Compte-rendu de la réunion de fin d'étape de la phase projet tenue le 6 février 2020

Un service opérationnel de référencement en 2020

L'après-midi a été introduite par **Joanna Janik***[1], CNRS Dist, qui a rappelé les objectifs de cette réunion rassemblant les participants à la phase projet, des premiers usagers et des collègues intéressés : partager l'avancement des travaux, présenter des exemples d'usages existants ou prévus et évoquer des pistes pour l'avenir.

1 Pourquoi Conditor ?

Serge Bauin*, CNRS-Dist, *un des initiateurs de Conditor*, a retracé la genèse du projet inscrit à l'origine dans la feuille de route du programme BSN (Bibliothèque Scientifique Numérique) du ministère de la recherche afin de créer une bibliographie de la production scientifique de la recherche publique française.
Un comité de pilotage présidé par Françoise Thibault alors pilote du programme BSN a été mis en place et un groupe comprenant les copilotes du groupe BSN3 portant sur le signalement (Serge Bauin et Raymond Bérard alors respectivement directeur de la Dist - CNRS et de l'Abes), Dominique Cavet (IRD), Francis André (CNRS Dist) et Alain Zasadzinski (CNRS Inist) a été constitué courant 2013 pour démontrer la faisabilité de Conditor (le dieu de l'entreposage du grain dans les granges dans la mythologie romaine).

2 Conditor aujourd'hui

Jérôme Poumeyrol*, Université de Bordeaux, *responsable du service "Soutien et services à la recherche" de la direction de la documentation,* a animé cette session et précisé en introduction qu'elle était destinée à rappeler les principes de construction, présenter le service pilote en place et illustrer le travail à mener par les membres du réseau métier.

2.1 Une « mise en commun »

par **Annie Coret***, CNRS Dist, *coordinatrice du projet Conditor*

Conditor a pour objectif de recenser la production des unités et établissements de l'ESR et fournir en métadonnées les dispositifs de l'ESR identifiés, avec l'objectif à terme de limiter les multiples travaux de recensement et de saisie.
De fait, il existe une multitude de dispositifs techniques et organisationnels de recensement, au niveau individuel, d'une équipe, d'une unité, d'un établissement, etc. d'où l'idée de "mettre en commun" (données, savoir-faire...) : bénéficier des travaux faits par chacun pour constituer un pot commun de métadonnées de qualité qui bénéficiera à tous**.**

Pour y parvenir, il était nécessaire de regrouper des organismes de recherche, des universités et des entités travaillant sur des sujets connexes.

[1] L'astérisque signale l'appartenance à l'équipe projet Conditor ou au groupe de préfiguration du réseau métier.

Le principe de construction est :
- d'utiliser des sources existantes ESR et non ESR ouvertes et de s'appuyer sur des référentiels communs (RNSR...),
- de développer l' « outillage » nécessaire à la collecte, la détection de doublons intra corpus et inter corpus, des enrichissements (ajout d'un code RNSR à toute adresse source notamment),
- d'intégrer le résultat de traitements effectués par des services partenaires à chaque fois que possible (proposition d'identifiants d'auteurs par exemple par l'Abes).

Aucune saisie ou correction de métadonnées sources n'est effectuée dans Conditor. Des alertes vers les dispositifs fournisseurs de l'ESR pourront être effectuées quand des anomalies seront détectées, afin que les corrections soient faites dans la source.
Le résultat incertain de traitements automatiques requiert par contre une intervention humaine pour valider ou invalider et nécessite la mise en place d'un réseau métier.

Le pot commun de métadonnées « qualifiées » [2] ainsi constitué est interrogé par des applicatifs ESR pour récupérer des métadonnées.
Une source ESR pourra récupérer par exemple des enrichissements provenant d'autres sources ou issus d'alignements avec des référentiels ainsi que des signalements de productions manquants dans la source.
Un SI (Système d'Information) recherche pourra récupérer la liste de productions d'une unité par exemple [3].

