Les référentiels dans la documentation scientifique : quelques exemples

Romain Vanel

LES RÉFÉRENTIELS DANS LA DOCUMENTATION SCIENTIFIQUE : QUELQUES EXEMPLES

ROMAIN VANEL

Résumé. L'objectif de cette communication est de sensibiliser au rôle déterminent que prennent les référentiels dans le paysage de l'information scientifique. Il s'agit de présenter la notion de référentiel et son application dans les bases documentaires de l'enseignement supérieur français. Après avoir défini les termes et évoqué brièvement les principaux référentiels utilisés nous verrons dans quelles mesures les données hébergées sont ouvertes, intéroperables et partagées.

Ce document est le texte remanié d'un atelier donné lors de l'action nationale de formation (ANF) intitulée *Accès ouvert : rêve ou réalité*[1], organisée par le réseau national des bibliothèques de mathématiques[2] du 16 au 20 octobre 2017, au Centre international de rencontres mathématiques de Marseille Luminy.

Table des matières

Aujourd'hui toutes les bases de données documentaires utilisent des référentiels. Ils permettent de faciliter la saisie des informations dans la base, de créer des liens entre divers éléments et ainsi de garantir la qualité des données produites. La consultation, la gestion et l'administration des bases est alors plus simple, plus rationnelle, et la qualité des données progresse.

Date: 17 novembre 2017.

Key words and phrases. référentiels, données, catalogues, identifiants.

Il m'est agréable de remercier Ariane Rolland pour ses nombreuses relectures ainsi que l'équipe du RNBM pour son accueil.

1. Programme complet de la formation et détails sur `http://www.rnbm.org/evenement/journees-rnbm-2017-acces-ouvert-reve-ou-realite/`.

2. RNBM, GDS 2755, `http://www.rnbm.org/`.

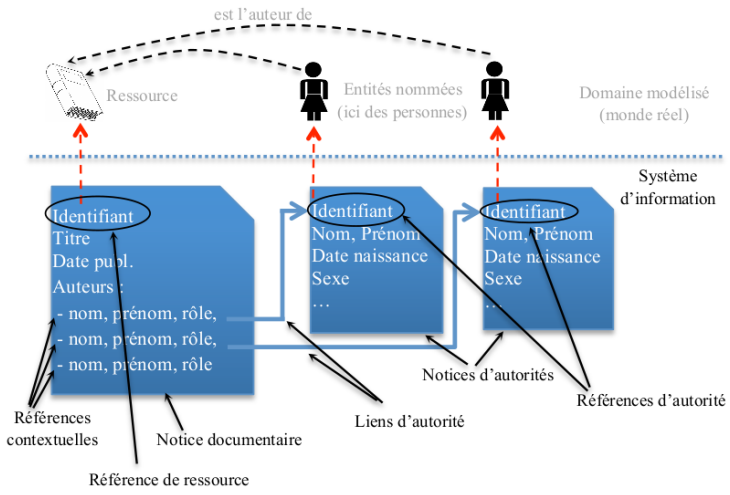


FIGURE 1. Modélisation d'un catalogue de bibliothèque. Source Chein 2015 [6] p. 4.

Par exemple, les bibliothèques gèrent, en parallèle de la description des documents, des notices d'autorités, décrivant les auteurs et sujets présents dans la base. Souvent, ces données sont contrôlées, et une notice d'autorité est ainsi liée à plusieurs documents. Les bases de données bibliographiques utilisent également des listes fermées pour certaines informations (par exemple le type de documents, les langues, les thèmes…).

Il y a encore quelques temps, chaque organisme gérait sa base, hébergeait et contrôlait ses données et ses propres référentiels. Ceux-ci contenaient souvent des informations équivalentes mais n'étaient pas connectés. La base d'autorité de la bibliothèque X n'était pas liée à celle de la bibliothèque Y, alors qu'elles contenaient souvent de nombreux auteurs en commun. L'enjeu documentaire est maintenant l'ouverture vers l'extérieur des données contenues dans ces référentiels. Le « Web de Données » a besoin de données structurées, fiables, vers lesquelles des liens peuvent être établis. Les structures documentaires peuvent fournir ces données et en garantir le contrôle et la qualité.

Nous allons voir dans quelle mesure l'utilisation des référentiels permet l'ouverture des données et facilite ainsi le partage de la connaissance.

Dans un premier temps, nous définirons les termes et tenterons de distinguer les notions de référentiels et d'autorités en les replaçant dans leurs contextes. Puis nous évoquerons quelques-uns des référentiels utilisés par les organisations documentaires et leurs principales caractéristiques. Enfin nous verrons que l'ouverture et l'exposition des données font des référentiels les pivots des systèmes d'informations documentaires [3].

1. DÉFINITION

Avant de définir ce qu'est un référentiel et de tenter d'en comprendre l'importance pour l'accès aux ressources, il convient de revenir sur l'architecture d'un catalogue de bibliothèque.

3. Nous ne cherchons pas ici l'exhaustivité, qui est d'ailleurs impossible à atteindre, tant pour la liste des référentiels présentés, que pour les liens possibles. L'objectif est de sensibiliser aux notions essentielles et aux enjeux à venir.

