyse statistique.

L'*open access*. Malgré l'absence de politique nationale, la question de l'*open access* est centrale pour les pays en voie de développement ou émergents, qui accèdent difficilement aux ressources payantes. La question du respect du droit d'auteur et la crainte du plagiat constituent parfois un frein à l'ouverture des thèses.

Les entrepôts de données de la recherche. Les enjeux de l'*open access* s'appliquent aux jeux de données. Face à la multitude de plateformes existantes, il semble indispensable de mettre en place des entrepôts nationaux centralisés afin de faciliter le travail des chercheurs. L'Inde souhaite par ailleurs proposer la mise en place d'un entrepôt de données international consacré à la question climatique.

La bibliométrie et les statistiques. Que ce soit dans le WoS ou dans Scopus, les thèses électroniques sont très fréquemment citées, mais leur impact sur la recherche reste méconnu. Or, mesurer l'impact des thèses sur la recherche scientifique a autant d'importance que mesurer l'impact des autres publications scientifiques.

Une présentation a également porté sur le **portail national des thèses iraniennes.** Le dépôt électronique des thèses a été rendu obligatoire en Iran en novembre 2016. Il s'effectue dans l'application Sabt et les thèses sont diffusées sur la plateforme Ganj, les deux outils étant gérés par IranDoc (Iranian Documentation Center).

Si la France a pu, pendant un temps, se prévaloir d'une certaine avance en matière de gestion centralisée des thèses électroniques, ce n'est plus le cas désormais avec la multiplication des programmes nationaux ailleurs dans le monde. Les problématiques sont aujourd'hui communes : qualité des métadonnées, accès ouvert, satisfaction des usagers, suivi bibliométrique, gestion des données de la recherche. Il est indispensable que l'Abes soit présente dans les instances internationales afin de se tenir informée, de partager son expérience avec les autres institutions, et, en retour, de s'inspirer des bonnes pratiques mises en place au-delà de ses frontières.

Maïté Roux

Service des Thèses, Département des métadonnées et services aux réseaux de l'Abes

m.roux@abes.f

INFLIBNET (Information and Library Network) est un opérateur national, rattaché au ministère de l'Éducation indien, qui centralise les missions relatives à la gestion de l'information scientifique et technique : formation des bibliothécaires, SIGB, catalogue collectif, portail national des thèses, achats mutualisés des ressources électroniques, portail d'accès aux ressources électronique, plateforme de diffusion des publications scientifiques, etc. Pour en savoir plus : https://www.inflibnet.ac.in

NDLTD (*Networked Digital Library of Theses and Dissertations*) est une organisation internationale à but non lucratif basée aux États-Unis qui promeut la création des thèses et mémoires sous forme électroniques, leur signalement, leur archivage et leur mise à disposition en accès libre. Elle finance et met notamment à disposition gratuitement le méta-moteur de recherche *Global ETD Search*[1] qui référence et donne accès à près de 6,5 millions de thèses électroniques. Au sein du « *board* » de NDLTD, la France est représentée par Joachim Schöpfel (université de Lille) et Maïté Roux (Abes).

[1] http://search.ndltd.org

(Portrait)

Jean-Hugues MORNEAU,

responsable du traitement, du signalement et de la valorisation des thèses d'exercice à la Bibliothèque Médecine – Pharmacie de l'Université Grenoble Alpes

Parlez-nous de vos fonctions actuelles

Au sein du réseau des bibliothèques universitaires de Grenoble, je m'occupe de la collecte et du catalogage des thèses d'exercice de médecine – pharmacie et des mémoires de maïeutique. La plupart sont ensuite mises en ligne sur l'archive ouverte Dumas. Je participe aussi aux formations destinées aux étudiants (recherche documentaire, Zotero) ainsi qu'aux professionnels (indexation Rameau). Enfin, je porte la casquette de correspondant Autorités Rameau pour mon établissement.