Un « retour » des dispositifs bénéficiaires est attendu pour enrichir ou améliorer Conditor, par exemple : la validation de l'affiliation d'une production à un établissement ou le signalement d'une référence manquante.

2.2 Un service pilote

par **Valérie Bonvallot***, CNRS-Inist, *responsable fonctionnel de la plateforme Conditor*

Actuellement les années 2014 à 2017 de Hal, Pubmed, Crossref, Sudoc sont chargées : des années plus récentes et une antériorité seront traitées et d'autres sources seront utilisées (en SHS notamment).
Toutes les sources sont mises au format TEI Conditor en sachant qu'aucun format d'entrée n'est imposé.
La détection des doublons certains via des règles documentaires strictes et la détection des doublons incertains via un algorithme de similarité sont opérationnels.
L'alignement certain avec le RNSR a été développé mais le résultat n'est pas encore intégré à la plateforme.
L'API sécurisée permet aux dispositifs bénéficiaires de récupérer des métadonnées via des recherches multicritères.
Deux outils sont à la disposition des gestionnaires de données, membres du réseau métier :
- Cornelius, une interface de validation des doublons incertains,
- une interface Kibana permettant de visualiser le contenu et d'accéder à des tableaux de bord paramétrables.

Un wiki est disponible en accès libre. Destiné à tous, il permet notamment d'accéder à des présentations, un glossaire, des documentations (algorithme de détection de doublons incertains et consignes de validation, modalités d'utilisation de l'API...), une FAQ.

Aujourd'hui, les doublons (ou signalements décrivant une même production) sont liés entre eux dans le réservoir Conditor. Il est prévu d'aller au-delà et de constituer un signalement de référence ou notice chapeau comprenant une sélection de champs communs, des champs spécifiques et des enrichissements réalisés dans Conditor.

Au 29 janvier 2020 (source Kibana):
- 1 084 798 signalements sources ont été ingérés,
- correspondant à au moins 796 229 productions distinctes.
- 123 212 signalements ont un ou plusieurs doublons incertains.

2.3 Un réseau métier pour gérer les données

par **Frédérique Flamerie***, Université de Bordeaux, *chargée de mission science ouverte / données de recherche et animatrice du groupe en charge de l'organisation du service opérationnel Conditor*

Un réseau métier est nécessaire pour effectuer une validation lorsque Conditor ne parvient pas à décider automatiquement.

[2] L'origine de toute donnée ou tout traitement effectué dans Conditor est tracée.
[3] Les personnels des unités verront Conditor au travers de leurs applicatifs habituels.

Cela concerne :
- les doublons incertains,
- les alignements RNSR incertains (à venir).

Un groupe de préfiguration a été lancé mi-2018 pour associer les collègues au paramétrage des interfaces et à la réflexion concernant le fonctionnement du travail collaboratif.
Ce groupe a testé Cornélius, l'interface web dédiée à la validation des doublons incertains de Conditor.

2.4 Discussion et commentaire

Charge de travail pour valider les doublons incertains
Le nombre de doublons incertains est de l'ordre de 120 000 avec 4 sources sur 4 années, ce qui se correspond à 60 000 validations à effectuer au maximum ; ce nombre ne prend toutefois pas en compte de récentes améliorations concernant les thèses, qui le feront diminuer.
Il est de fait difficile d'évaluer la charge de travail, car elle dépend de plusieurs facteurs tels que le paramétrage de l'algorithme de similarité, la qualité des données source, les sources prises en compte, etc. On peut anticiper que chaque établissement travaillera prioritairement sur les publications de ses laboratoires et chercheurs, et les facteurs précédemment cités auront un poids différent en fonction des cas.

Comparaison du corpus obtenu avec des sources payantes et non payantes
L'expérimentation, menée en 2013-2014, sur l'année 2011, avait conduit à un corpus de 160 000 signalements approximativement, dont la moitié environ était issue d'une source payante mais lors de la phase projet, cette comparaison n'a pas été faite.