1.1. **Qu'est ce qu'un catalogue ?** Dans un rapport technique de 2015, trois chercheurs du Lirmm [4] ont modélisé [5] la notion de catalogue documentaire (fig. 1). On distingue dans un catalogue plusieurs éléments :

La notice documentaire, qui a pour identifiant un RR (*référence ressource*) est composée de plusieurs RC (*références contextuelles*). Ce sont les références aux contributeurs du documents (auteurs, directeurs de thèses...).

La norme de catalogage (notamment dans le Sudoc) impose que ces RC, soient liées à des notices d'autorités, qui ont pour identifiants des RA (*références d'autorité*).

Dans certaines notices, il y a (comme dans la figure 1) plus de RC (*références contextuelles* que de RA. Les liens ne sont pas faits entre la notice bibliographique et les autorités. On ne fait le lien que sur l'identifiant de la notice. L'auteur ou le contributeur, via leurs identifiants, deviennent dans ce contexte, de simples numéros !

Dans la figure 1 les RC ne concernent que les auteurs. La norme est la même pour les sujets (indexation matière, lieux...).

1.2. **Qu'est-ce qu'un référentiel ?** Pour Wikipedia :

« C'est un ensemble de bases de données contenant les « références » d'un système d'information [6]. »

Ils organisent les données et les métadonnées des systèmes, en garantissant la qualité de l'information qu'ils contiennent. C'est l'information « de référence ». Ils constituent un jeu de données, suffisamment fiables, pour qu'elles puissent être ouvertes, réutilisées [7] et enrichir d'autres données, avec lesquels ils peuvent être *alignés*.

C'est la colonne vertébrale d'un système d'information. Dans la mesure du possible toute information qui doit être normalisée, doit y être liée.

En reprenant l'exemple de la figure 1 p. 2, on pourrait envisager de construire un référentiel à partir de la liste des RA.

Dans un contexte bibliographique on trouve de nombreux référentiels, qui n'apparaissent parfois que sous une forme de liste contrôlée : lieux, langue, type de document, dates, structures, personnes,...

1.3. **Autorités ou référentiels ?** Les autorités font partie des référentiels. Elles garantissent la qualité des données bibliographiques [8].

Une autorité est créée et contrôlée par un humain. Elle se base sur des sources [19]. Dans le contexte des bibliothèques travaillant dans le Sudoc, les sources des informations contenues dans une notice d'autorité doivent apparaître dans le champs 810 [9] de ladite notice.

Mais les différences pourraient s'estomper avec l'entrée dans le Web de données. Les autorités sont normalisées, liées. Tout comme les données des référentiels, elles entrent dans le champs des ressources, identifiées comme telle [10]. Elles ont, par exemple, un identifiant unique.

On rentre dans une logique de graphe [11] ou les liens entre les entités prennent de plus en plus en plus d'importance.

4. UMR 5506, CNRS/Université de Montpellier.
5. Chein 2015 [6] p.3-4.
6. Voir Wikipédia [22].
7. F. Mistral dans Rousseaux 2017 [19].
8. O. Rousseaux 2017 [19].
9. « Aucune information trouvée dans le corps d'une notice d'autorité (zones 1XX à 7XX) ne doit être présente sans être justifiée et explicitée par une source (au moins) citée en zone 810. » [1].
10. Voir Y. Nicolas dans Rousseaux 2017 [19].
11. Voir Boulet 2017 [4].

Pour les bibliothèques la notion de notice change. Les langages de structuration (Unimarc, Marc21,...) ont été conçu dans le but d'échanger les notices. Aujourd'hui la bibliothèque doit structurer, décrire et qualifier des données, pour les exposer et les partager. Les référentiels permettent de contrôler ces données en leur apportant une garantie de qualité.

Tout ceci devient nécessaire dans le cadre de l'adoption du modèle IFLA-LRM [12]. Toutes les entités vont avoir besoin de référentiels pour être contrôlées. V. Boulet [13] cite l'exemple de l'entité « manifestation », qui pour exister, a besoin des entités « expressions », « oeuvre » « personne » mais aussi « date » ou « langue » qui ne rentrent habituellement pas dans le giron des autorités telles qu'on les définit.

1.4. **La qualité au coeur du système.** Plus le nombre de liens est important, plus la ressource liée (par exemple une notice bibliographique) sera visible. La qualité des notices, des liens, et des données vers lesquelles ils pointent est donc primordiale. Ainsi, une thèse est un type de document qui concentre beaucoup d'informations, et permet de nombreux liens. Dans les métadonnées d'une thèse on peut trouver :

- doctorant / auteur : personnes (avec fonction)
- jury : personnes (avec fonction)
- lieu : lieu...
- établissement de soutenance : structure
- numéro national
- sujets...

Cet exemple comporte de nombreux liens, visibles dans le Sudoc et dans Theses.fr :

- `http://www.sudoc.fr/200911481`
- `http://www.theses.fr/2016TOUR2012`

L'Adum (accès au doctorat unique et mutualisé) utilise IdRef pour identifier les doctorants dans les outils de gestion des écoles doctorales [14], évitant ainsi, de potentielles erreurs dans l'identification des doctorants.

Les référentiels sont plus que jamais diffusés sur le web (par exemple les thèses dont les données peuvent voyager entre le Sudoc, Theses.fr et Hal...), dans des formats ouverts, réutilisables. Ils sont le plus souvent diffusés sous des licences ouvertes, qui permettent leur partage, leur réutilisation.

Cette exposition (qui est aussi celle de l'institution) exige de veiller à la qualité des données.

