L'audit du système documentaire d'un institut de recherche à l'heure des nouveaux enjeux de l'information scientifique et technique : l'exemple du SID Horizon de l'IRD

Sylvie Rouquier

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Equipe pédagogique Stratégies

INTD

MEMOIRE pour obtenir le Titre enregistré au RNCP
"Chef de projet en ingénierie documentaire et gestion des connaissances"
Niveau I

Présenté et soutenu par

Sylvie Rouquier

le 12 décembre 2018

L'audit du système documentaire d'un institut de recherche à l'heure des nouveaux enjeux de l'information scientifique et technique

L'exemple du SID Horizon de l'IRD

Jury :
Chartron Ghislaine, professeure du Cnam, chaire Ingénierie documentaire, directrice de l'INTD depuis 2007, directrice de mémoire
Goury Laurence, responsable de l'administration de la base Horizon Pleins textes, responsable de stage

Promotion 48 (2017-2018)

Remerciements

Je tiens à remercier tout spécialement Laurence Goury pour m'avoir pleinement accompagnée dans cette mission et dans la rédaction de ce mémoire, pour avoir partagé ses connaissances et compétences avec moi, et pour sa patience.

Je remercie particulièrement Ghislaine Chartron pour ses précieux conseils dans l'élaboration de ce mémoire.

Je remercie chaleureusement Caroline Doucouré, Emilie Brunet, Francine Delmas, Elisabeth Ambert, Pascal Aventurier, Luigi Rossi, pour le temps qu'ils m'ont consacré, et l'équipe IST en général pour son accueil, ce qui m'a permis de me sentir pleinement intégrée dans le service.

Je remercie enfin toutes les personnes qui ont accepté de m'accorder de leur temps pour un entretien, et pour les éléments précieux qui m'ont ainsi aidée à alimenter ma réflexion.

Notice

Description bibliographique

ROUQUIERSylvie. L'audit du système documentaire d'un institut de recherche à l'heure des nouveaux enjeux de l'information scientifique et technique - L'exemple du SID Horizon de l'IRD [en ligne]. Mémoire, CNAM-INTD, 2018, 129 p.

Résumé

Cette étude a pour objet un comparatif des systèmes de gestion documentaires pour une archive ouverte des EPST en France et a été réalisée pour l'IRD. L'institut souhaite faire évoluer le système documentaire en place gérant son archive ouverte institutionnelle Horizon.
Des entretiens ont été réalisés auprès de plusieurs instituts de recherche français, et une université étrangère d'un pays émergent, puis compilés en un tableau comparatif final accompagné d'une liste de préconisations.
La réingénierie du système gérant Horizon doit prendre en compte les nombreux partenariats avec les pays du Sud mis en place par l'IRD.
L'étude se place aussi dans un contexte de profondes transformations que connaissent l'information scientifique et technique, ainsi que le monde de la recherche depuis quelques années, et tente de montrer comment les nouveaux enjeux issus de ces changements sont au cœur de la réflexion qui sera menée dans le cadre de ce projet et pourront influer sur le choix final.
Cette étude s'attache à montrer le rôle des acteurs de l'IST à l'heure de ces nouveaux enjeux, et les nombreuses initiatives et pistes de réflexion en cours.
L'histoire de l'IST est brièvement retracée, ainsi qu'un état de la législation actuelle, puis les aspects de libre accès, d'édition scientifique, de mutualisation et centralisation des ressources, les enjeux bibliométriques et les évaluations qui en découlent, enfin, sont exposés, et replacés dans le contexte de notre mission. Les nouvelles interopérabilités, rendues possibles par les avancées technologiques, et qui permettent de nouvelles méthodes de stockage et de recherche de l'information sont également exposées.
Cette étude présente enfin des scénarios possibles pour l'institut, ainsi que des préconisations et recommandations à prendre en considération pour un choix éclairé.

Descripteurs

Indexation thématique :

Libre accès ; Open Access ; Archive ouverte institutionnelle ; Information scientifique et technique ; Recherche scientifique ; Agence d'évaluation de la recherche et de l'enseignement supérieur ; Publication scientifique ; Accès à l'information ; Web de données ; Web sémantique ; SIGB ; Système de gestion de bibliothèque mutualisé ; SWOT ; Réingénierie de système ; Etude comparative.

Indexation géographique :

France ; Europe ; Pays en développement.

Abstract

The purpose of this study is to compare the documentary management systems for an open archive of research institutes in France. The institute wishes to develop the existing documentary system managing its open repository Horizon.
Interviews were conducted with several French research institutes and a foreign university from an emerging country. The results were then compiled into a final comparative table with a list of recommendations.
The reengineering of the Horizon management system must take into account the many partnerships set up by the IRDwith Southern countries.
This study has been carried out for the IRD and is conducted in a context of important changes in the ways scientific communication is disseminated.
The study takes place in a context of profound transformations for the scientific and technical information and forthe scientific research. This study attempts to show how the new challenges resulting from these changes are at the heart of the reflection that will be carried out in this project and may influence the final choice.
This study also aims to show the role of the actors of scientific and technical information at this time of new challenges, and the many initiatives for reflection underway.
The history of the scientific and technical information is briefly retraced, as well as a review of the current legislation. Then the aspects of Open access, scientific publishing, mutualisation and centralisation of resources, bibliometric issues and the resulting evaluations will be reviewed and placed in the context of our mission. New interoperabilities, made possible by technological advances, and allowing new methods of storing and retrieving information are also presented.
Finally, this study presents possible scenarios for the institute, as well as recommendations to be taken into consideration for an informed choice.

Keywords

Open access; Institutional open archive; Open access repository ; Scientific and technical information; Scientific research; Scientific publication ;Bibliometric indicators; Access to information; Data web; Semantic web; Integrated Library System; Library management system ; SWOT ; System reengineering ; Comparative study.

France ; Europe ; Developing countries

Table des matières

Liste des figures

Introduction

« *La science ouverte vise à construire un écosystème dans lequel la science est plus cumulative, plus fortement étayée par des données, plus transparente, plus rapide et d'accès plus universel* ».[3]
Cette définition a été donnée par le plan national pour la science ouverte divulgué le 4 juillet dernier par le ministère de l'enseignement supérieur, de la recherche et de l'innovation.
Ce plan s'articule autour de trois axes : généraliser l'accès ouvert aux publications, structurer et ouvrir les données de la recherche, et inscrire la recherche française dans une dynamique durable, européenne et internationale.
Au niveau européen, l'Open science a pour objectif d'apporter des valeurs éthiques, promouvoir la science des données, et encourager la croissance économique de l'Union par l'innovation. (Chartron, 2018)[1]
Ces orientations relèvent des nouveaux enjeux que connait le monde de l'information scientifique et technique depuis plusieurs années. Grâce aux récentes avancées technologiques qu'a connu le monde l'information et de la communication, la circulation des connaissances scientifiques est plus rapide et accessible à tous.
L'enjeu d'un accès généralisé au savoir remonte pourtant bien avant l'invention du Web. Cette problématique a toujours été au cœur de la production scientifique, dont l'objectif était de diffuser les résultats de la science.
Cet objectif est toujours poursuivi aujourd'hui, non seulement par la communauté scientifique mais aussi par les institutions politiques de la plupart des pays qui encouragent la libre circulation des savoirs par le biais de dispositifs législatifs incitatifs et coercitifs.
En France, les différents EPST et Universités relaient ces recommandations pour les mettre en œuvre dans le monde de la recherche
Le Décret du 24 Novembre 1982 portant sur l'organisation et le fonctionnement du CNRS stipule dans son article 2 que l'organisme a une vocation nationale à "*développer l'information scientifique*".[2]
Elle rappelle en outre que le partage des connaissances passe par le partage des résultats de la science.
Le développement des archives ouvertes depuis le début des années 2000 poursuit le même objectif.
Cet objectif de partage implique aussi une prise en considération des besoins des chercheurs. Ceci est le travail que mènent les professionnels de la documentation scientifique.
La direction de l'information scientifique et technique de l'organisme insiste depuis longtemps déjà sur « *le potentiel des outils numériques pour révolutionner la production de l'information scientifique et technique et pour favoriser le partage des résultats de la science* »[1]
Nous nous efforcerons dans ce mémoire de montrer comment les récentes avancées réalisées dans ces deux domaines sont liées.
Il existe de puissants logiciels capables de stocker et diffuser une documentation considérable en libre accès.
L'IRD se trouve aujourd'hui dans un contexte de refonte de son système documentaire de gestion d'archive ouverte institutionnelle.,
La mission du stage dans lequel s'inscrit ce mémoire a eu pour objet de faire une étude comparative des outils et solutions proposés pour répondre à ces questions.

[1] http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf

Quel outil choisir à cette époque chargée de mutations que connait actuellement l'information scientifique et technique, tout en respectant les nombreux règlements et lois l'encadrant ?
Comment les enjeux de libre accès, d'évaluation des chercheurs, dont dépendent les subventions, et de publication de production scientifique peuvent influer le choix d'un outil ? Pourquoi faut-il les prendre en compte ?
Comment les nouveaux logiciels et les nouvelles fonctionnalités offertes répondent-elles à ces enjeux ? Comment peuvent-elles être déterminantes dans ce choix ?
Nous allons tenter de répondre à ces questions en commençant par un rappel de la définition de l'histoire de l'information scientifique et technique. Les enjeux énoncés étant encadrés par la législation, nous en rappellerons les grands principes.
Nous consacrerons ensuite une partie à décrire ces nouveaux enjeux, en expliquant comment leur prise en compte se répercute concrètement sur le choix d'un outil.
Dans une troisième et dernière partie, nous présenterons les résultats de notre étude comparative, et tenterons d'en dégager des préconisations.

Partie 1 : environnement, contexte et perspectives historiques

L'histoire de l'IST est liée à celles de la science et des évolutions technologiques des communications. Ces trois domaines ont évolué ensemble, l'un entrainant l'autre. L'IST a bénéficié de chaque nouvelle avancée technologique pour pouvoir progresser dans ses missions de valorisation de la production scientifique. L'archive ouverte de l'IRD s'est elle aussi développée au gré des évolutions technologiques et des changements qu'a connu l'IST ces dernières décennies. Les fonctionnements et les rouages de ces domaines posent par ailleurs des questions juridiques sur l'accessibilité à la connaissance et les droits d'auteur liés aux publications des chercheurs.

Dans cette première partie, nous allons montrer comment le contexte et les perspectives historiques nous amènent à ce qu'est l'IST aujourd'hui afin de mieux comprendre ses enjeux actuels. Nous tenterons d'expliquer comment les évolutions de la science et des progrès technologiques ont amené à une réflexion par les professionnels de l'IST sur les enjeux d'une science accessible par tous.

L'IST n'évolue en effet pas à la même vitesse dans tous les pays, et l'on retrouve dans notre domaine les problèmes rencontrés par les pays émergents. La mission de développement de L'IRD conduit l'institut à entretenir des liens étroits avec ses partenaires du sud. Cette particularité prend alors une place importante dans la réflexion et les actions menées en conséquence par le personnel de documentation.

Nous présenterons le système documentaire de l'IRD que cette équipe a en charge, et essayerons de comprendre le rôle que peut jouer un tel système dans les objectifs de l'institut à l'échelle internationale. Ce système, qui gère l'archive ouverte Horizon, fait l'objet du benchmark qui sera présenté en troisième partie. C'est un système complexe qu'il faut resituer dans son contexte pour comprendre les enjeux liés à son évolution.

Nous commencerons donc par un rappel de l'histoire de l'IST et un état de la législation actuelle encadrant ce domaine.

Enfin nous présenterons le système d'information derrière l'archive ouverte de l'IRD : Horizon.

1.1 Rappel sur l'histoire de l'IST

« L'information scientifique et technique (I.S.T.) regroupe l'ensemble des informations produites par la recherche et nécessaires à l'activité scientifique comme à l'industrie. De par sa nature, l'I.S.T. couvre tous les secteurs scientifiques et techniques et se présente sous de multiples formes : articles, revues et ouvrages scientifiques, spécifications techniques décrivant des processus de fabrication, documentation technique accompagnant les produits, notices de brevet, bases de données bibliographiques, littérature grise, banques de données brutes, archives ouvertes et entrepôts de données accessibles sur internet, portails, etc. »[2]

Cette définition que donne le Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche liste les formes que prend l'IST. Ces articles, revues, ouvrages scientifiques et autres représentent aujourd'hui une masse considérable d'informations et de données que les documentalistes de ce secteur doivent organiser, agréger, fédérer, valoriser.

Depuis un certain temps déjà des chercheurs et bibliothécaires aspirent à répertorier cette production. Le recensement de la production scientifique remonte au 19ème siècle, lorsque certains chercheurs ont souhaité recenser les articles scientifiques dans le but de valoriser leurs auteurs. La révolution industrielle survenue à cette époque a été la source d'une forte prolifération d'écrits scientifiques. Les périodes de forte croissance de la production scientifique correspondent en effet souvent aux périodes des grandes avancées technologiques.

L'expression d'IST est récente, mais dès les années 1900 ses éléments fondateurs sont déjà présents. Le développement des échanges scientifiques entre communautés de savants, les progrès techniques amorcés au XIXe siècle, la diffusion du livre, la multiplication des revues scientifiques, le développement de l'enseignement scientifique et le lancement de méthodes d'organisation de l'information comme l'indexation sont autant d'éléments qui vont commencer à alimenter des réflexions autour de la valorisation et de la diffusion des écrits scientifiques. Plus tard, la période de croissance économique que furent les trente glorieuses a incité la recherche à se rapprocher de l'industrie. Ces faits vont aboutir à une accélération de la science avec pour conséquence un foisonnement des écrits scientifiques.

Le nombre des revues scientifiques créées au niveau international, n'a fait que croitre depuis. Ce chiffre est passé de 1000 en 1951 à 71 000 en 1987 (source Meadow, 1998, p.15). Aujourd'hui, les estimations relatives à l'augmentation de la publication scientifique sont nombreuses. Derek de Solla Price estime que celle -ci double tous les 10 à 15 ans. Un rapport de l'UNESCO[5] estime que le nombre d'articles scientifiques est passé de de 1 029 471 articles à 1 270 425 entre 2008 et 2014. Cette étude de se base sur l'index de citations scientifiques de Thomson Reuters (Science Citation Index of Thomson Reuters du Web of Science).