Utilisation des métadonnées fournies par Elsevier en licence open pour des usages ESR
Le signalement exhaustif de la production scientifique française dans les revues d'Elsevier est prévu dans la licence nationale signée de 2019 à 2022. Elle permettra la récupération dès la publication des métadonnées de tous les articles avec un auteur affilié à une institution française. Elle pourra être une source de Conditor.

En conclusion de la session, **Jérôme Poumeyrol*** souligne l'apport spécifique du projet allant au-delà de la problématique de Conditor : faire travailler ensemble des structures qui n'en avaient pas l'habitude.

3 Les possibles usages des données de Conditor

Diane Le Hénaff*, Inraé, *chargée de mission sciences participatives au sein de la Direction pour la Science Ouverte* a animé cette session et prévu les présentations par type d'applicatifs (archives ouvertes, CRIS, Caplab et exemple de services ANR).

3.1 Alimenter une archive ouverte avec Conditor : les besoins de l'Université de Strasbourg

Par **Adeline Régé***, Unistra, *Responsable du Pôle Appui à la diffusion de la recherche*.

L'archive de l'Unistra est connectée à Hal depuis 2018.
Depuis 1er janvier 2020, le dépôt est obligatoire dans l'archive ouverte de l'université.
Conditor permettra :
- une fiabilisation des données en bénéficiant des enrichissements de Conditor,
- l'exhaustivité du fait du repérage dans d'autres sources de signalements (PubMed par exemple),
- une simplification du dépôt pour le rétrospectif : si les métadonnées sont fournies par Conditor, les chercheurs seront sollicités pour avoir le texte intégral.

Cela pose la question du positionnement des sources, où déposer, voir dans quoi pousser (tout dans Hal ? pas sûr).

3.2 HAL, dédoublonnage et curation avec Conditor

Par **Bénédicte Kuntziger***, CCSD

Conditor permettra :
- d'alimenter une partie référence bibliographique de Hal,
- de détecter des doublons intra Hal certains et incertains,
- d'enrichir des notices Hal.

Il est prévu de faire évoluer Hal pour assurer :

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

Les usagers de Hal ne devront pas avoir besoin de se connecter à Conditor pour effectuer ces opérations.

3.3 Modernisation du référencement des publications à l'EHESS et mise en place d'un CRIS

Par Joachim Dornbusch, EHESS, *responsable du Pôle numérique recherche de la DSI de l'EHESS*.

L'EHESS souhaite mettre en place un CRIS et a choisi Vivo.
Actuellement, la partie Production sera alimentée par Hal en attendant Conditor.

Un alignement des IdRef sur les IdHal des chercheurs de l'EHESS a été réalisé par l'Abes. Malheureusement peu de publications sont retrouvées par ce biais. Une autre recherche sur les noms d'auteurs a été réalisée. Les données Hal ont été intégrées dans Vivo qui permet de visualiser les données sous forme de réseau.

L'utilisation de la base documentaire Francis pour alimenter Conditor permettra de disposer d'une antériorité intéressante pour l'EHESS puisqu'il s'agit de sources SHS difficiles à repérer par ailleurs.

3.4 Système de pilotage de la recherche à l'Université de Limoges

Par **Damien Rieu***, Université de Limoges, membre de la DSI, projet Yuzu

Le projet Yuzu dédié au pilotage recherche consiste à recueillir des informations à partir de différentes sources de données, à stocker ses informations dans un entrepôt puis à produire des indicateurs et rapports pour la recherche. La partie RH est en phase de test avancé.
Pour la partie Publication, le fait que Conditor récupère des données de plusieurs sources et les met à disposition dans un format unique via l'API est intéressant pour le projet Yuzu. C'est pourquoi, l'API Conditor a été testée. Les données, récupérées à partir des identifiants RNSR présents, ont été chargées dans la plateforme : il reste à préciser avec les utilisateurs le contenu des rapports souhaités pour la partie Publication.