2. Quels référentiels ?

2.1. **Personnes et collectivités.**

2.1.1. *L'IdHal.* Le CCSD, avec la mise en production de la version 3 de l'archive ouverte Hal, a développé la base AuréHal (Accès unifié aux Référentiels Hal). Cet outil permet d'accéder aux référentiels des auteurs, des revues, des projets, des structures de l'archive ouverte.

L'IdHal est le référentiel auteurs de la base Hal. Un IdHal ne peut être crée que par un auteur, pour lui même. Il est impossible de créer l'IdHal de quelqu'un d'autre, le CCSD considérant que cette initiative reste du domaine, privé, de l'identité numérique du chercheur.

12. À propos du « nouveau » modèle IFLA-LRM, « successeur » des modèles FRBR, FRAD et FRSAD, voir les articles de P. Le Pape [10] et [9] et `https://www.ifla.org/publications/node/11412`.

13. Dans Boulet 2017 [4].

14. Pour des détails sur l'utilisation d'IdRef à l'Adum voir dans Morales 2017 [14].

Le référentiel, permet, sous un même identifiant, de rassembler des formes auteur existantes dans la base (sans forcément les fusionner), selon le même principe que les formes retenues et rejetées des bases d'autorités.

L'IdHal permet d'établir des liens avec d'autres référentiels. L'auteur peut ajouter ses autres identifiants : ArXiv, IdRef, Isni, Viaf, Orcid,... et ses identifiants de réseaux sociaux, blog...

Il permet enfin la création d'un CV du chercheur en affichant la liste des publications déposées dans Hal, les co-auteurs, les domaines, et les liens vers les autres bases.

2.1.2. *OrcId : Open Reseacher and Contributor.* Orcid est un référentiel international de personnes physiques qui publient ou collaborent dans les domaines de l'enseignement supérieur et de la recherche. Il est :

– non lucratif
– transdisciplinaire
– financé par ses membres (institutions, grands éditeurs...)

Il dispose de beaucoup d'adhérants et est utilisé pour identifier les chercheurs dans de nombreux systèmes d'informations. Certains organismes exigent un identifiant Orcid pour déposer des demandes de financement. On peut donc utiliser le référentiel Orcid dans son propre système.

Les Orcid sont crées sur la base du volontariat par les intéressés eux-mêmes. On peut ajouter ses publications, soit en important un fichier BibTEX, soit à la main, soit via les liens vers des bases existantes, en entrant l'Id de la publication (Scopus, BASE, MLA, PubMed...). Crossref met également à jour les données bibliographiques des chercheurs grâce aux DOI.

Un partenariat est en cours entre Orcid et l'Abes, pour que les données des thèses soient versées dans Orcid.

L'un des problèmes est que les données le base ne sont pas administrées par un humain. Il n'y a pas de contrôle ni de nettoyage des données. C'est une démarche volontaire du créateur. Des erreurs sont donc possibles. Elles peuvent avoir des effets indésirables une fois disséminées (doublons, homonymes,...). On trouve ainsi de nombreux Id pour lesquels seul un prénom est renseigné, sans nom de famille.

2.1.3. *Researcher ID.* C'est l'identifiant de Thomson Reuters. Il sert de référentiel au Web of Science, et dans EndNote.

Il ne concerne pas cette présentation, car il est peu connecté (et connectable) avec les SI des laboratoires et des bibliothèques.

2.1.4. *Viaf.* Virtual International Authority File

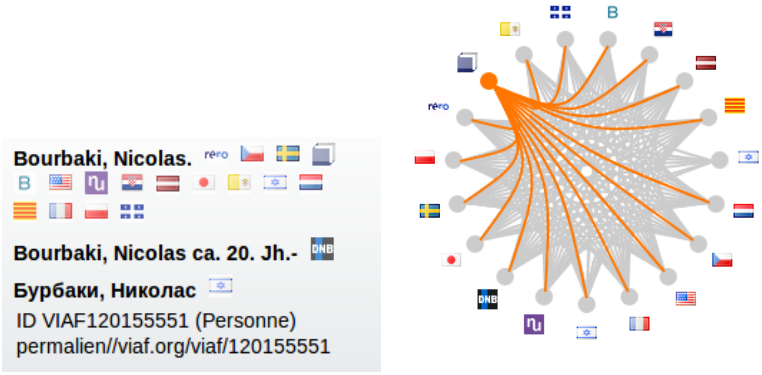
C'est une base d'autorité internationale. Viaf connecte des bases d'autorités nationales entre elles. Elle permet, notamment, de lier les noms ayant plusieurs graphies, plusieurs formes (fig. 2a).

Initialement, Viaf était limité aux personnes et collectivités. Puis le fichier s'est étendu aux œuvres, expressions et noms géographiques.

Viaf génère ainsi des grappes d'autorités. Par exemple :

– `http://www.viaf.org/viaf/68948357/#Demailly,_Jean-Pierre,_1957-.....`
– `https://viaf.org/viaf/120155551/#Bourbaki,_Nicolas.`

Viaf permet, à ce niveau de faire le lien entre les notices d'autorités des différentes institutions. Par exemple, on peut passer de la notice d'autorité du Sudoc à celle de la BN d'Allemagne, comme illustré par la figure 2b. Si un lien est fait avec Wikidata, on peut se rendre directement sur la page en question[15].