Ces chiffres sont des estimations, mais il en ressort un besoin de comptabiliser et gérer la production scientifique. Ce processus va débuter avec méthodologie dès les années 50. A cette époque, sont créés les centres de recherches et avec eux apparaissent les centres de documentation chargés de gérer la documentation produite par ces centres. Parallèlement,

[2] http://www.enseignementsup-recherche.gouv.fr/cid20438/les-missions-de-l-information-scientifique-et-technique.html

les grandes entreprises industrielles s'équipent elles aussi de centres de documentation. (Chartron, 2001) **[4]**

Les avancées technologiques depuis cette décennie sont nombreuses, la science progresse et propose de plus en plus d'outils de haute technologie pour les recherches. Une telle croissance de la production scientifique a vite amené à se poser la question de l'organisation des informations. Les structures visant à encadrer l'IST voient le jour dans les années 60 aux Etats-Unis, l'Europe lui emboitant le pas rapidement. Les pays anglo-saxons sont en effet les premiers à comprendre l'importance de l'information scientifique et technique, par sa puissance stratégique. (Chartron, 2001) [4]
En 1959, la Délégation générale à la recherche scientifique et technique créé un comité chargé d'étudier les problèmes posés par la rédaction et la diffusion des informations scientifiques, notamment leur fiabilité et leur exactitude.
Parallèlement, l'IRD dès cette époque affiche sa volonté de conserver et diffuser sa production scientifique en créant sa propre archive institutionnelle (ORSTOM 1955). (Rossi, 2018) **[13]**
Un centre de documentation, directement rattaché à la direction de l'institut est créé. Cette action est attestée par un arrêté du 1er décembre 1955 paru au Journal officiel de la République française du 21 décembre 1955 qui stipule, dans son article 6 que « *le Centre de documentation est chargé d'assurer le dépouillement, la conservation et la diffusion de la documentation scientifique et technique se rapportant aux activités de l'ORSTOM*[3] ». (Rossi, 2018) **[13]**
Les questions liées à la diffusion de l'information afin de la rendre visible et repérable par le plus grand nombre commencent à être abordées plus en profondeur. C'est le début d'une véritable réflexion menée par les instituts de recherche et les universités sur la manière la plus efficace de disséminer et repérer l'information scientifique. Cette réflexion est toujours en cours aujourd'hui.

Les institutions dans les années 60 ont déjà conscience de l'importance et de la complexité de la mission des centres de documentation scientifique et technique, qui commencent à se multiplier dans le pays. Cette mission consiste à mettre le plus rapidement possible à disposition des chercheurs et étudiants de l'information produite par leurs pairs, ce qui implique de traiter un volume considérable de documents en un temps limité.
Le centre de documentation du CNRS traite déjà à l'époque l'équivalent de plusieurs millions de pages par an. Ce traitement est encore manuel à l'époque, mais les années 60 sont marquées par le début de l'informatisation, et les récentes avancées dans le traitement automatique de l'information intéresse fortement le monde de l'IST.
L'informatisation va par conséquent trouver dans la documentation scientifique un des premiers champs d'application.

L'histoire de la science de l'information et la documentation scientifique électronique vont se retrouver étroitement liées et dès les années 70 la commission européenne associe clairement le développement d'un marché européen de l'information au développement des technologies.

Il faut attendre 1993 pour que le ministère en charge de l'enseignement supérieur et de la recherche implémente le réseau RENATER: le Réseau national de télécommunications pour la technologie, l'enseignement et la recherche. Ce maillage permet de relier les différentes

[3] L'ORSTOM est le nom sous lequel a été créé l'IRD en 1943 et qui signifie Office de la recherche scientifique et technique outre-mer.

universités et les différents centres de recherche entre eux en France métropolitaine et dans les départements d'outre-mer.

Puis l'avènement d'internet a engendré des avancées technologiques considérables permettant aux instituts de recherche et aux universités de diffuser et échanger en un temps record leurs productions.
Le système repose alors encore majoritairement sur les revues scientifiques, qui sont devenues innombrables, et malgré les progrès en termes de télécommunications, tous les pays n'accèdent pas de façon égalitaire aux connaissances, car toutes ne disposent pas d'un budget d'acquisition leur permettant de payer les abonnements.
Au début des années 90 nait donc le mouvement de l'open access qui a pour objectif de donner de la visibilité aux travaux de recherches des chercheurs. Le mouvement va se concrétiser par la création d'archives ouvertes que les institutions vont développer dans le courant des années 2000.
L'INIST propose la définition suivante de l'archive ouverte : « *Le terme archive ouverte désigne un réservoir où sont déposées des données issues de la recherche scientifique et de l'enseignement et dont l'accès se veut ouvert c'est-à-dire sans barrière. Cette ouverture est rendue possible par l'utilisation de protocoles communs qui facilitent l'accessibilité de contenus provenant de plusieurs entrepôts maintenus par différents fournisseurs de données* »[4].
L'INIST définit l'archive institutionnelle comme relevant« *d'une institution (université, grande école, organisme de recherche, association professionnelle) et qui a pour objectif de contenir, valoriser et conserver l'ensemble de la production scientifique de celle-ci*».

En 1999, lors de la convention de Santa Fe est créé le protocole OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting) qui permet aux réservoirs d'archives ouvertes d'échanger entre eux, d'être visibles pour les moteurs de recherches venant moissonner ces réservoirs, en les interrogeant par une même requête. Ce protocole représente une réelle avancée dans la visibilité des écrits scientifiques.
En 2001, le CNRS lance sa propre plateforme d'archive ouverte, Hal (Hyper articles en ligne) et parallèlement les institutions ouvrent des plateformes pour exposer leurs publications.

La première finalité d'une archive institutionnelle est indéniablement de mettre ses écrits scientifiques à l'égard du plus grand nombre, pour être avant tout partagé avec ses propres chercheurs, mais aussi avec le grand public, et enfin d'être plus visible par les agences de financement. Nous verrons dans la deuxième partie que les subventions dépendent largement de la production d'une institution, et une AOI permet de mesurer de manière exhaustive les publications apportées.
L'ouverture d'une AOI permet à une institution, en plus d'exposer les résultats de ses chercheurs d'en recenser plus facilement la production, d'assurer une pérennité à sa production, et enfin de partager des connaissances.

Aujourd'hui, l'IST tend à aller vers une généralisation des savoirs, encouragée en cela par la législation qui favorise la diffusion des résultats de la recherche.

1.2 Environnement législatif

L'avènement de l'ère du numérique et du libre accès offre aujourd'hui à la science des possibilités immenses de diffusion et de partage de l'information scientifique. Cette nouvelle donne implique un encadrement juridique solide qui aura pour objectifs d'assurer à la fois la

[4] http://openaccess.inist.fr/?+-Archive-ouverte-+

dissémination des résultats de la recherche tout en garantissant une protection du droit d'auteur.
Toute œuvre est protégée dès sa création par des droits moraux et patrimoniaux. Les premiers sont incessibles, perpétuels et imprescriptibles, alors que les seconds sont cessibles partiellement ou totalement concernant les droits de reproduction et les droits de représentation, c'est-à-dire la mise à disposition de l'œuvre sur un site.
Concernant les publications du chercheur, elles pourront être déposées dans une archive ouverte à condition que tous les co-auteurs soient d'accord, et notamment l'éditeur avec lequel l'auteur aura éventuellement signé un contrat d'édition.
Les chercheurs concluent en effet avec les éditeurs scientifiques des contrats par lesquels ils leur cèdent tout ou partie de leurs droits pour une durée déterminée que l'on appelle la période d'embargo. Ces contrats peuvent avoir pour conséquence d'empêcher ou retarder le dépôt, ce qui va à l'encontre du principe de libre accès.
Certains éditeurs acceptent toutefois l'auto-archivage depuis plusieurs années (l'acte par lequel les chercheurs déposent eux-mêmes leurs articles dans des archives ouvertes) dans certaines conditions. Le site Sherpa/Romeo recense les politiques des éditeurs en la matière.

La France a légiféré sur le sujet en 2016. L'article 30 de la loi pour une République numérique du 7 octobre 2016[5]autorise en effet les chercheurs à auto-archiver leurs publications dans le format manuscrit (c'est-à-dire la version acceptée pour publication) lorsque la recherche a été financée au moins par moitié sur les fonds publics et si l'éditeur diffuse l'article en libre accès, même s'il a cédé tous ses droits à un éditeur. Elle ne s'applique toutefois qu'aux écrits publiés dans des périodiques paraissant au moins une fois par an, c'est-à-dire aux articles scientifiques et aux actes de colloque publiés sous forme de revue.
D'autre part, quelle que soit la durée d'embargo imposée par l'éditeur, le dépôt pour être effectué six mois après la date de première parution pour les sciences techniques et la médecine ou douze mois pour les sciences humaines et sociales.

Par ailleurs, les licences Creative Commons précisent les droits des utilisateurs pour un document déposé en libre accès et permettent aux chercheurs d'exercer leur droit d'auteur tout en précisant la façon dont leur publication peut être réutilisée. Une étude réalisée par la fondation Creative Commons en 2015 recense 1.4 million d'articles scientifiques qui auraient été publiés sous une de leurs licences.[6]
Les auteurs ne sont pas légalement obligés d'attribuer une licence à leur œuvre, et elle ne se substituent pas à la législation du pays, mais les institutions et notamment le site Sherpa recommande leur utilisation afin de faciliter la circulation et le partage des publications.
La loi encourage donc la diffusion des connaissances et des résultats de la recherche en encadrant leur (ré)utilisation.

La Commission européenne s'est elle-même engagée dans la voie de la science ouverte, et les législations nationales se déclinent depuis pour assurer un relais à l'intérieur des pays de l'union européenne.

En France, la loi pour une République numérique du 7 octobre 2016[7] a établi les bases d'une politique favorisant l'ouverture des données et des connaissances.
Cette loi a été rédigée sur la base de travaux de réflexion menés par tous les acteurs concernés, et notamment, dans son volet concernant la recherche française, par le CNRS,

[5] Article consultable sur le site Legifrance.gouv.fr
[6] L'étude est consultable en ligne sur le site stateof.creativecommons.org
[7] Texte complet consultable sur le site Legifrance.gouv.fr

l'EPRIST (l'association des responsables IST des organismes de recherche), le Consortium Couperin[8]et l'ADBU (L'Association des directeurs et personnels de direction des bibliothèques universitaires et de la documentation).
Le résultat de cette réflexion a permis de dégager des stratégies pour la politique scientifique. L'objectif, en accordance avec la science ouverte, est donc de valoriser l'IST et de donner la priorité à « *l'accès aux données scientifiques mais surtout leur utilisation via des outils de traitement* ».(CNRS)[9]
La loi pour une République numérique a ainsi encadré certaines pratiques, restées vagues jusqu'à présent, et permis dans le même temps de mettre en vigueur des mesures importantes pour l'IST.
La pratique du TDM par exemple (Text and Data Mining, ou fouille de texte en français) est désormais encadrée par l'article 38[9], qui créé une exception au droit d'auteur en autorisant la fouille de texte à des fins de recherche publique.

Plus récemment, la directive sur le droit d'auteur adoptée le 12 septembre dernier par le Parlement européen propose de généraliser la pratique de la fouille de texte à l'ensemble de l'union européenne. Cette pratique est primordiale pour la science car elle permet de fouiller de très larges corpus de textes, d'en extraire des informations et d'établir des liens entre elles. Ceci suppose que les publications des chercheurs soient accompagnées des données brutes relevées pour réaliser leurs analyses et en extraire les résultats.
Cette pratique a déjà été reconnue par certaines législations dans de nombreux pays, au niveau européen (2012 en Irlande et 2014 pour le Royaume-Uni)) comme au niveau international (2009 pour le Japon).

Cette volonté d'ouvrir la science et d'en partager les résultats le plus largement possible s'inscrit dans une dynamique internationale. Les pays sont nombreux à s'engager dans la voie du libre accès, et de légiférer sur les délais d'embargo.
Les textes en vigueur sont souvent des recommandations, et nombreux sont les gouvernements à travers le monde qui légifèrent et rendent obligatoires ces directives, soutenus par les agences de financement de la recherche.
L'environnement législatif se construit donc peu à peu, et encourage la diffusion de la production scientifique tout en facilitant son traitement.

Nous reviendrons sur ces sujets dans la deuxième partie pour les détailler et expliquer comment ces mesures peuvent influer sur la façon dont une institution va exposer son archive ouverte.

1.3 Présentation de l'IRD

Nous allons présenter ici l'IRD, cadre dans lequel s'est déroulée la mission de stage, et décrire son archive ouverte.

1.3.1 présentation de l'institut

L'Institut de Recherche pour le Développement (IRD) est un établissement public à caractère scientifique et technologique (EPST). Il est placé sous la double tutelle du ministère de l'enseignement supérieur et de la recherche, et du ministère des affaires étrangères et européennes.

[8]Cette association dont la mission est de négocier auprès des fournisseurs des tarifs et conditions d'accès aux publications scientifiques et autres ressources documentaires numériques est subventionnée par le Ministère de l'Enseignement Supérieur et de la Recherche
[9] Article consultable sur le site Legifrance.gouv.fr

C'est un organisme pluridisciplinaire reconnu internationalement pour ses programmes scientifiques centrés sur les relations entre l'Homme et son environnement, dont l'objectif est de contribuer au développement durable des pays du Sud.

L'IRD compte aujourd'hui 66 unités de recherche, emploie 2013 agents IRD (805 chercheurs et 1214 ingénieurs et techniciens IRD). 31,4% des agents sont en poste à l'étranger. Près de1451 références d'articles publiés en 2016 par les chercheurs de l'IRD ont été signalés dans le Web of Science, dont61 % d'éco-publication avec un partenaire du Sud[10].

