Bilan à mi-2020

Référence : Conditor/Bilan mi 2020

Date : 30/07/2020

Auteurs : équipe Conditor

Sommaire

1. Pourquoi Conditor ?

Il existe une multitude de dispositifs techniques et organisationnels dans lesquels la production scientifique d'un chercheur, d'une équipe, d'un laboratoire, d'un établissement est décrite, par différents acteurs : une archive nationale HAL, des archives ouvertes et des bases bibliométriques institutionnelles, un dispositif opéré par l'OST (Iperu[1]), des Systèmes d'Informations (SI) « recherche », des bases de signalements et des archives internationales...
Pourtant nul n'a une vision complète de la production des laboratoires et établissements de l'ESR (Enseignement Supérieur et Recherche).
Conditor est né de ce constat pour contribuer à organiser la mutualisation et limiter les multiples travaux de recensement et de saisie.

Conditor vise à recenser la production scientifique des établissements et laboratoires relevant de l'ESR français en s'appuyant sur des sources existantes et à fournir en métadonnées des dispositifs de l'ESR identifiés.

C'est un service complémentaire des dispositifs existants, important à la fois pour la science ouverte et la cohérence entre SI, qui :

- capitalise le résultat du travail fait par tous pour alléger ensuite le travail fait par chacun,
- contribue à l'exhaustivité de la couverture de l'archive nationale HAL et des archives institutionnelles ainsi qu'à la qualité et la complétude des données bibliographiques,
- simplifie les processus de gestion de la recherche (évaluation, demandes de financement...),
- s'appuie sur des référentiels ou autorités reconnus collectivement.

Conditor est un « glaneur/unificateur » de métadonnées sources décrivant la production de l'ESR mettant à disposition dans un format unique des métadonnées qualifiées[2] réutilisables par tous les acteurs de l'ESR pour différents usages.

2. Quels étaient les objectifs de la phase projet ?

Après une expérimentation ayant démontré la faisabilité, l'intérêt et le potentiel de Conditor, la phase projet a été lancée en décembre 2016 avec le soutien du Mesri, l'objectif étant de mettre en place :

- un réservoir significatif de métadonnées décrivant la production scientifique, à partir d'un premier ensemble de sources identifiées, utilisables techniquement et juridiquement,
- un « outillage » aussi industrialisé que possible pour la collecte au fil de l'eau, les traitements (reformatage, détection de doublons, attribution d'un code RNSR à chaque adresse/affiliation française...), l'intégration d'enrichissements provenant de services partenaires, la diffusion des métadonnées aux acteurs de l'ESR identifiés,
- l'organisation multipartenaire pour le faire vivre, notamment un premier réseau métier réparti dans un ensemble d'établissements pour la validation du résultat incertain de traitements automatiques.

mais aussi d'engager le processus de mutualisation et d'intégration progressive dans les différents dispositifs de l'ESR.

3. Où en est-on ?

Grâce à l'équipe multipartenaire regroupant des professionnels de l'IST, gestionnaires d'archives, informaticiens... au sein d'organismes de recherche, universités et entités ESR travaillant sur des sujets connexes, et à l'implication de l'Inist-CNRS, un premier service opérationnel, présenté lors de la réunion du 6 février regroupant les participants au projet et collègues intéressés, est en place.
On trouvera en annexe la liste des partenaires et l'organisation du projet.

[1] Interface de repérage des publications : organisation et outil mis en place par l'OST de l'Hcéres afin de comptabiliser la production de chaque établissement dans le WoS et produire analyses et indicateurs.
[2] D'où viennent-elles ? quels traitements effectués ? qui est intervenu ? quelle est leur fiabilité ?