(A) Grappe d'autorités (B) Liens entre les sources

FIGURE 2. La notice de Bourbaki dans Viaf (sur `http://viaf.org/viaf/120155551/`)

En effet, Viaf est aussi connecté aux bases comme Perseus (hébergé par l'Université Tufts [15]) qui l'utilise pour sa gestion d'autorité tout comme le le Syriac reference Portal [16] [2].

Viaf sait parler aux machines. Il permet la négociation de contenus. Par exemple :

- En XML Viaf `http://viaf.org/viaf/52358786/viaf.xml`
- En RDF `http://viaf.org/viaf/52358786/rdf.xml`
- En Json `http://viaf.org/viaf/52358786/viaf.json`

Viaf joue donc un rôle essentiel car il permet l'échange et le lien entre des données issues de différentes sources et contrôlées. C'est un élément essentiel du web de donnée.

2.1.5. *Isni.* International Standard Name Identifier.

C'est un référentiel international des personnes et des organismes. Son objectif est l'identification univoque et sur le long terme dans un environnement numérique. Contrairement à Orcid il n'est pas limité à l'enseignement supérieur, et contient tous les auteurs ayant publié un document.

Il repose sur le concept d'identité publique, où une identité devient une notice Isni.

Contrairement à Viaf il n'est pas utilisé seulement par les structures documentaires, mais par toute institution qui a besoin d'identifier de manière univoque une personne ou une structure. Insi est utilisé par Orcid pour identifier les structures d'appartenance des chercheurs.

Les données sont disponibles en RDF en ajoutant .rdf à l'identifiant unique. Ex : `http://www.isni.org/isni/0000000002392694.rdf`

Pour Angjeli [2], Viaf est la « cours de récréation de données hétéroclites et Isni est le surveillant » ! Quelques différences entre les deux systèmes :

- Isni dispose d'un organe de surveillance et de contrôle permanent. Une équipe (BNF-BL) est chargée du contrôle qualité. Viaf est piloté par un conseil composé de ses contributeurs.
- Isni inclut des données privées.
- Viaf crée des grappes à partir de données de bibliothèques. Isni crée des liens avec des bases extérieures et attribue un identifiant ISO.

15. Disponible sur `http://www.perseus.tufts.edu/`.
16. Disponible sur `http://www.syriaca.org`.

2.1.6. *Idref.*
Aux origines de IdRef.

Idref est né de la diversité des projets de l'Abes [17]. La multiplication des bases gérées par l'agence a fait naître le besoin de mutualiser la gestion des autorités pour ses différents catalogues. Le référentiel IdRef ouvre alors en 2011. Il s'agit d'un outil extérieur et autonome par rapport aux applications comme le Sudoc, Theses.fr ou Calames.
Un référentiel interopérable.

Idref a été pensé pour être ouvert et interopérable. Philosophiquement, il a été développé dans un esprit d'ouverture vers l'extérieur. L'interface Web permet un accès simple à la production et à la recherche des données. Les partenaires peuvent écrire, modifier, enrichir les données du référentiel. Techniquement, il utilise tous les standards du Web (RDF natif, négociation de contenu, Solr, nombreuses API et webservices) qui permettent une réutilisation simple et une intégration dans son propre système. Juridiquement, il est placé sous licence Etalab qui permet la réutilisation des données.

Le projet de référentiel ouvert a fonctionné. Les partenaires sont maintenant nombreux (le CCSD, Okina, Persée, le Larhra, mais aussi Viaf, Isni, Orcid...) et les ressources accessibles aussi. Ainsi, dans les notices IdRef, on peut consulter les données de différentes bases. Par exemple : `https://www.idref.fr/02942349X`.
Des outils pour la qualité.

Plusieurs outils ont été développé pour aider les catalogueurs à travailler sur la qualité des notices et de l'ensemble du référentiel [18]. Par exemple :

- AlgoLiens qui permet de détecter les notices du Sudoc non liées
- Algodoublons qui, grâce à la masse de données de Viaf, de détecter des doublons potentiels [19]
- Qualinca (voir 3.2.1 p. 10)

Quelques chiffres. En avril 2017 IdRef comptait 3,4 millions notices d'autorités [20]

- 2,69 millions d'autorités personne physique
- 77000 lieux

Le schéma 3 résume l'ouverture des données d'IdRef vers d'autres systèmes, avec lesquels des alignements ont eu lieu ou sont en cours :

- 2 600 000 notices Viaf liées
- 1 900 000 Isni liés
- 280 000 notices liées dans Theses.fr
- 50 000 notices liées dans Persée
- 750 000 notices liées avec la BnF
- ...

2.2. **Identifier les structures.**

2.2.1. *Le Référentiel national des structures.* Le RNSR est le référentiel national des structures de l'enseignement supérieur et de la recherche. Tous les organismes de recherche sont concernés, quel que soit le ministère auquel ils sont rattachés.

La figure 4 montre la place du RNSR dans l'écosystème français.

17. On trouvera dans Mistral 2017 [13] les détails sur IdRef qui ont servi à l'écriture de ce paragraphe.
18. Des exemples d'utilisation des outils sont données dans Mistral 2017b [12] et sur les pages de documentation de l'Abes `http://documentation.abes.fr`.
19. Il y a doublon quand, par exemple, deux notices d'autorités renvoient à la même personne, au même sujet...
20. Tous les chiffre sont extraits de Mistral 2017a [13].