L'activité scientifique de l'IRD se déploie sur cinq départements, balayant les sciences des techniques et de la matière, mais aussi les sciences sociales :

- Milieux et environnement
- Ressources vivantes
- Société et santé
- Expertise et valorisation
- Soutien et formation des communautés scientifiques du Sud

Les premières disciplines étudiées à l'IRD étaient la pédologie et l'hydrologie. Les sujets d'étude se sont depuis diversifiés, et aujourd'hui s'étendent à la biologie des plantes, l'halieutique, la chimie, l'hydrologie, la santé, la société…

Son activité dans ces domaines s'articule autour de trois missions fondamentales :

- La recherche
- L'expertise et la valorisation
- La formation

La stratégie de recherche de l'IRD est inscrite dans une dimension internationale. L'institut mène en effet ses missions en partenariat avec les universités et les structures de recherche des Pays du Sud.
Ces partenariats sont nombreux et nécessitent donc des canaux de communication solides et fiables pour correspondre et échanger des informations.
L'institut pilote de nombreux programmes à l'international, et notamment dans les zones tropicales et organise des partenariats de recherche avec de nombreux pays. Ses laboratoires de recherche sont présents partout dans le monde.

Sa mission va au-delà de la recherche pour le développement, car l'institut a également vocation à participer au développement des capacités de recherche des partenaires du Sud.
L'équipe en charge de la documentation fait partie du service de l'information technique et scientifique, et la formation des chercheurs et doctorants des pays du Sud fait aussi partie de ses missions.

La carte ci-dessous illustre le positionnement de l'IRD dans le monde.

[10] http://www.ird.fr/l-ird/rapports-d-activite-annuels/2017

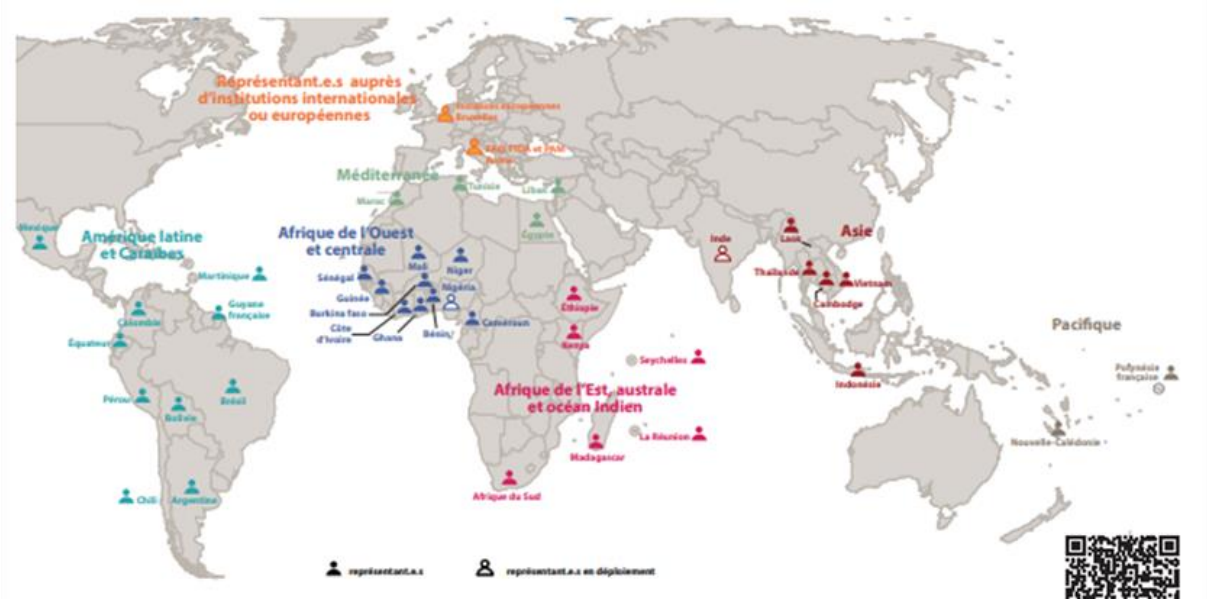


Figure 1 : L'IRD dans le monde[12]

1.3.2 spécificités de l'IRD

- Pluridisciplinarité de l'IRD

La première spécificité de l'IRD est sa pluridisciplinarité. L'institut conduit des recherches en STM aussi bien qu'en SHS.

Cette singularité est à prendre en compte car elle pose une difficulté supplémentaire pour les administrateurs d'une archive ouverte qui doivent porter un effort supplémentaire sur ces publications afin de les rendre ostensibles et mesurables.

- Publications multilingues

L'IRD présente également la particularité de publier un nombre important de documents en français, dont la consultation est importante. Ceci apporte une limite à la croyance que seuls les articles publiés en anglais ont vocation à être exposés dans une archive ouverte.

Les pays du Sud téléchargent à eux seuls 70% des documents publiés sur Horizon[11] dont l'Afrique est la plus grande utilisatrice, représentant à elle seule 47% du total des accès à travers le monde, ce qui correspond à 1 107 007 accès. (Rossi, 2017)[14]
Les statistiques de consultation remontées montrent que les internautes d'Afrique francophones recherchent de la documentation en langue française.

- Les partenariats avec les pays du Sud

La particularité la plus importante à prendre en compte est l'étroite collaboration de l'IRD avec les pays du Sud.

[11] Chiffre mis en ligne sur la page d'accueil d'Horizon, rubrique Consultations par zone géographique.

L'institut a pour mission de participer au déploiement des capacités de recherche des partenaires du Sud.
La DDUNI est d'ailleurs en charge d'un service informatique scientifique dont la mission consiste à apporter un support aux représentations outre-mer, par exemple en mettant des espaces d'hébergement ou des machines virtuelles à leur disposition, ou encore en offrant un appui au volet informatique de projets scientifiques.

La solution documentaire doit être accessible aux pays du Sud pour qui l'accès à internet est plus compliquée, et qui, d'autre part ne disposent pas des budgets suffisants pour accéder aux abonnements des revues, notamment en STM.
L'IRD encourage donc le libre accès, qui reste un moyen efficace de diffuser et valoriser la production scientifique dans les pays émergeants. Les faibles budgets d'acquisition empêchent les bibliothécaires et documentalistes de ces pays de s'abonner aux revues aux prix prohibitifs pratiqués par les très grands éditeurs. Des programmes de diffusion des revues scientifiques à des coûts modérés ou nul pour les pays du Sud existent[12] mais ils sont loin de permettre un accès équivalent à ce à quoi un chercheur d'un pays 'du Nord' peut prétendre via les abonnements de son institution. Malheureusement, Les archives ouvertes institutionnelles en Afrique sont encore minoritaires, et plus encore en Afrique francophone qu'en Afrique anglophone. (Rossi, 2018) [13]
Les professionnels de l'information de ces pays ont bien conscience du potentiel qu'offrent les ressources numériques en libre accès, mais cet accès est encore difficile. Ces obstacles sont principalement dus au sérieux manque de moyens auquel ces pays doivent faire face, moyens humains, le personnel étant trop limité, comme techniques, les infrastructures de communication étant pauvres et défaillantes.
Même lorsque les documentalistes parviennent à accéder aux notices, l'accès au texte intégral n'est pas garanti pour autant. (Rossi, 2018) [13]

Le progrès technologique a néanmoins permis, depuis les années 2000 de faire reposer l'accès à internet sur le réseau mobile, plutôt que sur un réseau filaire faible. (Vicart, 2015)**[18]**
La volonté de créer et diffuser de la connaissance prend tout son sens dans ce contexte. Rendre accessible les ressources de façon libre et gratuite est primordial pour ces pays.
L'initiative de Budapest pour l'accès ouvert recommande « *une mise à disposition gratuite sur l'Internet public, permettant à tout un chacun de lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces articles, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale, sans barrière financière, légale ou technique autre que celles indissociables de l'accès et l'utilisation d'Internet. La seule contrainte sur la reproduction et la distribution, et le seul rôle du copyright dans ce domaine devrait être de garantir aux auteurs un contrôle sur l'intégrité de leurs travaux et le droit à être correctement reconnus et cités* ».
L'UNESCO pour sa part a proclamé dans sa décision du 28 septembre 2015 proclame un droit universel d'accès à l'information et rappelle que « *La liberté d'information ou le droit à l'information fait partie intégrante du droit fondamental de la liberté d'expression* ».

L'IRD s'inscrit dans ce courant et œuvre pour que les professionnels IST des pays émergeants puissent s'approprier et maitriser les outils numériques afin d'avoir accès au savoir.
L'institut participe donc à des programmes de transfert de compétences destinées aux professionnels de l'information dans les pays émergeants pour pallier le déficit d'information

[12] Research4life, par exemple, est un programme offrant un accès gratuit ou à faible coût à un contenu en ligne universitaire et professionnel revu par des pairs pour les étudiants, chercheurs et scientifiques des pays en développement.

auquel doivent faire face les étudiants et chercheurs du Sud. En numérisant son fond papier d'autre part, il permet leur permet d'avoir accès à des ressources historiques.
La plateforme NumeriSud[13] en est en autre exemple. En plus d'un lien direct vers Horizon, ce site met aussi à disposition 150 000 documents de supports différents, ainsi que des espaces collaboratifs que des groupes de chercheurs ou d'étudiants peuvent utiliser pour animer des formations ou partager des contenus.
Des projets de formation à la numérisation sont aussi en cours, afin de mettre en accès libre des collections de documents. Grâce à l'outil libre Greenstone notamment, une suite de logiciels développée et distribuée par l'UNESCO, les partenaires du Sud peuvent ainsi construire leurs propres bibliothèques numériques. Greenstone permet en outre d'indexer les champs bibliographiques ainsi que le texte intégral. (Rossi, 2018) [13]

Toutefois, le problème que rencontrent ces partenaires n'est pas le stockage ou la numérisation, mais l'hébergement. Une bande passante et un service informatique conséquent sont indispensables pour rendre visibles sur internet ces ressources.
C'est pourquoi le site BEEP[14](Bibliothèques électroniques en partenariat) a été créé dans le cadre des projets de partenariats entre l'Afrique et l'IRD, dans le but d'offrir aux bibliothèques numériques créées dans ce contexte un service d'hébergement provisoire, en attendant une solution plus durable. (Rossi, 2018) [13]
Dans tous les cas, un portail ayant vocation à être consulté par des utilisateurs des pays du Sud doit proposer des contenus simples et facilement téléchargeables.

Cette caractéristique de l'IRD consistant à travailler avec les pays émergents doit être prise en compte lors de la refonte du logiciel hébergeant son archive ouverte. Cette mission est inhérente à l'IRD, et sa présence dans le monde doit être aussi assurée par un système de diffusion de la production scientifique consultable par tous, et utilisable par tous.

1.3.3 Présentation du SID de l'IRD

Un système d'information documentaire est *« un dispositif global d'accès à la documentation et à l'information multimédia, depuis un poste banalisé, sur place ou à distance, en intranet ou extranet et sur l'internet, selon une habilitation personnalisée pour chaque usager, construit selon un schéma d'intégration de ressources et de services, dans une totale compatibilité avec le Système global d'information […] ont le corollaire indispensable est une interface universelle et conviviale, dans le respect des normes (interopérabilité)* »[15]

Un portail documentaire fait partie d'un SID, il en constitue l'interface d'interrogation pour l'utilisateur. Il doit représenter un point unique d'entrée à toutes les bases de données qui se situent en arrière-plan.

La base Horizon est accessible via un portail internet, accessible à tous. Derrière l'adresse URL www.**documentation**.**ird**.frse trouve un système complexe et maitrisé, ainsi qu'une équipe de personnes gestionnaires et administratrices qui travaillent dans une même direction, le but de cette base étant d'organiser, publier, diffuser et gérer des ressources le plus largement possible.
Horizon est l'archive ouverte institutionnelle de l'IRD et s'inscrit dans la logique d'une diffusion de l'information scientifique et technique. Cette base une des plus anciennes

[13] https://numerisud.ird.fr/
[14] http://www.beep.ird.fr/cgi-bin/library.cgi
[15] La définition du SID par Marie-Thérèse Rebat consultable sur https://slideplayer.fr/slide/2660924/

archives ouvertes institutionnelles d'EPST. Sa particularité est de gérer à la fois un fonds papier et une AOI, tous deux accessibles par un point d'entrée unique.
La collecte de la production scientifique de l'IRD remonte aux années 1960, avec la création de la collection de référence à Bondy qui prend le nom de « fonds documentaire » en 1982.

En 1986 l'IRD entame l'informatisation du catalogue avec le logiciel Texto. Dix ans plus tard, en 1996, l'informatisation du fonds documentaire est lancée, accompagnée d'une grande campagne de numérisation du fonds papier existant.

En 2006, Texto est abandonné, et le logiciel Exlibris, de la société Cadic[16], est adopté. C'est un des seuls à l'époque capable de gérer la multi-exemplarisation. L'IRD assure en effet sa présence à l'étranger par des « représentations » : Certaines de ces représentations disposent d'un centre de documentation. Les représentations se situent à Cayenne, Nouméa, Abidjan, Niamey, Ouagadougou, et La Paz. Chaque centre de documentation possède ses collections et le catalogue est commun à tous. Ce catalogue est en ligne, et alimenté par les documentalistes depuis leur centre.

La même année la base bibliométrique est créée, permettant d'établir des indicateurs pertinents pour mesurer la production scientifique de l'IRD à partir d'Horizon.

Par ailleurs, afin de développer des fonctionnalités indispensables aux chercheurs pour la gestion de leurs publications (liste dynamiques, exports, rapports, etc.), le système Horizon s'enrichit d'une base MySQL qui est mise à jour quotidiennement à partir du catalogue géré par Cadic.

Enfin, en 2007 le portail HAL-IRD est ouvert.

- Ressources informationnelles :

Figure 2 : Ressources informationnelles d'Horizon

Le catalogue de ce fonds compte aujourd'hui 99 210 notices et référence 3 000 nouveaux documents chaque année.
Environ 80 000 de ces notices renvoient à un document en texte intégral téléchargeable au format PDF, dont presque 65 000 sont en libre accès.
La base est consultable par tous publics, mais seuls les chercheurs IRD peuvent y déposer leurs travaux.