3.5 Les usages de Conditor dans Caplab

Par
Gilles Scotto Di Carlo*, Amue, *Consultant SI Recherche, Projet Caplab*
Frédéric Ollier, Amue, *Consultant SI Recherche, Référent Produit Projet Caplab*
Caplab vise à :

- Aider à la veille et au montage des projets
- Recenser, décrire et suivre les projets et activités de recherche
- Partager et lier l'information
- Répondre aux besoins de pilotage en unité de recherche et en établissement

La version pilote est accessible à 7 tutelles (5 universités, le CNRS et l'Inserm) et à 11 de leurs structures de recherche pilotes.
Caplab interroge l'API Conditor à partir des IdRef, IdHal, OrcId déclarés par le chercheur dans Caplab et les affiche dans l'interface accessible au chercheur dans Caplab.
Par ailleurs, une expérimentation est en cours dans le cadre de l'AAPG 2020 de l'ANR. Le chercheur déposant un projet sous Iris (SI de l'ANR) renseigne son OrcId, Caplab interroge alors l'API Conditor à partir de cet OrcId et des identifiants IdRef et IdHal déclarés par le chercheur dans Caplab, Conditor renvoie les notices de production à Caplab qui les met à disposition du chercheur dans Iris.

Pour l'Amue, cette étape a été mise en place dans de très bonnes conditions : utilisation commune d'Elasticsearch, API simple à interroger, équipe Conditor à l'écoute.
Les souhaits d'amélioration portent sur :

- la profondeur, nombre d'années disponibles des notices de publications,
- la largeur, notices issues d'archives institutionnelles ou d'éditeurs privés,
- l' enrichissement des notices : Identifiant structure RNSR, identifiants chercheurs OrcId, IdHal, IdRef,
- une notice chapeau Conditor pour les notices en doublons.

3.6 Preuve de concept : intégration du « service » Conditor à l'AAPG

Par **Cyril Demange**, ANR, *Maître d'ouvrage des Systèmes d'Information, Direction des Opérations Scientifiques*

Comme cela a été dit, Iris récupère les données de Conditor au travers de Caplab pour les chercheurs des unités pilotes.
Pour déposer un projet, 2 formulaires en ligne sont à compléter :

- Le projet
- Le CV des chercheurs avec les 5 publications majeures.

Ces CV sont persistants et non liés à un projet.

Le POC (preuve de concept) permet à un chercheur d'une des 12 unités pilotes ayant complété son OrcId de pré-remplir une case « Publication » en commençant à saisir une partie du titre puis en sélectionnant le titre de la production recherchée.

Le service proposé simplifie « la vie du chercheur ». Il est en test mais le souhait de l'ANR est de passer en vraie grandeur sur les laboratoires utilisateurs de Caplab (environ 0,5% des dépôts AAPG) et sur l'ensemble de l'AAPG.
Le gain utilisateur est évident … mais les informations récupérées doivent être plus complètes.

3.7 Discussion et commentaires

Actuellement, la base de CV de l'ANR n'est pas ouverte.
Caplab utilise les codes RNSR pour récupérer de Conditor des "pools" de publications par unité de recherche. Les équipes ne sont pas systématiquement décrites dans le RNSR alors qu'il est souvent nécessaire de lister les productions d'une équipe au sein d'une unité : elles peuvent être décrites dans Auréhal et sont prévues dans Caplab...

4 Table ronde - Et après ?

Serge Bauin*, CNRS-Dist, animateur de la table ronde a proposé aux intervenants **d'imaginer un monde idéal**.

Pour **François Mistral***, Abes, *responsable d'IdRef et coanimateur du groupe travaillant sur la méthodologie de construction du réservoir de métadonnées Conditor,* la dimension multipartenaire est une réussite du projet et il ne faut pas perdre cette dimension au niveau opérationnel car cela fait grandement avancer les projets. Pour lui, la mutualisation est une vraie force et même une planche de salut.
Seules les publications sont décrites dans Conditor (du fait des sources disponibles) mais il serait intéressant d'élargir le périmètre aux données de recherche notamment.
Il fait également remarquer que lorsque l'usage d'identifiants se généralisera, les données dispersées dans les différents systèmes pourront être liées et certaines validations ne seront plus nécessaires.