Quelles sont les étapes qui vous semblent les plus importantes dans votre parcours professionnel ?

Mon parcours est le résultat de rencontres. C'est en suivant les traces de Fabien, un copain de fac d'histoire, que j'ai intégré l'année spéciale du DUT info-com de Grenoble. J'y ai rencontré Frédéric Saby, qui m'a incité à candidater sur un emploi au sein du réseau culturel français à l'étranger. Recruté par Renée Herbouze, j'ai ensuite dirigé la médiathèque de l'Alliance française de Quito, en Équateur. Après être devenu bibliothécaire adjoint spécialisé, j'ai exercé cinq ans au SCD de l'Université des Antilles et de la Guyane. En Martinique, j'ai été formé à l'indexation Rameau par Lucien Pavilla ; Sylvain Houdebert m'a quant à lui confié la responsabilité de la réinformatisation pour la section Guyane. Enfin, à Grenoble, Agnès Souchon a accepté d'ouvrir l'archive ouverte Dumas aux thèses d'exercice. Je tenais à remercier ici ces collègues pour la confiance qu'ils m'ont accordée. Et pour finir, spéciale dédicace à mes supers collègues cyclistes de Cyclo-Biblio !

Quelles sont vos relations avec l'Abes ?

J'ai participé plusieurs fois, et avec grand plaisir, aux journées de l'Abes. Je n'ai pas de relations régulières avec l'agence mais il m'est arrivé de la solliciter sur des points précis. Avec le recul, je réalise que j'ai souvent posé à l'Abes des questions insolubles. Par exemple, celle des thèses d'exercice, une entité aux contours très flous. L'Abes n'a pas réponse à tout et il faut faire preuve de bon sens pour continuer d'avancer sur certains dossiers.

Quels défis majeurs l'Abes aura-t-elle, selon vous, à relever dans les prochaines années ?

Mettre à la retraite CBS, le socle technique du Sudoc sur lequel je travaille depuis 2002 ! Comment ne pas évoquer ensuite l'éléphant au milieu de la pièce : une version LRMisée du catalogue Sudoc ? Cela fait des années que nous nous sommes engagés dans la transition bibliographique. Il est primordial qu'elle se matérialise enfin pour le plus grand bénéfice de nos usagers. Enfin, il faudra sûrement mener de nouvelles expérimentations autour de l'intelligence artificielle appliquée aux données catalographiques.

Qu'appréciez-vous le plus dans votre métier ?

Son étonnante diversité. Ma carrière atypique m'a permis de toucher à tant de choses : encadrement d'équipe, acquisition, gestion de collections, réinformatisation, formation, archives ouvertes… Côté catalogage, le métier n'a cessé d'évoluer avec la transition bibliographique, ce qui est très stimulant. Et avec le web sémantique, le travail collectif des catalogueurs constitue désormais un élément clef du nouvel écosystème mondial des données. C'est une source de fierté et de motivation !

Qu'est-ce qui vous énerve le plus ?

Le manque de reconnaissance de l'importance du catalogage et des compétences techniques qui l'accompagnent. Les conséquences sont concrètes : en 15 ans, j'ai vu fondre les postes de catégorie B dédiés au catalogage, ceux-là même dont nous avons besoin pour mener à bien la transition bibliographique et produire des métadonnées de qualité pour alimenter la science ouverte. J'avais évoqué ce fait au micro lors des journées de l'Abes et cela avait provoqué une salve spontanée d'applaudissements dans la salle.

Quelle image donneriez-vous pour définir l'Abes ?

L'Abes est le sanctuaire où officient les expert(e)s de la technique des bibliothèques. Un pèlerinage annuel permet de les rencontrer ! Plus sérieusement, L'Abes représente pour moi la coopération fructueuse et au long cours entre bibliothécaires et informaticiens.

Votre expression favorite ?

« Et en français, ça donne quoi ? ». C'est la pique que j'envoie aux gens qui abusent des termes en anglais.