3.1 Les corpus sources aujourd'hui et demain

Le choix des sources est un sujet complexe.
Plusieurs critères de sélection des sources sont à prendre en compte : disponibilité technique, juridique, qualité des métadonnées, critères d'identification des signalements pertinents dans ces sources, complémentarité entre sources, apport spécifique[3], etc.
Le groupe en charge de la stratégie de construction du corpus de signalements[4] a commencé par travailler sur des corpus sources connus et disponibles sur l'année 2014 (HAL, WoS, ProdInra et des thèses du Sudoc) pour définir un format pivot.
Ce format pivot est une extension de la TEI HAL afin de faciliter l'articulation avec l'archive nationale : il comprend des métadonnées bibliographiques mais aussi de gestion pour assurer la traçabilité de toute métadonnée.

En attendant l'élaboration des feuilles de style XSLT nécessaires à l'automatisation des ingestions au fil de l'eau, des programmes de reformatage en masse de quelques corpus ont tout d'abord été réalisés afin de disposer de données pour mettre en place la plateforme applicative et les premiers modules.

Aujourd'hui des feuilles de styles ont été développées pour 4 sources : HAL, PubMed, les thèses du Sudoc, CrossRef. L'élaboration de la feuille de style pour les ouvrages du Sudoc est en cours.

D'autres sources sont actuellement à l'étude afin de compléter la couverture :

- le réservoir de métadonnées constitué pour le BSO,
- des sources complémentaires en SHS, les ouvrages du Sudoc.

En fait, l'un des problèmes à résoudre, pour toute nouvelle source, est notamment d'identifier les signalements pertinents en l'absence d'affiliation[5].

Il faut également noter que dans la mesure où aucun format d'entrée n'est imposé, l'introduction de tout nouveau corpus nécessite après sélection, un investissement non négligeable pour l'étude, les spécifications et la création des feuilles de styles XSLT. Une tâche complexe nécessitant des compétences spécifiques peu répandues[6].

3.2 Une plateforme applicative évolutive et ouverte

Le groupe en charge de la conception/développement itératif/déploiement de la plateforme applicative[7] a tout d'abord étudié les solutions techniques possibles et s'est appuyé sur les principes suivants :

- Utiliser la méthode Agile Scrum[8],
- Se baser sur des briques du projet ISTEX pour le développement de la plateforme applicative,
- Développer de façon modulaire et sous licence CeCILL[9],
- Héberger la plateforme sur des serveurs Inist[10].

A ce jour, grâce à l'équipe Inist, associant informaticiens et acteurs métier mise en place, et à la contribution de tous les collègues au sein des établissements, il a été réalisé :

- 4 corpus sources (HAL, CrossRef, PubMed et les thèses du Sudoc) ont été collectés, reformatés et ingérés.

[3] Par exemple : signalements spécifiques à un domaine scientifique, métadonnées particulières...
[4] Lot 3 copiloté par François Mistral de l'Abes et Christiane Stock de l'Inist
[5] L'affiliation facilite grandement le repérage de travaux de recherche effectués dans un établissement ou un laboratoire de l'ESR ou du moins en « France ».
[6] Ces feuilles de style ont été élaborées par Catherine Morel dans un premier temps puis Christiane Stock et Stéphanie Gregorio toutes les trois de l'Inist.
[7] Lot4 copiloté par Claude Niederlender jusqu'en août 2019 puis Pascal Cuxac de l'Inist et Yannick Barborini du CCSD jusqu'au début 2018.
[8] Ce qui permet d'ajuster la feuille de route des développements au fil de l'eau, en fonction des retours de l'équipe projet.
[9] Cela permet l'éventuelle appropriation par les partenaires et pourquoi pas une future réutilisation ou des améliorations par d'autres personnes ou projets que Conditor. Il est hébergé sur https://github.com/conditor-project/
[10] D'autres solutions ont été envisagées : offre cloud du CNRS, hébergement CCSD/In2P3, futur SI recherche