[16] qui deviendra 'Cadic Intégrale' dans les années 2000

Les fonctionnalités proposées couvrent notamment le dépôt de documents, la gestion d'un thésaurus, le suivi des statistiques d'accès, les recherches simples en texte intégral sur l'ensemble des documents ; les recherches plus fines sur les titres, les auteurs, les mots-clefs, le téléchargement de documents au format PDF.
L'interrogation peut porter, au choix, sur les seules publications de l'IRD ou sur les publications *et* les acquisitions des centres de documentation de l'IRD.
Ces chiffres concernent les publications des chercheurs IRD. Si l'on ajoute la bibliographie d'acquisition, qui comprend aussi des ouvrages produits par des chercheurs non employés par l'IRD, le nombre de notices s'élève à 147 598 (chiffres au 14 novembre 2018). 7 500 documents sont téléchargés chaque jour[17].

Les ressources informationnelles gérées par le service IST sont nombreuses et diverses[18].On peut recenser les articles (publications scientifiques), les livres, les chapitres, les documents de travail (la littérature grise), les diplômes, thèses, mémoires, la communication de colloque, de congrès ou de séminaire, les posters, les manuels, les cours, les revues électroniques et papier, les cartes, les atlas, des pages Web (Scoop It), des photos, des vidéos.
Le format PDF a été choisi dès le départ, afin de garantir une pérennité.
Les notices peuvent être exportables aux formats EndNote, EndNote XML, BibTeX, txt.
Le format Dublin Core, enfin, va permettre à d'autres institutions de moissonner l'entrepôt de données de la base Horizon grâce à ces notices stockées dans un entrepôt OAI.

- Types de processus opérationnels

Le fonctionnement et l'alimentation d'Horizon repose sur des processus opérationnels à la fois complets et complexes qui vont permettre de découper les étapes de traitement d'une publication de chercheurs.
C'est le processus de récolte et de traitement des publications qui permet d'alimenter la base Horizon.
Ce processus a pour objectifs d'obtenir un recensement exhaustif des publications, d'assurer la diffusion et le partage des résultats de la recherche et de l'information scientifique, de rendre visible les publications sur internet, mais aussi de dégager des indicateurs permettant de quantifier la communication scientifique de l'IRD.

Horizon est alimenté via une veille sur le WoS d'un côté, et par le dépôt par les chercheurs eux-mêmes sur une boite mail.
Dans le premier cas les notices sont collectées automatiquement et dans le second elles sont créées manuellement. Elles seront ensuite enrichies, indexées, et dupliquées dans d'autres formats.

    - La collecte des notices

        - La collecte à partir du WoS : Rappelons tout d'abord les fonctions du WOS :

C'est en premier une base de références bibliographiques, qui opère un signalement des articles parus. On ne trouve pas le document, seulement sa notice. C'est une information structurée. Il n'y a donc pas d'indexation plein texte, mais une indexation sur la notice.

[17] Chiffre disponible sur la page d'accueil d'Horizon

[18] Nous n'aborderons pas ici des ressources électroniques externes (abonnements aux revues, bases de données accessibles via le bureau du chercheur IRD) ni des pages intranet et internet IST gérées également par le service.

C’est également un producteur des indicateurs de consultation. Le WOS classe les revues selon leur nombre de citation, c’est-à-dire chaque fois qu’un article est cité par un autre article. Ce classement génère ensuite l’Impact Factor de chaque revue, qui indique leur niveau de notoriété. Cet indicateur est géré par le JCR (Journal of Citation Reports) qui rassemble également d’autres indicateurs sur les revues.

La veille WoS est une collecte des publications qui se fait par une requête rédigée et corrigée si nécessaire par le service IST. Cette collecte est automatique, mais nécessite toutefois un travail humain.
L'enjeu de cette fonction est très important, car le *ranking* et l'ensemble des indicateurs vont déterminer l’impact factor qui va ensuite servir à l’évaluation des travaux de recherche.
L’objectif de la veille réalisée sur le WOS est de savoir ce que publient les chercheurs IRD. On utilise cette méthode plutôt que de demander aux chercheurs d'envoyer leurs PDF au service IST, car ce n’est pas fait systématiquement.

La collecte à partir du WoS permet de récolter jusqu’à 350 notices par mois. Ces notices sont exportées au format EndNote avant d'être importées dans Cadic intégrale.
Cette veille représente la moitié des notices enregistrées, les autres notices provenant de dépôts manuels.
Ce processus consiste tout d’abord à formuler l’équation de recherche permettant de récupérer les notices des publications des chercheurs IRD. L’équation doit prendre en compte toutes les formes possibles qu’utilisent les chercheurs pour décrire l’unité mixte de recherche à laquelle ils sont affiliés, c’est pourquoi c’est une équation longue (37 lignes).
Cette étape est importante, car les résultats de la bibliométrie dépendent des signatures d'articles. Ils peuvent être faussés si l'auteur cite l'IRD mais oublie de citer une UMR. Si la signature ne mentionne pas l’UMR, la publication ne lui est pas attribuée

Il faut ensuite passer en revue chaque notice pour vérifier si l’affiliation est correcte. Puis ces notices sont importées, et une opération automatisée de dédoublonnage s’ensuit pour marquer les notices déjà présentes dans Horizon.

L’indexation manuelle de la notice se fait à l’étape suivante. Les notices bibliographiques sont enrichies d’une indexation thématique**,** éventuellement une sous-thématique**,** et géographique, utile par la suite à la fois aux études bibliométriques et aux recherches d’information dans la base Horizon.

L’étape suivante consiste à définir les droits d’auteur attachés à la publication, grâce à l’outil public SHERPA/RoMEO[19](Publisher copyright policies & self-archiving), les documentalistes peuvent différencier les documents autorisés à être mis en ligne en libre accès de ceux pour lesquels les éditeurs ne permettent pas la diffusion publique.
Les documents autorisés à être mis en ligne en libre accès sont systématiquement déposés en accès ouvert sur internet dans la base Horizon, en début de chaque mois.
Jusqu’en 2015 tout était déposé en intranet, avant que ne commencent les campagnes de mise en accès libre des PDF libres de droit.
Les fichiers soumis à droits d’auteur quant à eux sont envoyés en accès réservé sur l’intranet quotidiennement, et une authentification sera nécessaire pour les consulter.

Le texte intégral est recherché par les documentalistes via les abonnements aux revues, à environ 90%. Les 10% restants sont récupérés en ayant recours à des demandes aux auteurs, au réseau de collègues documentalistes ou à l’utilisation de réseaux sociaux.

---

[19] http://www.sherpa.ac.uk/romeo/index.php

L’étape de recherche du texte est primordiale car elle permet d’assurer une des spécificités fortes de l’IRD qui consiste à garantir l’accès au texte intégral des publications de ses chercheurs déposées dans la base Horizon. Horizon ne se veut pas seulement une base de signalisation, mais une bibliothèque dans laquelle on trouve le document.
Tous les PDF sont imprimés et déposés ensuite au FDI, le fonds papier localisé à la délégation régionale de Bondy servant ainsi d’archive durable et pérenne

Les notices sont ensuite injectées dans CADIC Intégrale via une procédure d’import réalisée grâce au module ExImport du logiciel.

Ce processus est schématisé ci-dessous

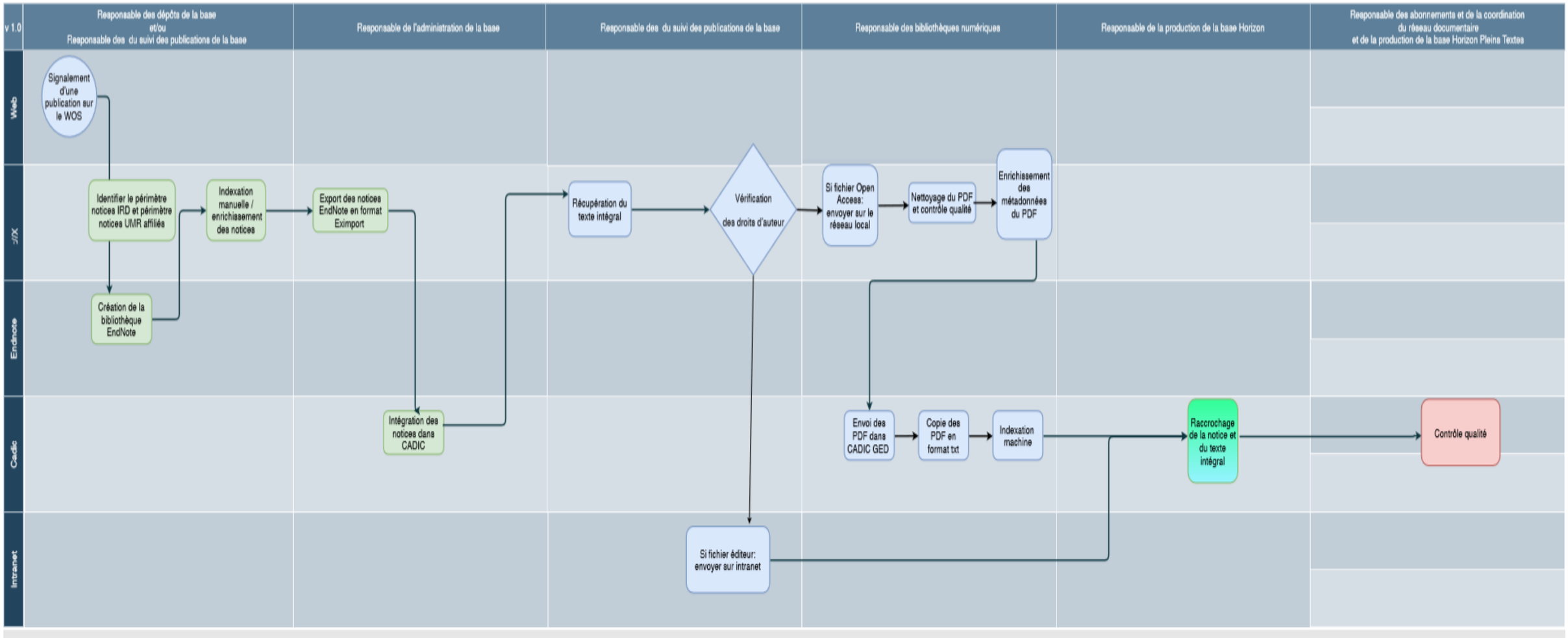


Figure 3 : Collecte des notices à partir du WoS

- Le dépôt par mail est la seconde source d'alimentation de la base Horizon.

Contrairement à la veille WOS qui permet de récolter uniquement des notices, ce dépôt sur boite mail permet de collecter des documents en texte intégral à partir desquels sont créées les notices. Parfois, les chercheurs n'envoient que la référence bibliographique, ou même qu'un signalement, cela implique d'aller récupérer le texte intégral (achat, prêt entre bibliothèques…).
De plus, ce dépôt contribue à garantir l'exhaustivité des publications collectées, notamment dans le domaine des SHS.
Les notices issues de ce dépôt par mail sont saisies manuellement dans CADIC Intégrale.

- Toutes les notices sont ensuite copiées quotidiennement dans la base MySQL grâce à un script PHP.

MySQL envoie un message au chercheur pour lui notifier qu'une de ses publications a été référencée dans Horizon.
Enfin, toutes les notices subissent un contrôle qualité, en plus de toutes les différentes opérations liées à leur enrichissement.

- La base Horizon est aussi alimentée par la voie de la numérisation.

Tous les documents numérisés dans ce cas de figure doivent être déjà référencés dans Horizon.
Ce processus consiste à numériser le rétrospectif (le fonds papier), c'est-à-dire les ouvrages acquis depuis la création de l'IRD.
Il résulte des étapes de ce processus une chaine de numérisation très formatée. Les enjeux de stockage et d'indexation sont importants, les premiers permettent au document de ne pas occuper trop de place, les seconds vont permettre aux documents et faire l'objet de fouille de texte via le text and data mining[20].

- La dernière étape de ces trois modes d'alimentation consiste à indexer les PDF qui seront exposés en libre accès.

Cette opération est automatique et permet une indexation sur l'ensemble du texte, ainsi que sur sa notice.
Cette indexation est capable de repérer les occurrences dans le texte et permet d'extraire les métadonnées liées au titre, à l'auteur, la date du document, son code de classement, et des mots-clés.
A l'issue de cette opération Les PDF sont visibles sur le web ce qui signifie que Google y a accès et peut venir à son tour les indexer et les rendre ainsi visibles sur internet.
Google permet d'envoyer directement l'internaute sur l'URL du PDF sans le faire passer sur la page d'accueil et l'interface CADIC.
L'indexation Google est optimisée grâce à la déclaration de *Sitemap*[21] pour le référencement des PDF.

- Le travail de bibliométrie peut ensuite se faire grâce à ces minutieuses opérations successives.

---

[20] Le TDM, ou *exploration de données* consiste à extraire automatiquement des informations à partir de grandes quantités de données. Nous abordons ce concept dans la partie 2.
[21] fichier XML qui permet d'indiquer aux robots d'indexation des moteurs de recherche une liste des URL à indexer pour un site donné.

- Les acteurs du SID :

Les acteurs impliqués dans le projet sont l'équipe IST et la direction informatique (DDUNI).
Les missions générales de la DDUNI consistent à assurer des services support aux utilisateurs, prestataires, délégations régionales et aux représentations, à garantir qualité, performance et valorisation, offrir un appui au développement de projets innovants, et garantir l'exploitation, les infrastructures informatiques, et la sécurité du système d'information dans son ensemble.

1.3.4 Le système d'information documentaire derrière Horizon: analyse technique et fonctionnelle du système existant

La base Horizon se présente à l'internaute sous forme d'un portail derrière lequel se trouvent deux applications, CADIC Intégrale et une base MySQL, toutes deux hébergées par le service informatique (la DDUNI). Nous les décrivons dans le tableau ci-dessous.