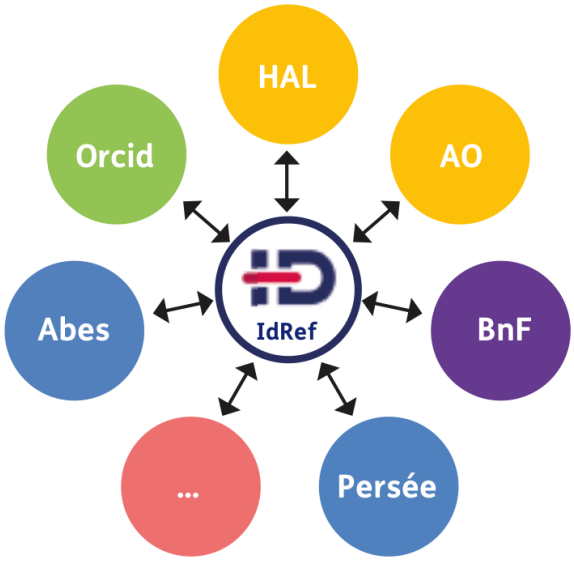


FIGURE 3. Idref, un référentiel interopérable. Source [13]

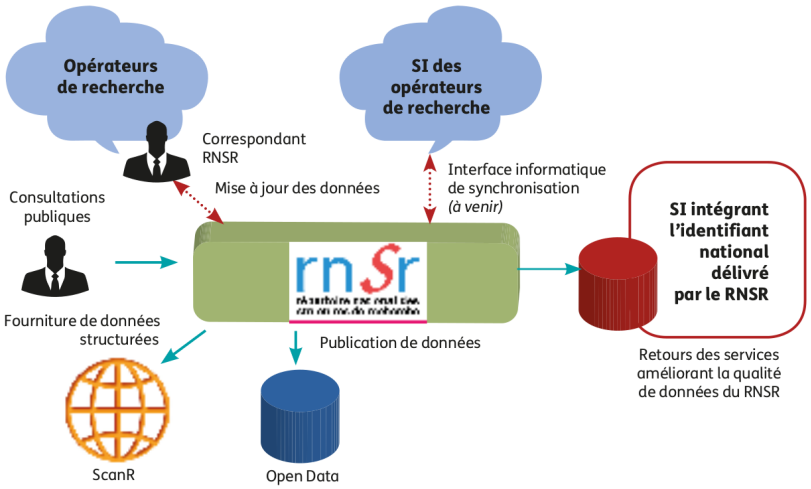


FIGURE 4. Le RNSR et son écosystème. Source [17]

Les données sont librement accessibles en consultation, à l'aide d'une interface le Web. Il n'est possible d'exporter les données aux formats XLS, JSON ou CSV ainsi qu'à l'aide d'une API [21].

Les données sont gérées par des correspondants dans chaque établissement. Une fois authentifiés, les gestionnaires peuvent créer et modifier des structures, rapprocher les structures suite à des fusions, exporter les données en XML, Excel...

Le fiabilité des données dépend donc des mises à jours des correspondants dans les établissements.

2.2.2. *Les structures dans AuréHal.* En plus du référentiel de personnes, (voir 2.1.1 p. 4), AuréHal héberge d'autres listes contrôlées : projets ANR et européens, revues, disciplines, et structures.

21. Code et documentation sur : `https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-repertoire-national-structures-recherche/`.

Tout comme les auteurs, les structures de recherche font l'objet d'un référentiel. Il permet d'identifier les auteurs et leurs appartenances. L'auteur d'un papier déposé dans Hal doit nécessairement être affilié à une structure.

La granularité est fine. Les structures s'entendent du regroupement d'institutions (de type Comue) aux équipes de recherche des laboratoires. Les structures peuvent être reliées entre elles, afin de reconstituer l'organisation administrative de la recherche. Une unité mixte est ainsi liée à ses tutelles. Chaque entité dispose d'un identifiant unique.

En plus de permettre d'extraire les données de la base (par exemple la liste des publication d'un laboratoire, d'un établissement...), le référentiel permet de lier les informations avec celles d'autres référentiels. Les liens sont possibles avec IdRef, Isni et bien entendu le RNSR. On peut donc, par exemple avec IdRef, faire le lien entre les publications dans Hal d'un laboratoire et les thèses soutenues dans ce dernier et qui seraient présentes dans Theses.fr.

Tous les outils évoqués sont donc largement utilisés par les organismes ayant en charge la production et la diffusion de données. Ils répondent chacun à un besoin de qualité, de données de référence. Après avoir présentés ces outils et leurs caractéristiques principales, il convient de voir comment ces derniers sont connectés et ouverts et participent ainsi au partages de données et de l'information scientifique.

3. Les référentiels : ouverture et partage

3.1. **Un Fichier National des Entités.**

L'Abes et la BnF prévoient ensembles, à l'horizon 2020, la constitution d'un fichier mutualisé regroupant toutes leurs autorités... et même plus !

3.1.1. *La notion d'entité.* On ne parle pas ici de référentiel d'autorité. Le périmètre choisi est plus large que les « simples » autorités. On est dans le contexte des données liées.

Le FNE couvrira les éléments suivants :
- Toutes les autorités dans un premier temps
- les lieux
- les expressions
- laps de temps [22]
- structures (intégration du RNSR ?)

3.1.2. *Mutualiser la production des données ?* Pourquoi un fichier national ?

Le FNE sera co-produit par l'Abes et la BnF. La mutualisation de la production devrait favoriser des économies d'échelles. Un seul et même outil devrait ainsi être utilisé par les deux établissements.