- La détection de doublons « certains » via des règles documentaires strictes est en production.
- Le repérage de doublons « incertains » s'appuyant sur des règles de similitude est opérationnelle :
    - L'interface Cornelius de validation des doublons « incertains » a été testée et utilisée par le groupe de préfiguration du réseau métier.
    - Des améliorations sont envisagées pour alléger la tâche de validation.
- Un module d'alignement « certain » avec le RNSR a été développé (en dehors de la plateforme Conditor) mais n'a pas pu être exploité à ce jour.
- Différentes études sont en cours pour réaliser des alignements incertains avec le RNSR[11] qui nécessiteront une validation par les membres du réseau métier.
- Des tests d'intégration du résultat d'enrichissements effectués hors plateforme Conditor par un service partenaire (alignement avec référentiel auteur par l'Abes notamment) ont été effectués avec succès.
- Un mécanisme de génération à la volée d'une « notice de référence » à partir de signalements sources distincts décrivant une même production a été conçu, développé et testé avec succès :
    - Il s'appuie sur un ensemble de règles documentaires de sélection des métadonnées qui seront à terme personnalisables par un partenaire consommateur de données, pour mieux répondre à ses besoins.
- Outre Cornelius, 2 autres interfaces ont été mises en place et sont utilisées en production par des collègues au profil administrateur de données :
    - Concerto pour lancer et suivre les traitements opérés par les différentes briques logicielles,
    - Kibana pour la visualisation du contenu du réservoir de métadonnées avec des tableaux de bord paramétrables.

Au total :

- 1 430 256 signalements issus des années de publications sont intégrés :
    - 2014-2019 de Hal et PubMed,
    - 2014-2017 de CrossRef et Sudoc thèses.
- Ils décrivent a minima 1 108 307 productions distinctes.
- 145 011 signalements ont un ou plusieurs doublons incertains.

Parmi les développements restant à faire :

- l'alignement incertain avec le RNSR et l'interface de validation,
- la gestion des accès via la fédération d'identités.

3.3 Un processus de mutualisation et d'intégration dans les SI de l'ESR engagé

Conditor est un glaneur/unificateur/distributeur de données de qualité, permettant à un partenaire :

- d'éviter de refaire ce qui est fait ailleurs par d'autres acteurs de l'ESR,
- de faire bénéficier tout l'ESR de ce qu'il produit.

Les SI de l'ESR peuvent dans ce cadre être fournisseurs et/ou consommateurs.

[11] Des méthodes par apprentissage automatique non supervisé couplées à des méthodes de découpage d'adresses permettant l'identification, entre autre, de la ville et du pays. Ces méthodes sont à base de plongements lexicaux ('words embeddings', cf https://fr.wikipedia.org/wiki/Word_embedding). Une approche par apprentissage supervisé avec des corpus d'affiliations alignées avec le RNSR sera également testée

L'API sécurisée, accessible uniquement à des établissements et entités de l'ESR bien identifiés, qui a été développée, permet des recherches multicritères et la récupération de métadonnées.
Elle est notamment utilisée dans une version pilote de Caplab et de façon expérimentale dans l'appel à projet générique 2020 lancé par l'ANR ainsi que par l'université de Limoges. Elle est testée par l'université de la Réunion et l'ENPC.

Un SI ESR fournisseur peut récupérer :

- des enrichissements provenant d'autres fournisseurs ou issus d'alignements avec des référentiels, qu'ils soient réalisés dans Conditor ou par un service partenaire (exemple des propositions d'identifiants auteurs par l'Abes),
- des signalements de productions potentiellement pertinents non présents .

Un SI recherche peut récupérer et afficher une liste de productions potentiellement pertinente pour une unité ou un auteur, par exemple. Les personnels des unités pourront ainsi voir Conditor au travers de leurs applicatifs de gestion habituel et leur « retour » permettra d'enrichir ou améliorer Conditor.

Le fait d'utiliser le RNSR est aussi un facteur de mutualisation supplémentaire : pour une UMR donnée ayant 2 tutelles par exemple (Angers et Strasbourg), il « suffira » que le signalement soit présent dans l'archive d'Angers et versé dans Conditor pour que Strasbourg puisse le récupérer via l'API Conditor.

Concernant HAL en particulier, Conditor permet de :

- détecter des productions non décrites dans HAL, les doublons entre HAL et une source, préalablement à une importation "en masse" ou au fil de l'eau, des doublons intra HAL certains et incertains,
- fournir des enrichissements provenant d'autres sources que HAL ou d'alignements faits dans Conditor ou via un service partenaire.