Portail	Description	Administration, alimentation et public	Documents et ressources stockés
Base Horizon Pleins Textes (vue CADIC) : **http://horizon.documentation.ird.fr**	- Archive ouverte institutionnelle IRD / Gestion du fonds documentaire IRD (FDI) - Gestion du catalogue et des fonds d'acquisition des centres de doc	- Base reposant sur le SIGB CADIC - Fourniture, hébergement sécurisé et maintien en condition opérationnelle (maintenance corrective et évolutive) : direction informatique - Administré et alimenté par l'IST - Public cible : chercheurs et grand public	PDF en libre accès + toutes les notices: - Articles de revue - Ouvrage - Chapitre ou partie d'ouvrage - Colloque, congrès, séminaire - Diplôme, thèse, mémoire - Littérature grise - Notices
Base Horizon (vue MySQL) : **http://www.documentation.ird.fr**	- Url dynamiques permettant de faire remonter automatiquement des listes de publication - URL pérennes pour chaque notice - Services Web (flux RSS, formats d'export des bibliographies, bulletin mensuel de veille, entrepôts OAI)	- PHP-MySQL - Administré par l'IST - Alimenté par CADIC - Public cible : chercheurs IRD (authentification demandée)	Notices bibliographiques (copie de la base Horizon)

Figure 4 : Tableau descriptif de la base Horizon

Horizon est donc géré par un système d'information composé de deux applications: **le SIGB Cadic Intégrale (version 11) et MySQL.**

Le catalogage, l'exemplarisation, le stockage et l'indexation, la recherche des documents reposent sur le Système Intégré de Gestion de Bibliothèque CADIC Intégrale, mais certaines fonctionnalités propres à la gestion d'une AOI ne sont possibles que via une copie de la base Horizon CADIC sur une base gérée par MySQL.

Le moteur de recherche du logiciel Cadic Intégrale permet toutes les requêtes sur la page Horizon, dans les notices et dans les PDF.

Une base miroir des publications la base MySQL permet de proposer des fonctionnalités plus avancées, comme l'affichage de listes à partir de requêtes implicites gérées en PHP pour faire remonter toutes les publications par le biais d'une requête lancée directement dans l'URL.
Cette base de données relationnelle permet de gérer des données qui ne sont pas dans CADIC, mais de les croiser avec les données de CADIC, comme par exemple les affiliations aux UMRS gérées dans l'annuaire
Des scripts PHP permettent d'interroger cette base pour en faire ressortir des données intéressantes pour les chercheurs, comme les listes de leurs publications en période d'évaluation. Chaque service correspond à un script PHP. La liste complète des services sera développée plus bas.

Cette dualité n'est pas visible pour l'utilisateur qui n'a à faire qu'à une seule interface, graphique, qu'il soit sur la page d'accueil d'Horizon ou sur les listes de publications, par exemple.

→http://horizon.documentation.ird.fr/exl-php/cadcgp.php?CMD=CHERCHE&query=1&MODELE=vues/horizon/accueil.html&AUTH=1

→http://www.documentation.ird.fr/hor/FROMENT,ALAIN/tout

Le tableau ci-dessous dresse une liste des fonctionnalités que propose le SID existant, en attribuant chacune à la base la rendant possible. Certaines sont utiles aux chercheurs, et d'autres aux administrateurs de la base, surtout pour son alimentation.

Thématique de fonctionnalité	Nom de la fonctionnalité	A quelle application la fonctionnalité est-elle rattachée	Bénéficiaire de la fonctionnalité
Référencement Google	Protocole Sitemap	CADIC	Documentalistes
Fonction indexation et référencement	Indexation des listes dynamiques	Non	Documentalistes
	Indexation du texte intégral des PDF	CADIC	Documentalistes
	Indexation des notices	CADIC	Documentalistes
	Gestion d'un thesaurus	CADIC	Documentalistes
	URL pérenne des PDF	MySQL	Documentalistes
	Permalien des notices	MySQL	Documentalistes
Fonction authentification des usagers	Service SSO	Indépendant	Chercheurs
	Accès différencié	CADIC	Documentalistes
Fonction notices	Exporter des notices au format: XML TXT Zotero	MySQL	Documentalistes
	Export des métadonnées des notices CADIC en Dublin Core, Bibtex et EndNote	MySQL	Chercheurs
	Import des notices Endnote issues du WOS avec le module Eximport	CADIC	Documentalistes
	Gestion des affiliations des notices du WOS	MySQL	Documentalistes
	Repérage des doublons	MySQL	Documentalistes
Fonction statistiques	Tableaux d'indicateurs	MySQL	Chercheurs
Fonction veille/fidélisation	Flux RSS	MySQL	Chercheurs
	Bulletin de veille WOS	MySQL	Chercheurs
	Bulletin des acquisitions	CADIC	Chercheurs
	Envoi de mail automatique	PC Bondy	Chercheurs

Fonction identification des publications	N° FDI	CADIC	Documentalistes
Fonction bibliométrie	Suivi des statistiques d'accès	Log Apache de CADIC	Documentalistes
	Indicateurs bibliométriques	MySQL	Documentalistes / Direction de l'institution
Fonction workflow	Gestion des workflows	CADIC	Documentalistes
Fonction SIGB classique	Catalogage	CADIC	Documentalistes
	Téléchargement de documents au format PDF si en libre accés	CADIC	Documentalistes
Fonction recherche	Recherches simples en texte intégral sur l'ensemble des documents	CADIC	Chercheurs
	Recherches avancées sur les titres, les auteurs, les mots-clefs	CADIC	Chercheurs
	Créer des bibliographies dynamiques par nom, auteur IRD, unités Mixtes de Recherche, N° FDI, année	MySQL	Chercheurs

Figure 5 : Liste des fonctionnalités d'Horizon

Le schéma ci-dessous reproduit de façon simplifiée l'architecture technique du SID.

Une représentation plus détaillée de l'architecture du système a été documentée par l'ancien directeur du service IST, Dominique Cavet, disponible en annexe 1.

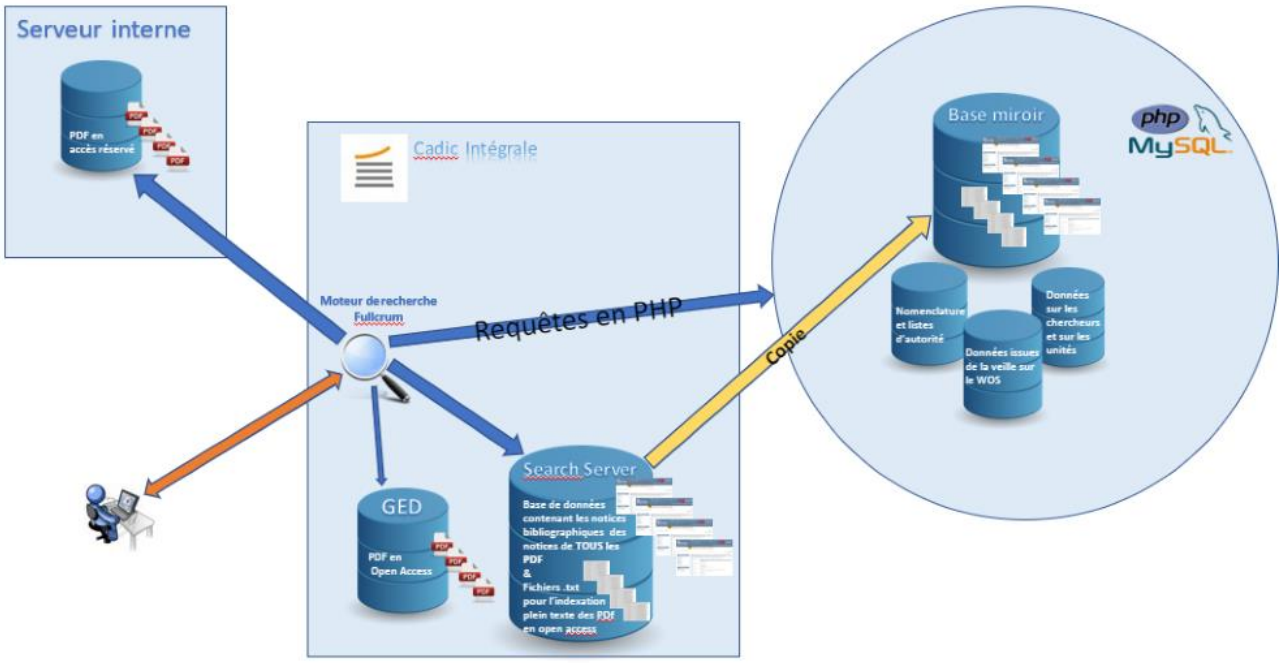


Figure 6 : Vue simplifiée de l'architecture technique d'Horizon

1.3.5 l'évolution souhaitée du SID

Ce système présente toutefois ses limites. Le SIGB a été enrichi de briques et decouches successives, développées en interne.
Ce système a, certes, su suivre les progrès du Web, ainsi que les évolutions et les contraintes du domaine de l'IST, qui est performant et qui rend de précieux services, mais également complexe, et dont la sécurité et la pérennité sont remises en question aujourd'hui.
La personne ayant mis en place ce système est depuis partie à la retraite en ne laissant que très peu de documentation technique.

Aujourd'hui, la DDUNI souhaite donc mettre en place un système plus solide et fiable.
Ses motivations sont, d'une part, de sécuriser ce système que le manque de documentation technique rend fragile, et de mettre en place un outil clé en main.

Toutes les solutions sont envisageables : logiciels propriétaires ou libres, optimisation du système de GED Alfresco déjà en place dans le reste de l'institut, ou encore la possibilité de confier la base Horizon Pleins Textes à Hal et centraliser ainsi le stockage.

La capacité du système de gérer à la fois un fond papier et une archive ouverte ne sera pas remise en cause, mais de nouveaux critères viennent s'ajouter, afin de satisfaire les exigences des informaticiens d'une part, et les nouveaux enjeux du monde de la recherche scientifique d'autre part.

Le service IST souhaite de son côté un outil plus performant, capable d'évolution.

Les faiblesses du système actuel reposent en effet sur un logiciel qui n'évolue plus assez vite par rapport aux exigences du monde de l'IST aujourd'hui. Son moteur de recherche par exemple n'est pas très performant comparé à ceux qu'offrent le marché actuel.

Toutefois d'autres finalités rentrent en ligne de compte. L'IRD, et en particulier les documentalistes aimeraient un outil plus moderne, connecté avec les outils IST, comme l'outil ORCID qui permet aux chercheurs de déposer leurs publications sur une plateforme ou encore pouvoir implémenter un module de versement dans HAL.

Le marché offre aujourd'hui des outils bien plus performants que ceux proposés à l'époque ou CADIC Intégrale a été choisi, et l'IRD pourrait s'équiper d'un outil métier qui comprendrait la dualité de son système (gérer à la fois un fond papier et une AOI) et permettrait de gérer une base de données complexe et capable de produire des métadonnées riches.
L'enjeu est d'exposer des notices de haute qualité, de garantir ainsi une visibilité à la production scientifique de l'IRD, et d'assurer le moissonnage des publications des chercheurs.

La solution pourra être libre et suivre la tendance encouragée par le monde de la recherche. Elle pourra aussi être propriétaire, l'offre des SIGB éditeurs s'étant largement développée ces dernières années, et proposant plus de fonctionnalités spécifiques aux bibliothèques que les premières.
Enfin, bien sûr, la solution choisie devra impérativement répondre aux enjeux de libre accès et continuer de mettre en avant le texte intégral de la publication, des critères qui sont chers à l'IRD.
Nous aborderons dans la deuxième partie les outils et projets IST en cours à l'heure actuelle qui vont influer sur le choix d'une nouvelle solution.

Pour finir, le nouveau SID devra évoluer avec les engagements qu'à l'IRD envers les pays du Sud. Il existe une contrainte forte de rendre accessible ces ressources à ces pays.
L'institut cherche en effet à optimiser l'accès des chercheurs des pays émergeants aux connaissances scientifiques, et est fortement engagé dans des projets de mutualisation de moyens technologiques destinés à faciliter la recherche.
Cette mission ne peut pas être ignorée au moment de choisir un logiciel qui a pour tâche de diffuser de la connaissance scientifique le plus largement possible.
Certains logiciels étant plus ou moins fermés, ou accessibles, il sera souhaitable de mesurer leur accessibilité dans ce contexte.

Nous avons pu dans cette première partie faire le constat selon lequel l'IST est devenu un domaine vaste et complexe, en pleine mutation et qui se construit à un niveau national et international.
La législation dans le monde entier va dans le sens de la valorisation et de la diffusion des résultats de la science. Les archives ouvertes ont un rôle important à jouer dans ce mécanisme, d'où la dimension forte d'un SID capable d'étendre la portée de ces archives. C'est en effet ce système qui garantit la visibilité et la notoriété d'une institution.

Cette description de l'IST aujourd'hui permet également de mettre en perspective le rôle des pays bénéficiant des grandes avancées technologiques, et de montrer comment les évolutions de l'IST, notamment le libre accès peuvent aider à pallier ces différences.
L'alliance de l'IRD avec les pays émergents est à prendre en compte dans sa recherche d'un nouvel outil pour exposer sa production.
Toutefois, les particularités liées à l'IRD ne sont pas les seules à devoir être intégrées dans cette réflexion.

Nous allons aborder dans une deuxième partie les nouveaux objectifs de l'IST.

Partie 2: Les enjeux actuels du monde de l’IST pour penser un SID

Les nouvelles tendances et évolutions de l'IST s'inscrivent dans un contexte de science ouverte. Libre accès, évaluation, édition scientifique, Web de données, Web sémantique, sont autant d'enjeux dont il faut tenir compte lors de la mise en place d'une archive ouverte, ou, comme dans notre cas, d'un projet visant à faire évoluer une archive ouverte d'un institut de recherche.

Si les institutions politiques encouragent ces différents développements, les acteurs de l'IST accomplissent un travail considérable pour les mettre en place à un niveau plus local. De nombreuses initiatives sont ainsi lancées pour offrir aux chercheurs la possibilité de bénéficier de ces avancées.

L'IRD doit en prime envisager ces évolutions dans le cadre d'une de ses missions spécifiques, qui est de rendre la science accessible aux pays en développement.