Mutualiser les données entre l'Abes et la BnF semble logique. Les deux périmètres sont très complémentaires. L'Abes (notamment grâce au circuit et au dépôt des thèses) dispose d'une masse importante de données concernant l'enseignement supérieur. La BnF couvre l'ensemble de la production éditoriale, notamment grâce au dépôt légal [8]. À terme, l'idée est qu'il s'ouvre davantage. Ainsi, d'autres établissements ayant pour mission de gérer des « collections » – et donc des catalogues ou des bases de données dont le nom peut être différent selon la structure – pourraient être amenés à l'utiliser. Il pourrait bénéficier des enrichissements des bibliothèques territoriales, de services d'archives, de musées...

Il devrait être intérrogeable seulement par les professionnels, sans interface publique [23], mais des API permettront les échanges de données.

22. À propos de la notion de la laps de temps, voir dans Le Pape 2017 [10].
23. Voir Johannic-Seta 2017 [8].

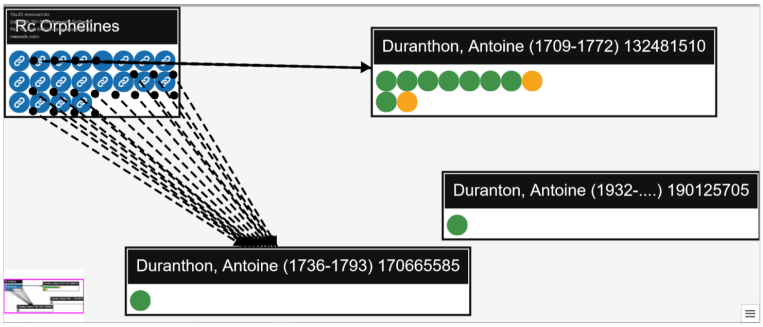


FIGURE 5. Écran de SudoQual : création et vérification de liens. Source [11]

La licence Etalab sera utilisée pour permettre une réutilisation large. L'ouverture est prévue pour 2020 [24].

3.2. **Les référentiels au service de la qualité des données.**

3.2.1. *Au service de la qualité.*

L'expérience Qualinca.

Le projet Qualinca, porté par l'Abes, a pour objectif de favoriser les liens et les associations entre les autorités et les ressources. Le projet est compliqué car il s'agit de le faire en masse, et d'en automatiser le maximum de processus.

Aujourd'hui, Qualinca travaille selon trois axes [25] :

Alignement des référentiels.

C'est le préalable ! Un outil a été développé pour aligner les référentiels avec des entrepôts de données. Par exemple, pour lier en masse, 50 000 auteurs de la base Persée à IdRef.

Travailler sur la qualité des données.

Pour ne pas multiplier les erreurs de liens –toujours possibles dans le travail en masse– l'Abes a développé l'outil SudoQual. Il s'agit d'une série d'algorithmes vérifiant la fiabilité et la pertinence des liens entre Idref et le Sudoc. Il va tenter de repérer les faux liens, les doublons... Il permet aussi la création et la vérification des liens grâce à une interface graphique en cours de développement. La figure 5 montre un écran de vérification des liens Sudoqual.

L'interface graphique sera un outil pour les catalogueurs. L'Abes souhaite ainsi leur faciliter le travail en rendant plus simples, plus rapides et plus fiables les opérations favorisant la qualité des données.

De l'importance du catalogage.

Les documentalistes ont un rôle fondamental à jouer dans les processus de maintient de la qualité des référentiels.

Le rôle des référentiels est aujourd'hui déterminant. Personne ne peut se permettre d'avoir des données de mauvaises qualités ou incomplètes. Une donnée erronée c'est une donnée sur laquelle il faudra revenir, qui créera peut-être des doublons, des liens faux... Si elle est exposée, l'erreur sera propagée à d'autres bases, empêchant, peut-être, l'accès à une ressource documentaire.

Pour garantir cette qualité, les réseaux sont essentiels. La communauté, en travaillant grâce à des outils communs et des pratiques commune se donne les moyens de conserver des bases de grande qualité. La découverte d'une erreur, doit être une occasion de la corriger ou de la signaler.

24. Date dans Johannic-Seta 2017 [8].
25. On trouvera tous les détails sur le projet dans Le Provost 2017 [18] et [11] et dans Nicolas 2017 [16].

Les outils bibliographiques fournis aux chercheurs se basent sur les données produites et contrôlées par les réseaux. Garantir la qualité c'est garantir l'autonomie des chercheurs dans leur travail.

3.2.2. *Au service de la diffusion et de l'enrichissement, du patrimoine.*
Faire des liens ! Le principe est simple : dès que l'on dispose d'un identifiant pérenne pour une ressource, il faut l'ajouter à une base. On ajoute ainsi les DOI, les identifiants Numdam des publications déposées dans Hal, les PPN des auteurs...