HAL pourra assurer :

- la validation des doublons incertains, le résultat étant transmis à Conditor ensuite,
- le dédoublonnage (ou fusion) des notices HAL en doublons,
- l'enrichissement de notices HAL avec des données provenant d'autres sources.

Un travail en commun CCSD/Inist est lancé pour éviter tout redéveloppement inutile, l'objectif général étant cependant de faciliter autant que faire se peut le dépôt dans HAL par les chercheurs.

3.4 Une organisation du service Conditor à consolider

Le groupe en charge de la réflexion sur l'organisation du service[12] a défini les grandes lignes de l'organisation multipartenaire du service :

- Une structure de pilotage opérationnel (évolutions et organisation).
- Une infrastructure et une équipe informatiques (hébergement, maintien en condition opérationnelle et maintenance évolutive de la plateforme applicative).
- Un premier réseau métier constitué de gestionnaires de données (validation des doublons et alignements incertains) répartis dans les établissements partenaires.

Un groupe de préfiguration du réseau métier impliquant des collègues des universités d'Angers, Lorraine, Montpellier, Nice, Paris Diderot, Strasbourg, du CNRS, Inria a été constitué pour associer les collègues concernés au paramétrage des interfaces et à la réflexion concernant le fonctionnement du travail collaboratif.

[12] Lot 5 piloté par Nathalie Reymonet de l'université Paris Diderot jusqu'en septembre 2018 puis copiloté par Maxence Larrieu de l'université d'Angers jusqu'en janvier 2019 et Frédérique Flamerie de l'université de Bordeaux

Un wiki, entièrement ouvert et régulièrement enrichi, met à disposition toutes les documentations utiles pour les membres du réseau métier ou toute personne intéressée par le projet (algorithme de similarité pour le repérage de doublons incertains, guide pour l'utilisation de l'interface Cornelius et bonnes pratiques pour le travail collaboratif de validation, API pour les informaticiens, API pour les non spécialistes), une FAQ.

4. Annexe : partenaires et organisation du projet

4.1 Les établissements et entités partenaires

- des organismes de recherche : CNRS (DAPP, Dist, Inist, InSHS), Inra et Irstea puis Inrae, Inria, IRD,
- des universités : Universités d'Angers, Bordeaux, Grenoble (juin 2017), Paris Dauphine, Lorraine, Montpellier, Nice, Paris Diderot, Sorbonne Université....
- des opérateurs et entités de l'ESR : Abes, Amue, CCSD, Hcéres (OST, DSI), Huma-Num, MESRI RNSR-ScanR.

4.2 L'organisation du projet

Les travaux à effectuer ont été regroupés en six « lots » menés par des groupes multipartenaires pilotés ou par un ou deux d'entre eux :

Lot	Objet du lot	Pilotes
Lot 1	Gestion de projet	CNRS (Dist)
Lot 2	Négociation des sources et formalisation des usages des données collectées et produites	Irstea puis IAVFF Agreenium[13] CNRS (Dist Inist)
Lot 3	Stratégie de construction et constitution « itérative » d'un corpus de signalements	Abes CNRS (Inist)
Lot 4	Conception / développement itératif / déploiement de l'applicatif	CNRS (Inist) *CCSD*[14]
Lot 5	Mise en place du service opérationnel	Université Paris Diderot[15] Angers[16], Bordeaux
Lot 6	Communication	Université de Bordeaux

Le comité de suivi opérationnel (cosop) assure la coordination d'ensemble des travaux : y participent les pilotes et/ou copilotes des lots et des acteurs clés.

Le comité de pilotage veille au bon déroulement du projet et à l'harmonisation de la position des acteurs.

L'équipe projet comprend une cinquantaine de collègues participant aux travaux d'un ou plusieurs lots auxquels s'ajoutent les membres du groupe de préfiguration du réseau métier.

[13] jusqu'en novembre 2018
[14] jusqu'en mars 2018
[15] jusqu'en mai 2018
[16] jusqu'en janvier 2019