Nous exposerons ces sujets dans cette partie, en tentant de les replacer dans le contexte du projet en cours, consistant à choisir un outil de gestion documentaire pour une archive ouverte, et montrer comment la prise en compte de ces enjeux actuels est cruciale pour développer un SID performant dans le futur ;

2.1 Enjeux d'Open Access

Le mouvement du libre accès ou open access consiste à mettre à disposition, en ligne, du contenu. Ce contenu peut être libre de droits ou soumis à un régime de propriété intellectuelle. Il repose sur les principes d'accessibilité de toute la littérature scientifique, de gratuité, et de pérennité, en proposant des formats et des protocoles communs pour garantir l'interopérabilité.
Dans le cas de la production scientifique, ce principe a vocation à mettre en ligne gratuitement les publications scientifiques produites par les chercheurs et issues de fonds publics.

Le libre accès commence au début des années 90 avec la création de l'archive ouverte de prépublications ArXiv, destinée à la production scientifique des physiciens théoriciens.
Il s'est développé à travers la création d'archives ouvertes, de déclarations, d'incitations politiques, et par le développement de nouveaux modes de publication.
La communauté des développeurs élabore de son côté des moyens de mettre techniquement en œuvre ce mouvement. En 1999, l'Open Archive Initiative conçoit ainsi le protocole OAI-PMH, qui permet de rendre interopérable les réservoirs d'archives ouvertes entre eux (voire partie 1).

Trois ans plus tard, l'Initiative de Budapest[22]en 2002 recommande deux stratégies complémentaires: la première consiste à encourager les chercheurs à déposer leurs articles en auto-archivage dans des archives électroniques ouvertes. Ces articles peuvent être sous la forme de prépublications (la version soumise par un auteur avant la revue par ses pairs, ou peer-reviewing) et sous la forme de post-publications (après la revue par les pairs, et acceptation par le comité éditorial).
La seconde stratégie consiste à encourager la création de nouvelles revues engagées dans le libre accès ou d'accompagner des revues existantes qui choisissent de s'orienter vers cette voie.

La Déclaration de Bethesda[23], un an plus tard, donne une définition de la publication en libre accès et attribue des rôles aux différents acteurs de la communication scientifique.

La Déclaration de Berlin[24], enfin, signée en 2003, définit «*le libre accès comme une source universelle de la connaissance humaine et du patrimoine culturel ayant recueilli l'approbation de la communauté scientifique* » et insiste sur le rôle du Web dans la mise à disposition de ces connaissances.

Le mouvement est encadré par des recommandations, plus que par des obligations. Les instituts et universités relaient ces recommandations par des politiques incitatives ou coercitives de dépôt, qui seront plus ou moins suivies par les chercheurs.
La France s'inscrit dans ce mouvement, par l'action du CNRS, dont une de ses unités, le Centre pour la Communication Scientifique Directe (CCSD) lance des plateformes d'archives ouvertes et notamment Hal (Hyper article en ligne) créée en 2001, une archive interdisciplinaire accueillant des articles publiés ou non.

Historiquement, l'IRD s'est rapidement prononcé en faveur du mouvement du libre accès et a été parmi les premiers instituts de recherche européens signataires de la Déclaration de Berlin. Son action se poursuit aujourd'hui, et l'année dernière l'IRD a également signé l'appel

[22] https://www.budapestopenaccessinitiative.org/read
[23] http://legacy.earlham.edu/~peters/fos/bethesda.htm
[24] https://openaccess.mpg.de/Berlin-Declaration

de Jussieu[25] pour une science ouverte qui plaide en faveur d'un modèle de publications diversifié et en libre accès.
Par ailleurs, l'institut s'est engagé à augmenter son taux de publications en libre accès de 25% dans son archive ouverte d'ici 2020. Le service IST est moteur de cette politique et a commencé une campagne de mise en accès libre de publications qui consiste à mettre en libre accès des articles jusqu'alors en accès réservé.
Sa politique de dépôt dans Hal n'oblige pas les chercheurs à verser leurs publications dans Hal. Toutefois, un module peut être implémenté sur l'outil gérant son AOI pour entrainer le versement automatique d'une publication dans Hal à partir d'Horizon. L'IRD envisage d'ailleurs de mettre en place dès que possible le reversement automatique dans HAL de toutes les notices créées dans Horizon.

Le mouvement de libre accès s'est concrétisé par l'ouverture d'archives ouvertes, institutionnelles lorsqu'il s'agit de plateformes créées et gérées par les organismes de recherches. Aussi est-il important de penser aux outils techniques derrière ces plateformes.
Nous verrons dans la troisième partie qu'il existe aussi des dispositifs permettant de rendre plus visibles les publications. Le protocole SWORD, par exemple, permettra le transfert d'un document vers des archives centrales ou thématiques, comme Hal ou PubMed. Une politique institutionnelle en faveur du libre accès doit accompagner son AOI de moyens techniques qui permettront d'accroitre la visibilité de sa production scientifique.
La mise en place de services et d'actions autour des archives contribuent à la valorisation des connaissances, il faut par conséquent bien définir ces services et actions que l'institut voudra mettre en place lors de la mise en place de son AOI. De même, la politique documentaire établira le type de publications que l'archive ouverte acceptera. Certaines AOI aujourd'hui n'acceptent que des articles postprint, alors que d'autres tolèrent également des versions preprint. Les agence Allenvi[26] et Aviesan[27] considèrent désormais le preprint comme une forme acceptable de communication scientifique. Cette stratégie est à définir lors de la mise en place d'un outil.

Les avancées technologiques permettent ainsi de favoriser le mouvement du libre accès, et ce mouvement est étroitement lié au progrès de l'informatique.
L'Union internationale des télécommunications (UIT) est une agence de l'ONU dont la mission est de démontrer l'importance des technologies de l'information et de la communication dans la lutte contre les inégalités, après le constat d'un retard de l'Afrique notamment dans l'accès à internet. Consciente de la fracture numérique que connaissent les pays émergeants, l'Union souhaite élargir les bienfaits de la révolution numérique à tous.
Dans sa déclaration de principes du 12 décembre 2003, l'Union affirme que les « *TIC devraient être considérées comme un moyen, et non comme une fin en soi. Dans des conditions favorables, elles peuvent être un puissant outil, accroissant la productivité, stimulant la croissance économique, favorisant la création d'emplois et l'employabilité et améliorant la qualité de vie de tous. Elles peuvent par ailleurs contribuer au dialogue entre les personnes, les nations et les civilisations* »[28].
Le principe du libre accès repose en effet sur des télécommunications performantes, mais nous avons vu, dans la première partie, que les pays émergents ont plus de difficultés à accéder aux bases de connaissances mises en ligne sur le Web. Le libre accès contribue à diminuer les différences d'accès à l'information, mais encore faut-il pouvoir techniquement y

[25] https://jussieucall.org/
[26] Alliance nationale de recherche pour l'environnement dont la mission est de coordonner les recherches françaises.
[27] Alliance nationale pour les sciences de la vie et de la santé dont la mission est d'assurer une coordination scientifique des grandes thématiques de recherche dans ces domaines.
[28] http://openaccess.inist.fr/?Declaration-de-principes-du-Sommet

accéder. Le libre accès à la connaissance passe donc nécessairement par un accès aux plateformes numériques et donc à la technologie.
Or, la demande pour des publications en libre accès est grande dans les pays en développement. Leur production scientifique est conséquente : l'Inde, la Chine, le Brésil, la Russie et l'Afrique du Sud représentent à eux seuls 24% de la production scientifique, publient 16% de l'ensemble des revues en libre accès. Ces cinq pays n'hébergent pourtant que 10% des archives ouvertes. (Schöpfel, 2017)[15]
Pour aider à résorber ces défaillances, le programme research4life, cité plus haut, a été lancé pour améliorer les conditions de recherche des scientifiques et étudiants des pays en développement à travers des plateformes gratuites à faible coût mettant en ligne du contenu approuvé par la communauté scientifique.
D'autres plateformes ont été montées afin de permettre la diffusion d'articles sous la forme de working papers considérés comme des publications dont la plateforme RePEc [29], par exemple, fait partie.
Certains logiciels, nous le verrons permettent le versement automatique des publications vers ces plateformes, et favorisent ainsi la visibilité de la production de l'organisme de recherche.
De son côté l'IRD œuvre également pour que ses partenaires du Sud puissent avoir accès aux publications. Aux côtés d'EPRIST[30], l'association des responsables des services d'information scientifique et technique des organismes de recherche, le personnel IST de l'institut mène des négociations et agit pour l'intérêt de ses partenaires en défendant le principe de gratuité pour la publication par les scientifiques des pays du Sud dans les revues auteur-payeur.
En outre, le libre accès facilite la conduite de projets liés à l'IST, projets que nous développerons dans cette partie. Le libre accès est par exemple un outil très utile de bibliométrie. Dès le début du mouvement, un informaticien a montré que les articles en libre accès étaient plus cités (Lawrence, 2001)[21]. Dans certaines sciences, un article peut être cité deux fois plus. (Tennant, 2016)[23]
Le libre accès favorise également la fouille de texte, en mettant à disposition plus de littérature scientifique, les revues Open Access mettant à disposition des chercheurs des moyens techniques afin que leurs publications puissent être fouillées efficacement. (Tennant, 2016)[23]

Le libre accès continue de progresser à travers le monde grâce à des projets et initiatives issus des financeurs publics: SCIELO[31] est une base de données bibliographiques et une bibliothèque numérique fonctionnant sur un modèle de publication électronique coopérative de revues à libre accès.
Des actions de sensibilisation, de communication et de formation sont régulièrement lancées par le personnel IST de l'ensemble des instituts de recherche afin d'accompagner les chercheurs sur les questions juridiques et techniques. Le service IST de l'IRD mène beaucoup d'actions dans ce sens.
Une des particularités de l'IRD est d'être pluridisciplinaire, l'institut expose des publications dans le domaine des STM, mais aussi en SHS. Or, le libre accès dans le domaine des sciences humaines et sociales est légèrement différent.
L'offre de revues en SHS est en effet plus fragmentée, et bénéficie d'un nombre moins important d'abonnés, contrairement aux revues spécialisées en STM. Leur support est encore attaché au papier. Les articles SHS sont par ailleurs plus souvent publiés dans la langue du chercheur, réduisant la part de lecteurs potentiels, contrairement aux STM qui

[29] RePEc est un projet collaboratif de plusieurs centaines de volontaires de 79 pays destiné à permettre la plus large diffusion de la recherche en économie.
[30] EPRIST est l'association des responsables IST des organismes de recherche.
[31] http://www.scielo.org/php/index.php

sont quasiment toutes accessibles en anglais. En outre les SHS sont plus largement diffusées par la voie de la monographie, et ne bénéficient donc pas de la visibilité offerte par les revues scientifiques. (Chartron, 2014)[25]
Des publications en SHS exposées sur une AOI doivent par conséquent bénéficier d'une bonne visibilité. Les outils disponibles sur le marché aujourd'hui permettent de verser automatiquement les publications sur des plateformes spécialisées afin d'augmenter leur visibilité, et ce critère sera sans doute à prendre en compte pour une archive pluridisciplinaire.
Cet aspect est important à l'heure où les institutions encouragent le croisement des disciplines. Le programme cadre de recherche et de développement, lancé par l'union européenne, et dont Horizon 2020[32] fait partie, prône une science ouverte, innovante et interdisciplinaire.

Le libre accès est un moyen récent de publier mais la voie de la publication classique reste avant tout l'édition scientifique, que nous allons aborder maintenant.

2.2 Problématique d'édition scientifique

La publication scientifique a toujours été le moyen de diffuser les travaux scientifiques des chercheurs et d'en faire connaitre les résultats par leurs pairs. C'est au XVIIème siècle qu'apparaissent les premières revues scientifiques, jusque-là la correspondance entre les savants était le seul moyen de transmettre ces connaissances. Les revues permettent donc de véhiculer à travers le monde les découvertes scientifiques, grâce notamment à un système d'évaluation des chercheurs par d'autres chercheurs, qui aide à estimer et valider le travail du scientifique.
Le nombre de ces revues n'a cessé d'augmenter, et ce phénomène s'est encore accentué depuis l'avènement d'internet. Cette tendance est en outre accompagnée d'une forte croissance des prix d'abonnement pratiqués par les plus grands éditeurs, que les universités et instituts de recherche modestes, notamment des pays émergents, n'ont pas pu suivre. Les résultats de la science n'étaient donc plus accessibles à tous.

C'est à la suite de ce constat qu'est né le mouvement du libre accès est né dans les années 90. Au même moment, l'explosion des TIC permet au monde de la recherche de développer ses propres outils pour favoriser la diffusion de l'information. Les versions preprint constituent alors les premières archives ouvertes. Ces versions n'avaient pas vocation à remplacer les revues scientifiques. Mais la conjoncture pousse le monde de l'IST à penser de nouvelles voies de publication. Ce domaine est depuis en pleine mutation.
De nouveaux modèles de publication issus du mouvement du libre accès apparaissent, que nous allons énumérer ci-dessous[33].

- Le Free Gold Open Access, d'abord, ou « voie dorée gratuite » est un modèle entièrement gratuit qui repose sur le financement par les institutions de recherche.
- La Platinum Road, ou « accès platinum » est un modèle Freemium, qui propose du contenu gratuit et des services plus avancés mais payants.
- Le modèle hybride est adopté lorsque la revue diffuse ses numéros récents par la voie de l'abonnement, qui seront ensuite exposés gratuitement après six à vingt-quatre mois d'embargo.

[32] Texte complet consultable sur le site EUR-Lex

[33] Les définitions suivantes sont issues de l'article de Lise Verlaet **[31]**

- La Green Open Access, ou « voie verte », est la voie de l'auto-archivage par l'auteur. Dans ce cas, certains éditeurs n'autorisent toutefois l'archivage de la version éditeur qu'après une période d'embargo. La liste de ces éditeurs est disponible sur le site Sherpa/Romeo.

Ces périodes d'embargo font l'objet de recommandations par les pays. La commission européenne recommande dans son programme Horizon 2020 des délais maximums d'embargos de 6 mois en STM et de 12 mois en SHS. Ce plan comporte l'obligation pour les pays de l'Union européenne d'atteindre l'objectif de 100%depublicationsissues de recherches financées sur fond public disponible en libre-accès en 2020sous peine de sanctions financières.