L'exemple du projet de reprise de l'institut Fourier. L'idée est de remettre au propre l'intégralité de la base de publication des chercheurs de l'institut Fourier. On utilise pour cela la base Hal comme pivot des données et comme entrepôt de stockage des articles. C'est à partir des données de Hal que l'on extraira, à terme, les listes de publications demandées pour les évaluations. Les vérifications sont à la fois manuelles, automatiques (OcdHal) et individuelles. Les chercheurs créent leur IdHal. Leurs formes auteurs sont ainsi rassemblées. Dans l'IdHal, tous les liens possibles sont faits : Idref, Orcid, Isni, Viaf... Ensuite, chaque publication est vérifiée et complétée, avec le chercheur, et sa liste de publication. On ajoute les publications manquantes, manuellement ou grâce à Bib2Hal. Enfin, s'il dispose d'une notice d'autorité, on ajoute l'IdHal dans IdRef, manuellement, à partir de la nouvelle interface Web d'IdRef. L'objectif est d'aboutir à une base propre et complète, tant pour des raisons administratives que patrimoniales.
Mutualiser pour enrichir : les référentiels au milieu du système d'information. Le but est de construire un écosystème complet, où tout est lié. Aujourd'hui, tout est ressource. Et toutes les ressources doivent être liées. Ainsi, chaque référentiel assure sa part de garantie qualité. Chaque référentiel est au centre des données qu'il expose. Les autres systèmes gravitent à côté de ces données de références. Ils réutilisent les données « valides » et ajoutent leurs informations. Un système externe peut utiliser IdRef pour identifier ses auteurs, en complément de ses propres données. Il s'agit d'établir la correspondance entre l'auteur dans un système tiers et l'auteur dans le référentiel (alignement). L'identifiant auteur de ce système externe peut être ajouté à IdRef. Dans certains cas, des données sur l'auteur peuvent être ajoutées au référentiel (enrichissement).

La figure 6 tente de résumer les liens existants (mais non exhaustifs) entre quelques référentiels et quelques bases et catalogues. On remarque alors à quel point ils sont imbriqués. Les formats ouverts et documentés laissent ainsi la possibilité de réutiliser les données d'un référentiel dans la base voisine et réciproquement.

3.3. **Disséminer, ouvrir : des exemples de réutilisation.** Le numéro 85 d'*Arabesque* détails justement plusieurs projets et outils, utilisant IdRef comme référentiel. Quelques exemples :

- Persée utilise IdRef pour gérer la base des auteurs des documents numérisés [5].
- La base de prosopographie bretonne Prelib utilise IdRef, notamment pour permettre des alignements avec d'autres fourisseurs de données, comme Wikipédia [3].
- Le Larhra [26] a aligné des autorités « Acteurs » de la base SyMoGIH avec IdRef, pour permettre l'ouverture vers d'autres bases et l'enrichissement des données [21].
- Dans l'outil CapLab, IdRef devrait servir de pivot car il est un réel référentiel de chercheurs [20].
- Le référentiel de la production scientifique Conditor va utiliser IdRef pour les mêmes raisons, ainsi que le RNRS pour gérer les stuctures [7].

26. UMR 5190.

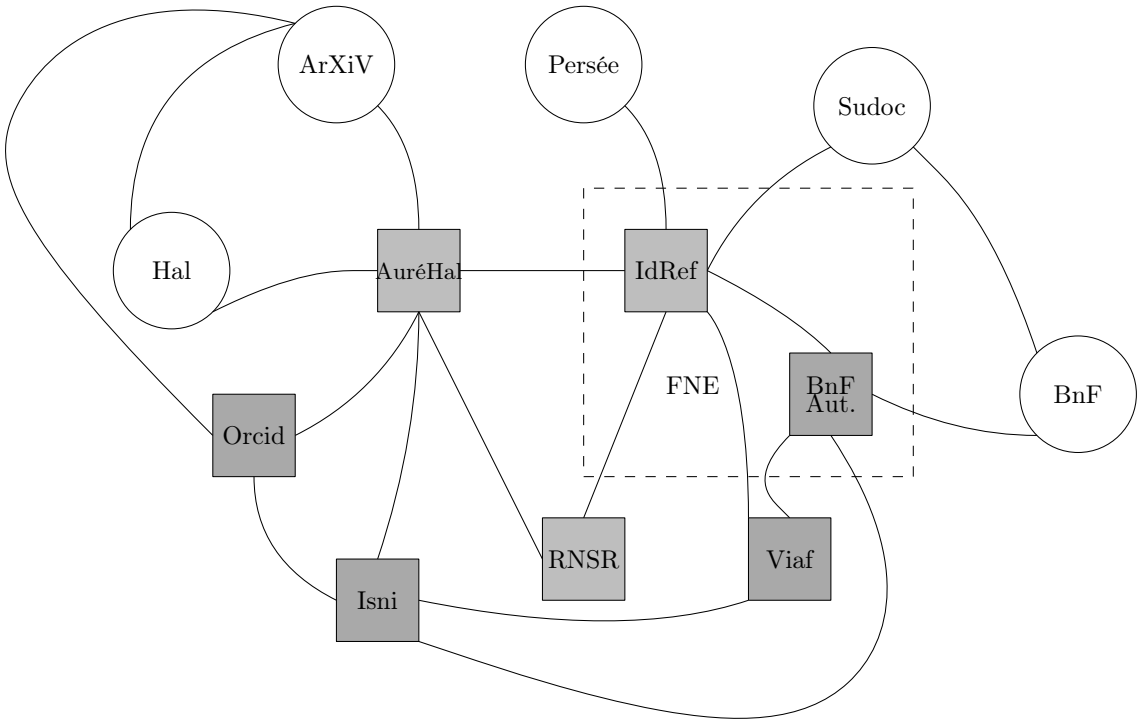


FIGURE 6. Les liens entre quelques référentiels

4. Conclusion

Étant donnée la complexité des systèmes d'informations documentaires actuels, on ne peut plus se passer des référentiels. Toutes les connexions sont possibles grâce aux API, aux standards, aux webservices.