Comme le souligne **Frédéric Ollier**, Amue, *Consultant SI Recherche, Référent Produit Projet Caplab,* Caplab vise justement à agréger des informations au niveau national sur les projets en s'appuyant sur des clés (RNSR par exemple).
Il ajoute que Caplab est « consommateur » de données Conditor et fait remarquer que le versement de données Conditor vers Caplab et vers le SI de l'ANR a pris moins d'un an notamment du fait de la belle qualité opérationnelle de Conditor.
La fraîcheur des données mises à disposition par Conditor est importante pour Caplab qui pourrait devenir une source car les utilisateurs de Caplab sont les personnels des unités : il n'est pas du tout exclu de faire non seulement des validations mais aussi de pouvoir déclarer des produits dans Caplab pour alimenter Conditor.[4]

Comme il a été dit, Conditor permettra d'améliorer le RNSR mais le souhait de **Xiaofeng Chen*,** MESRI, *responsable fonctionnel du RNSR* est que Conditor ne signale plus de lacunes ou anomalies.
Le dispositif technique et organisationnel RNSR va s'améliorer via l'échange avec d'autres SI permettant des mises à jour automatiques et l'équipe nécessaire va être renforcée.

Odile Hologne, Inraé, *directrice de la Direction pour la Science Ouverte,* rappelle que l'intérêt initial de l'Inra pour Conditor était de récupérer les métadonnées des bases WoS et PubMed afin de faciliter le travail de dépôt du full text par le chercheur. Il semble que la situation s'est inversée : Conditor va s'appuyer sur les archives.

[4] Tirer parti de tout acte de gestion pour enrichir les données Conditor

De son point de vue, il serait bien de penser un système de la métadonnée, fédéré et sémantique, permettant à chaque dispositif de se relier aux métadonnées de base (Dublin core) et à des référentiels.
Tracer le cycle de vie de la publication est important : le lien entre preprints (et le peer review) et l'article final.

Serge Bauin* fait remarquer que c'est l'un des travaux actuels de COAR.
Voir à ce sujet le communiqué : "Meeting participants agree to work together on a technical architecture for distributed peer review on repository resources."

Il note également que la plus-value de Conditor est d'utiliser des métadonnées qui viennent d'ailleurs (CrossRef...) : Conditor est un dispositif de filtrage des données avant de les ingérer dans les dispositifs bénéficiaires.

Lors des échanges avec la salle **:**
David Aymonin, directeur de l'Abes qui coordonne le « new consortium OrcId » préconise d'utiliser OrcId comme source.
Il fait remarquer par ailleurs que les éditeurs ont compris que les métadonnées devaient mieux circuler et qu'il était nécessaire de lâcher un peu prise afin de garder une certaine attractivité.

Suite à une question concernant le lien entre IPERU et Conditor, **Françoise Rojouan**, Hcéres, OST, rappelle que la base OST s'appuie sur la base commerciale WoS de Clarivate analytics qui est sous contrat de licence et qu'une proposition de convention [5] pour la fourniture des repérages IPERU, n'a pas abouti.

Pour conclure **:**
Serge Bauin fait remarquer que la plus-value actuelle de Conditor est de « faire le ménage » dans les métadonnées : il faudrait de son point de vue que les flux de métadonnées passent déjà dans Conditor avant d'être ingérées dans les dispositifs ESR.
Il regrette un grand absent : le BSO et souligne la nécessité d' « articulation » entre les deux outils.

David Aymonin revient sur l'ouverture des données des éditeurs à rechercher (exemple d'Elsevier).
Odile Hologne revient sur le fait que la circulation des métadonnées est un enjeu, que de plus en plus de bases de preprints existent et qu'il est nécessaire de tracer l'évolution vers la publication.

Serge Bauin souligne la nécessité de préciser les acteurs de cette évolution.
Conditor a été "développé" à l'Inist : la question est « comment organiser concrètement la suite ? »

[5] Elle portait sur la fourniture des repérages de l'année 2014 dans le cadre du pilote.