Nous avons vu dans la première partie que la législation française autorise les chercheurs, depuis la loi pour une république numérique, à auto-archiver la version manuscrite de leur article dès lors qu'elles sont issues d'une activité de recherche financée au moins pour moitié par des fonds publics.
Ces notions sont à considérer dans la recherche d'un nouvel outil de gestion d'archive ouverte institutionnelle. La plupart des logiciels aujourd'hui permettent de prendre en compte ces délais d'embargos et de projeter automatiquement la publication en libre accès dès ce laps de temps atteint. Il est également possible d'intégrer une API Sherpa/Romeo qui définit systématiquement les conditions d'archivage d'un document.
La difficulté pour les documentalistes réside dans le suivi de ces politiques pour se mettre à jour afin d'informer et d'accompagner correctement les chercheurs.

La gestion des droits d'auteur implique aussi que le logiciel puisse gérer des accès différenciés selon le statut d'une publication, puisque le texte intégral d'un article sous embargo ne devra être visible que par les chercheurs de l'institut.
Les grands éditeurs ont réagi en créant le modèle économique du Gold Open Acces ou « voie dorée » qui repose sur le modèle de l'auteur-payeur et sur l'Article Processing Charge (APC) qui correspondent à des frais de traitement de l'article.

Nous avons indiqué que l'inflation des coûts de certaines revues STM publiées par les grands éditeurs internationaux est à l'origine du mouvement du libre accès.
Toutefois ce phénomène ne prend pas en compte les éditeurs plus modestes, et notamment des éditeurs de revues SHS, dont les coûts d'abonnement ne sont pas aussi prohibitifs. (Chartron, 2010)[26]
Le débat tend en effet à se focaliser uniquement sur les dérives des grands groupes internationaux de l'édition en STM et oublier la fragilité des modèles économiques des plus petits éditeurs. (Chartron, 2014)[25]
Or l'édition est le modèle historique de diffusion des connaissances scientifiques. Sa force est dans la diversité des éditeurs eux-mêmes qui peuvent être issus du secteur privé (laboratoire de recherche, éditeurs privés à but commercial par exemple) et public (presses universitaires, instituts de recherche, comme l'IRD qui a sa propre édition).
Une réflexion est en cours aujourd'hui pour adapter le modèle historique de publication scientifique aux progrès technologiques.
La lettre ouverte de Public Library of Science[34] (PLoS), a été écrite dans ce sens par des chercheurs prenant position pour « *la création d'une bibliothèque publique en ligne qui fournirait le contenu intégral des résultats publiés de la recherche et des textes scientifiques dans le domaine de la médecine et des sciences du vivant sous une forme en libre accès* » tout en reconnaissant le droit aux éditeurs à une rémunération équitable.

[34] http://openaccess.inist.fr/?La-lettre-ouverte-de-Public

Le ministère de l'enseignement supérieur, de la Recherche et de l'innovation a mis en place en 2017 une stratégie nationale de soutien à l'édition scientifique dont l'objectif est d'évaluer les effets de la loi pour une République numérique sur l'évolution des revues scientifiques. Sa mission est de concilier à la fois les exigences de libre accès tout en respectant la diversité, la compétitivité et la richesse d'un paysage éditorial complexe[35].

Face à une dynamique de science ouverte voulue par les politiques au niveau international, il convient donc de se questionner sur l'avenir des éditeurs. (Chartron, Schöpfel, 2017)[24]
Le modèle éditorial propose une vraie diversité, quel part faut-il lui laisser face à la mise en accès libre ?
Les institutions publiques et les instituts de recherche tels que l'IRD peuvent-ils prendre en charge les coûts de diffusion des travaux scientifiques ?

Certains professionnels de l'IST plaident pour un modèle qui combinerait la diffusion de l'information scientifique et la publication par des éditeurs qui prendraient en charge des problématiques délaissées par les grands éditeurs d'un côté, et l'auto-archivage en OA de l'autre. (Vanholsbeeck, 2017)[30]
La valeur ajoutée du marché de l'édition se trouve dans sa diversité et sa qualité. Il doit désormais s'inscrire dans la dynamique créée par Internet et répondre aux enjeux d'un accès plus ouvert aux résultats de la recherche, sans toutefois être mis en danger par les exigences de libre accès. (Chartron, 2007)[27]

Face à ses problématiques, des alternatives sont proposées par les acteurs de l'IST dont certains essayent de mettre en place des modèles économiques et projets éditoriaux avec les petits éditeurs.
Le projet NumeRev[36], pour commencer est un projet de portail interdisciplinaire de ressources scientifiques numériques en libre accès dont la mission sera de recenser les projets éditoriaux et leurs contenus et de les rendre interopérables. Cet outil de publication serait gratuit pour les contributeurs qui suivraient la voie de l'accès libre, et payant, sous forme d'abonnement, pour ceux qui préfèreraient la voie éditoriale classique commerciale. (Verlaet, 2017)[31]
Le projet Episciences[37] ensuite, piloté par le CCSD, consiste en une plateforme sur laquelle sont hébergées et gérées des épi-revues, c'est-à-dire des revues électroniques composées d'articles nativement en libre accès. Les articles sont publiés lorsqu'ils ont été validés par des comités de lecture, et disponibles en preprint dans le cas contraire. Ce projet est mené en partenariat avec des instituts de recherche, comme l'INRIA.
Une autre initiative, SCOAP3[38], a permis de convertir les revues les plus importantes du domaine de la physique des hautes énergies en accès libre, sans frais pour les auteurs.

Les problématiques d'édition sont fortement liées à celles du libre accès et influent sur la politique documentaire des établissements. Le choix d'un logiciel de gestion d'une AOI doit prendre en compte ces éléments et ne peut s'orienter vers une solution qui n'intégrerait pas d'applications pour gérer ces aspects.
Les archives ouvertes sont accessibles à partir des plateformes mises en place par les institutions elles-mêmes, mais la production scientifique peut aussi être centralisée sur des sites comme Hal. Nous allons regarder de plus près ces aspects dans la sous-partie suivante, consacrée à la mutualisation.

---

[35] http://www.enseignementsup-recherche.gouv.fr/cid136723/le-soutien-a-l-edition-scientifique.html
[36] http://numerev.com
[37] https://www.ccsd.cnrs.fr/epi-revues/
[38] https://scoap3.org/

2.3 La mutualisation

On observe une tendance à la mutualisation dans le monde de la recherche. Des projets sont en cours. Un des premiers a été le lancement de l'archive Hal créée en 2001 par le CCSD, une unité créée par le CNRS qui était signataire de la déclaration de Berlin sur le libre accès.

D'autres projets de mutualisation sont en cours dans le monde de l'enseignement supérieur et de la recherche, comme le projet de Système de Gestion de Bibliothèque mutualisé (SGBM) notamment, initié par l'agence bibliographique de l'enseignement supérieur (l'ABES) en 2012. Les systèmes de gestion des bibliothèques de l'enseignement supérieur et de la recherche sont actuellement en questionnement, d'après le constat selon lequel les nouvelles possibilités qu'offre le Web n'auraient pas été pleinement mises à profit par les SIGB, faute de moyens. Avec l'arrivée d'internet, si les SIGB proposent aujourd'hui une interface web, ils n'en tirent pas pour autant le meilleur parti, comme rendre interopérables les catalogues entre eux.

Ce projet de grande ampleur a donc pour objectif de créer une chaine collective de signalement, qui pourra être enrichie de façon collective. Concrètement, il consiste à synchroniser le futur SGBM des bibliothèques adhérentes avec les systèmes locaux et celui de l'ABES. Il a pour but d'intégrer nativement la gestion des ressources électroniques dans une chaine collective de signalement.
Cette mutualisation a l'avantage de réduire les coûts tout en assurant une plus grande efficacité technique. Il permet en outre de repenser les circuits de travail, et d'optimiser les process.
Si cette démarche est économique, elle engendre en contrepartie d'autres coûts liés au déploiement de la solution. D'autre part, les moyens déployés par les services informatiques de ces sociétés sont très importants, les personnels des bibliothèques sont donc largement mis à contribution.

Pour autant un système mutualisé a l'avantage de rassembler toute une communauté autour d'un même projet, unissant les résultats de son travail. Les efforts collectifs permettent par la diversité des acteurs de proposer des solutions plus créatives.
C'est toutefois un système contradictoire qui laisse peu de place à la spécificité, or les bibliothèques universitaires sont de plus en plus autonomes dans leur gestion, et disposent d'une communauté de pratiques bien établie.
D'autres inquiétudes peuvent se poser par rapport à la solution technique, car l'éditeur qui se verra attribuer le marché sera renforcé commercialement, pouvant entrainer une situation de monopole lui permettant par la suite d'imposer ses conditions.

La mutualisation peut s'entendre sur d'autres plans. C'est ce que propose le projet Plume[39](Promouvoir les Logiciels Utiles, Maîtrisés et Economiques) lancé en 2006, dans la communauté de l'Enseignement Supérieur et de la Recherche, qui a, parmi ses objectifs, celui de mutualiser les compétences sur les logiciels.
Ce projet, rebaptisé projet Fenix[40] (Fiches d'Évaluation Normalisées Issues de l'eXpérience) en 2016 vise à promouvoir les développements internes, fédérer une communauté autour du logiciel, et promouvoir l'usage des logiciels libres et la contribution à leur élaboration.

[39] https://projet-plume.org/taxonomie/202/fr?page=1
[40] http://fenix.resinfo.org/

L'utilisateur trouvera sur ce serveur des fiches descriptives de logiciels et de ressources rédigées par des personnes qui utilisent régulièrement le logiciel et qui disposent de compétences suffisantes pour une aide à l'installation.
Cette forme de mutualisation permet de réaliser des économies grâce à une mise en commun des expériences et des compétences techniques qui existent dans les universités et les laboratoires.
L'état français lui-même suit cette orientation, à travers sa mission Etalab[41] en encourageant ses agents publics à contribuer à des logiciels libres dans le cadre de leurs missions.

Pour revenir au monde de la recherche, certains établissements ont ouvert un portail Hal en plus de leur propre AOI, comme l'IRD, considérant alors Hal comme une solution complémentaire, tandis que d'autres s'en remettent à Hal pour l'ensemble de leurs communications scientifiques, comme l'institut Pasteur. Leurs cas sont abordés de plus près dans la troisième partie, avec notamment le cas de l'INRA qui a décidé récemment de fermer son AOI Prodinra pour verser l'intégralité de sa production dans Hal, centralisant ainsi toute sa production en un seul emplacement.
La mutualisation induit en effet souvent de centraliser les ressources, faut-il donc préférer un système centralisé plutôt qu'un système distribué ?
Les raisons sont souvent politiques d'une part, ces décisions étant prises par la gouvernance des établissements, et financières d'autre part, car les choix de centraliser ou mutualiser permet d'économiser les coûts considérables engendrés par la gestion d'une AOI.
Les avantages techniques sont parfois mis en avant, comme offrant une plus grande efficacité documentaire, quoique nous reviendrons sur ce point dans la troisième partie, dans laquelle nous présenterons les fonctionnalités techniques de Hal.
La centralisation suscite toutefois quelques inquiétudes, on peut notamment se questionner sur la pérennité d'une archive centrale, si celle-ci venait à disparaitre, pour des raisons techniques, économiques ou politiques.

Face à ces interrogations, des sites refuges voient le jour, comme le site Datarefuge[42].
Des études questionnent aussi l'intégrité des contenus dans une archive centrale, et notamment de ce qu'il advient des versions antérieures, et préconisent que seul un système décentralisé peut garantir la conservation des différentes copies. (Girard, 2017)[33]
Des initiatives existent pour contrer ces risques, comme celle prise par le site LOCKSS, qui consiste à mettre en place un réseau distribué de réservoirs indépendants. Ces réservoirs moissonnent les dépôts des adhérents au programme et récoltent ainsi les nouveautés ou les fichiers modifiés, assurant une pérennité aux utilisateurs sur leurs collections.

Centraliser la production scientifique permet néanmoins d'assurer une meilleure visibilité, et d'assurer plus facilement les calculs métriques.
Un des principaux objectifs d'une archive centrale est d'assurer l'exhaustivité, à condition que le dépôt soit obligatoire.
Ce sujet fait débat à l'heure actuelle, car il est question de rendre obligatoire le dépôt des articles qui sont issus d'une recherche financée par les fonds publics[43] alors que jusqu'à présent le code de la recherche protégeait l'autonomie du chercheur[44].
Faut-il donc obliger ou simplement encourager le dépôt sur Hal ?

[41] https://www.etalab.gouv.fr/publication-de-la-politique-de-contribution-de-letat-aux-logiciels-libres
[42] https://www.datarefuge.org/
[43] Cette mesure est énoncée dans le premier axe du plan national pour la science ouverte.
[44] Article L411-3 du Code de la recherche

Certaines institutions encouragent le dépôt en associant le dépôt aux chances de promotion des chercheurs. Ainsi, une candidature pour postuler sur un nouveau poste pourrait être considérée en fonction des dépôts que le chercheur aura effectués sur Hal.
D'autres instituts rendent obligatoire ce dépôt en l'inscrivant dans les contrats de travail, comme c'est le cas pour l'INRIA. Le dépôt ne doit pas forcément se faire dans HAL, mais la publication doit être accessible.
Le dépôt obligatoire a aussi l'avantage de faciliter le travail du personnel des services IST.

Ces questions ont été soulevées lors de la journée des AOI organisée par Sciences Po en juin dernier. L'un des intervenants avait notamment fait remarquer que la prolifération des réservoirs et la concurrence qui en découlait pouvait toutefois conduire à décourager les chercheurs à déposer dans Hal, comme sur l'AOI de leur établissement.
Les chercheurs peuvent ressentir aussi l'obligation de dépôt comme une pression supplémentaire liée aux questions d'évaluation et de bibliométrie, ce sujet est abordé dans la sous-partie suivante. (Valluy, 2017) [35]

2.4 Evaluation, subventions et bibliométrie

La visibilité des publications des chercheurs est un enjeu primordial, car c'est à partir de ce travail que se fait leur évaluation.
Les chercheurs sont évalués tous les deux ans, et cette évaluation se fait grâce à des indicateurs qui permettent de repérer leurs articles publiés. Ces chiffres sont ensuite remontés au Haut conseil de l'évaluation de la recherche et de l'enseignement supérieur (HCERES).
Les impératifs d'évaluation sont très importants, et les établissements exposant la production de leurs chercheurs doivent disposer d'outils performants permettant de mesurer cette production
Nous avons abordé ce sujet brièvement dans la première partie lors de la description du processus d'alimentation de la base Horizon. L'attribution d'une publication à un chercheur doit être précise car de ses publications découle son évaluation.
Les indicateurs sont l'objet de la bibliométrie qui mesure de manière quantitative la production scientifique d'un chercheur, d'un groupe de chercheurs, d'un institut ou d'un pays. Un des principaux usages de la bibliométrie aujourd'hui est l'évaluation de la recherche.