Mais plusieurs questions se posent. Les référentiels renvoient à de nombreuses personnes physiques. Qu'en est-il alors de la question des données personnelles ? Chaque référentiel contrôle un domaine bien spécifique. Mais y a-t-il trop de référentiels ? Aura-t-on les moyens de tous les maintenir à long terme ? Que faut-il faire pour convaincre les institutions qui disposent de bases données de les ouvrir et de les lier ?

Tout ceci va dans le sens d'une ouverture des données et d'un partage des métadonnées entre les bases, et entre les organisations. Faciliter ces échanges, c'est participer à la diffusion de l'information scientifique et à la circulation de la connaissance.

Références

[1] Abes, *Guide méthodologique : Zone 810 : Source consultée avec profit*, Guide méthodologique, 2017, URL `http://documentation.abes.fr/sudoc/formats/unma/zones/810.htm`.

[2] Anila Angjeli, Andrew Mac Ewan, and Vincent Boulet, *SNI et VIAF transforment le paysage : pour des identités fiables et solides*, Research report, IFLA 2014, 2014, URL `http://library.ifla.org/985/7/086-angjeli-fr.pdf`.

[3] Nelly Blanchard, Jean-Baptiste pressac, and Mannaig Thomas, *Aurehal et son idhal rassembleur*, Arabesque **85** (2017), 14.

[4] Vincent Boulet, *Des référentiels et de leur usage aujourd'hui*, Arabesque **85** (2017), 4–5.

[5] Viviane Boulétreau, *Persée, partenaire confirmé*, Arabesque **85** (2017), 17.

[6] Michel Chein, Alain Gutierrez, and Michel Leclère, *Un problème d'identification d'entités nommées dans des bases de donnés documentaires*, Research report, LIRMM, 05 2015, URL `https://hal-lirmm.ccsd.cnrs.fr/lirmm-01187747`.

[7] Annie Coret, *CONDITOR, le futur pot commun de métadonnées*, Arabesque **85** (2017), 12.

[8] Frédérique Johannic-Seta, *Le futur FNE : vers une vraie coproduction*, Arabesque **85** (2017), 16–17.

[9] Phillipe Le Pape, *Vingt ans après : LRM, le cinquième mousquetaire*, Arabesque **87** (2017), 18–19.

[10] ———, *Vingt ans après : LRM, un pour tous !*, Blog RDA, 8 2017, URL https://rda.abes.fr/2017/08/21/vingt-ans-apres-lrm-un-pour-tous/.

[11] Aline Le Provost, *Qualinca : diagnostiquer, améliorer. le signalement augmenté*, Présentation, Journées Abes 2017, 5 2017, URL http://www.abes.fr/Media/Fichiers/Footer/Journees-ABES/Jabes17/Qualinca.

[12] François Mistral, *IdRef, pour des données de qualité*, Présentation, Journées Abes 2017, 5 2017, URL http://www.abes.fr/Media/Fichiers/Footer/Journees-ABES/Jabes17/IdRef-session-parallele.

[13] François Mistral and Yann Nicolas, *IdRef, les autorités en conquête et en partage*, Arabesque **85** (2017), 8–9.

[14] Catherine Morales, *Adum, plus de visibilité pour les doctorants*, Arabesque **85** (2017), 13.

[15] Yann Nicolas, *IdRef dans VIAF et après ...1 passer d'un identifiant à l'autre (VIAF, IdRef, LC, BnF, Wikipedia, ...)*, Punktokomo, 5 2012, URL https://punktokomo.abes.fr/2012/05/11/idref-dans-viaf-identifiants/.

[16] ———, *Aligner, le signalement augmenté*, Présentation, Journées Abes 2017, 5 2017, URL http://www.abes.fr/Media/Fichiers/Footer/Journees-ABES/Jabes17/Qualince-alignements.

[17] Isabelle Pouliquen, Isabelle Kabla-Langlois, and Xiaofeng Chen, *Le RNSR, colonne vertébrale du SI recherche*, Arabesque **85** (2017), 15.

[18] Aline Le Provost, *Qualinca et Idref : l'intégration est en cours !*, Arabesque **85** (2017), 1O.

[19] Olivier Rousseau, François Mistral, Yann Nicolas, and Philippe Le Pape, *Trois fois sur le métier remettons les notions : 3 questions aux experts de l'ABES*, Arabesque **85** (2017), 6–7.

[20] Romain Thouy, *Caplab, pour un suivi de l'activité des unités*, Arabesque **85** (2017), 14.

[21] Pierre Vernus, *Symogih, de l'umr5190 - larhra et les objets historiques*, Arabesque **85** (2017), 14.

[22] Wikipédia, *Référentiel (base de données) — Wikipédia, l'encyclopédie libre*, 2017, [En ligne ; Page disponible le 13-mai-2017], URL \url{http://fr.wikipedia.org/w/index.php?title=R%C3%A9f%C3%A9rentiel_(base_de_donn%C3%A9es)&oldid=137306797}.

Institut Fourier, UMR5582 CNRS, Université Grenoble Alpes, CS 40700, 38058 Grenoble cedex 09, France

E-mail address: romain.vanel@univ-grenoble-alpes.fr

URL: http://vanelro.perso.math.cnrs.fr/