Les revues scientifiques pèsent beaucoup dans le système d'évaluation. C'est Eugene Garfield qui, en 1950, a l'idée d'utiliser les citations présentes dans les articles scientifiques, c'est-à-dire les renvois faits à d'autres articles, pour lier les articles entre eux. Il créé le SCI (Science Citation Index) qui permet donc d'estimer le facteur d'impact d'une revue, c'est-à-dire sa visibilité par rapport à la fréquence de citations de ses articles.
A la même époque, un physicien anglais, spécialiste de l'histoire des sciences et de la science de l'information Derek John de Solla Price, décide d'utiliser les articles scientifiques comme indicateurs quantitatifs de l'activité de recherche. Il créé ainsi la scientométrie, science de la mesure et l'analyse de la science. Elle est souvent en partie liée avec la bibliométrie
Cette méthode se base sur des données quantitatives, en ignorant le contenu des articles. Elle va donc se servir de l'ensemble des publications scientifiques et de leurs propriétés statistiques comme d'un indicateur de l'activité scientifique.
Ces deux procédés se basent sur les articles publiés dans des revues, et aujourd'hui le référentiel des articles recensés dans le WOS a pris le relais du SCI. Sa fonction première était de servir comme outil de repérage par les bibliothécaires, mais représente désormais un important indicateur.

L'IRD, tout comme les autres instituts de recherche, évalue ses chercheurs et mesure la diffusion de sa production scientifique par le moyen d'indicateurs. La DDUNI a développé à cet effet une application nommée MAP (moyens d'aide au pilotage) pour réaliser ce travail,

indispensable pour tenir à jour les différents classements dans lesquels l'IRD se positionne, pour mesurer la production scientifique globale, et notamment les indicateurs permettant de mesurer sa diffusion, dont les publications exposées en libre accès.
Le libre accès contribue à rendre visible la production scientifique, et permet d'augmenter le taux de citation, et constitue par là-même un instrument de mesure très intéressant de bibliométrie et d'évaluation des chercheurs. Les politiques d'encouragement au dépôt dans les archives ouvertes visent aussi à augmenter la visibilité d'une publication à des fins de bibliométrie.
Le marché des logiciels de gestion d'AOI proposent quasiment tous aujourd'hui des modules de bibliométrie, ou dashboard, qui permettent d'extraire des statistiques à partir de corpus de documents.

Toutefois cette pression induite par les indicateurs est remise en cause par certains acteurs de l'IST, et certains instituts de recherche.
On observe en effet des craintes sur l'avenir du libre accès et les modes de publications. L'objectif premier de diffusion de la science risque de devenir secondaire face à la pression des indicateurs que connaissent les chercheurs pour leur évaluation. (Chartron, 2010)[26]
Même le dépôt sur une archive ouverte institutionnelle, est parfois ressentie par le chercheur comme un moyen de mesurer sa production, le dépôt devient alors un outil managérial de mesure.(Valluy 2017)[35]
Il faut noter toutefois que l'évaluation des chercheurs tend à évoluer. Les orientations aujourd'hui vont dans le sens d'une évaluation qui ne dépendrait plus seulement d'indicateurs chiffrés. Ainsi, en 2012, un groupe d'éditeurs et des responsables de maisons d'édition de journaux scientifiques s'est réuni afin de réfléchir à la question.
Cette déclaration[45] recommande notamment de ne pas seulement baser l'évaluation de la recherche sur l'impact factor d'un journal, mais aussi d'attacher de l'importance au contenu de l'article et à sa qualité.

Les fonctionnalités liées à la bibliométrie sont donc indispensables pour un logiciel gérant des archives ouvertes. L'IST connait une ère de changement, et la bibliométrie, jusqu'alors basée sur des critères quantitatifs, semble s'orienter vers des critères qualitatifs. Certains instituts de recherche comme l'INRA projettent d'abandonner les données chiffrées pour accorder plus d'importance à d'autres éléments, comme le nombre de doctorants qu'aura encadré un laboratoire de recherche. (information recueillie lors de nos entretiens).
La Commission européenne recommande également, depuis 2012, d'élaborer de «*nouveaux modèles, critères et indicateurs alternatifs pour l'évaluation des carrières* » et encourage l'examen « *de nouveaux indicateurs de la recherche et de nouveaux indicateurs bibliométriques englobant non seulement les publications scientifiques, mais aussi les ensembles de données et d'autres types de résultats issus des activités de recherche et les réalisations de chaque chercheur*.»[46]
En 2016, la Commission a également lancé une étude menée par un groupe d'experts, dans le cadre de l'Open Science, sur les métriques alternatives. (Vanholsbeeck, 2017)[30]
Les professionnels de l'IST et les chercheurs espèrent que le soutien de la Commission pour les répertoires institutionnels permettra la visibilité de toute la diversité qu'offre la communication scientifique dans les analyses bibliométriques qui prennent trop souvent en compte les seuls articles scientifiques publiés dans des revues de renommée.
Enfin, depuis 2012 une association britannique s'est constituée, Open Scholar[47], qui a pour objectif de développer un plugin, l'Open Peer Review Module for Repositories (OPRM) permettant aux dépôts d'archives en libre accès d'intégrer les fonctionnalités nécessaires

[45] https://sfdora.org/read/
[46] Article 7 de la Recommandation de la Commission du 17 juillet 2012 relative à l'accès aux informations scientifiques et à leur conservation
[47] http://www.openscholar.org.uk/open-peer-review-module-for-repositories/

pour devenir également des plateformes d'évaluation. Le but poursuivi par cette association est de rapprocher de la communauté des chercheurs le contrôle de la qualité de la recherche. Ce plugin sera compatible avec le logiciel Dspace, mais a été conçu pour s'adapter sur des d'autres logiciels de gestion de référentiels, tels Eprints.
Ce projet est à suivre, et selon la politique bibliométrique suivie par l'IRD, il pourrait être intéressant de porter son choix sur un logiciel qui aura la capacité d'intégrer de nouveaux modèles de calcul de la production scientifique.
Le traitement de ces données et leur mise en relation, tous objectifs confondus, bibliométriques ou autres, est désormais possible grâce aux avancées technologiques de l'informatique, mais aussi du Web. Nous allons aborder dans cette dernière sous-partie les applications possibles du Web sémantique dans notre contexte.

2.5 Les nouvelles interopérabilités et ses applications IST

L'expression « Web sémantique » a été utilisée pour la première fois par Tim Berners Lee en 2001 pour désigner «*une évolution du web qui permettrait aux données disponibles (contenus, liens) d'être plus facilement utilisables et interprétables automatiquement, par des agents logiciels*»[48].
Ce concept repose sur des standards et des technologies qui vont permettre de faire communiquer entre elles d'immenses bases de données.
Le Web sémantique est composé de langages infiniment plus riches que ce que permet le langage HTML, grâce auxquels la signification et la structure des contenus sont représentés de manière à être accessibles à des logiciels et machines.

Dans notre contexte, le Web sémantique est intéressant dans le sens où il offre des possibilités d'interopérabilité entre les différents réservoirs.
Dans le cadre de notre mission, nous avons interrogé la responsable du Pôle informatique de DataPersée, qui a participé au projet récent de mise en place d'un triplestore. Cette initiative fait suite aux comportements nouveaux des chercheurs en matière de recherche documentaire, qui consistent de plus en plus à télécharger des corpus pour en croiser les métadonnées ou les données de recherche.
Le Web sémantique peut bien sûr être appliqué aux archives ouvertes. Hal a d'ailleurs ouvert un triplestore[49].

Ces évolutions amènent à se questionner sur l'opportunité de les appliquer à une archive ouverte, et dans quelle mesure.
L'interopérabilité apportée par le Web sémantique est d'abord rendue possible par un système d'identification des ressources à travers un URI (Uniform Resource Identifier). Le format RDF (Ressource Description Framework) va permettre de structurer les métadonnées pour rendre leur traitement automatique. Ces deux composants sont intégrés par certains logiciels d'archives ouvertes aujourd'hui.
Selon l'orientation et les possibilités d'évolution qu'un établissement souhaite donner à son archive ouverte, choisir un logiciel supportant le format RDF et un système d'identification des ressources peut s'avérer intéressant pour l'avenir, afin de pouvoir bénéficier de toutes les opportunités que pourra apporter les avancées dues au Web sémantique.

Le Web sémantique rend aussi possible l'interconnexion des données, dès que celles-ci sont disponibles en libre accès.
Or, parmi les projets portés par le monde international de l'IST dans le cadre d'une science ouverte figure le plan de la mise en accès libre les données de la recherche. Le plan

[48] Définition consultable sur le site de la BnF
[49] https://data.archives-ouvertes.fr/

national pour la science ouverte encourage « *la diffusion sans entrave des publications et des données de la recherche* ».
Le rapport OCDE de 2007 donne la définition suivante des données de la recherche: les données de la recherche sont définies comme « *des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche*»[50].
Les avancées dans le monde du numérique rendent désormais possible la réutilisation des données originales collectées par les chercheurs pour leurs études. Chaque publication a à son origine des tableaux et autres éléments sur lesquels s'appuie les résultats de son étude. Ces données sont généralement stockées sur des serveurs internes, et, par manque d'exposition, ne sont pas réutilisées.
Certains éditeurs demandent déjà aux auteurs de publier les jeux de données en lien avec leur publication.
L'OCDE dans son rapport encourage le partage des résultats des recherches issues de fonds publics, y compris sur les données donnant lieu aux publications, et attache à ce projet les principes d'ouverture, de flexibilité, et de transparence.

Il existe donc un lien certain entre l'ouverture des données de la science et le principe des archives ouvertes.
Le principe de flexibilité en particulier implique la prise en compte, entre autres, des évolutions rapides du monde des TIC.
Ce principe renvoie dans notre cas aux fonctionnalités que doit présenter le logiciel qui gèrera la base Horizon. A l'heure de l'ouverture des données de la science, il serait souhaitable de pouvoir implémenter une API qui s'assurera du dépôt automatique des données liées à la publication dans une base de données. Certains logiciels comme PolarisOS permettent de déposer avec la publication, des documents d'étude liés à celle-ci.
L'IRD a lancé cette année un projet d'entrepôt de données qui doit être réalisé grâce au logiciel Dataverse. Ce logiciel, développé par l'université de Harvard et adopté par de nombreuses institutions en France et dans le monde, permet à l'IRD d'intégrer un écosystème d'entrepôts interopérables. Ce projet a pour objectif la diffusion et le partage des jeux de données.

Selon le principe d'ouverture recommandé par le rapport de l'OCDE, les coûts d'accès à ces données doivent être le plus bas possible. Le rapport préconise aussi que les organisations productrices de ces données soient transparentes sur l'information liée à ces données, ainsi que leurs conditions de réutilisation.
Ces principes, tout comme l'accès libre aux publications, s'inscrivent dans une science ouverte qui œuvre pour la circulation et la diffusion les plus larges des publications scientifiques et des jeux de données qui y sont rattachés. Tout comme pour les publications, il est possible de rattacher des droits à ces jeux de données pour en faciliter la réutilisation.
Le statut juridique des données de la recherche est encadré par l'article 30 de la loi pour une République numérique qui prévoit leur libre réutilisation lorsqu'elles sont issues d'une activité de recherche financée au moins à moitié par un financement public, si elles ne sont pas protégées par un droit spécifique, et si elles ont été rendues publiques par chercheur ou l'établissement.
Enfin, c'est l'établissement de rattachement du chercheur qui décide quelles données seront ouvertes, où elles doivent être déposées et sous quelles conditions. C'est l'institution qui

[50] http://www.oecd.org/fr/science/inno/38500823.pdf

dispose de droits sur les données, et non le chercheur, contrairement aux publications qui restent sa propriété intellectuelle.[51]

L'Ifremer a ainsi optimisé la réutilisation de ses données en connectant des bases les unes aux autres. Son archive Archimer est connectée avec SEANOE, un sous-ensemble d'Archimer, et accueille non seulement les jeux de données produites par les chercheurs d'Ifremer, mais également les données bancarisées à la demande d'autres instituts[52]. Une archive ouverte peut ainsi englober des jeux de données, ou être mise en lien avec des banques de données.
Ces données de la science marine peuvent être des données récoltées par des flotteurs disséminés dans les océans, mais aussi des images, des vidéos sous-marines. Elles sont interopérables avec d'autres systèmes documentaires, comme le catalogue des campagnes à la mer de la flotte océanographique française.
Ces données sont visibles depuis Archimer, elles sont interopérables et chaque jeu est identifié par un DOI qui permet de rendre visible l'auteur et de le citer correctement.
Outre le stockage de données, ces jeux peuvent ensuite être fouillés pour être mis en lien entre eux. Il est possible aussi d'extraire des informations parmi un très large corpus, grâce à des outils de TDM. Cette opération serait impossible humainement, mais les avancées technologiques ont permis d'optimiser la recherche sur des bases de données volumineuses.

Une archive ouverte offre ainsi à une institution l'opportunité de centraliser la production de ses chercheurs et d'en exploiter les données de façon optimale.
Au niveau français, le CNRS a créé l'INIST, l'institut de l'information scientifique et technique, une plateforme d'information et de services numériques pour la recherche. Les missions de cet institut tournent autour de trois axes : l'analyse et la fouille de l'information, la valorisation des données de la recherche, et l'accès à l'information scientifique.
L'INIST a contribué au développement du projet ISTEX, une plateforme pluridisciplinaire proposant un accès en ligne aux collections rétrospectives de la littérature scientifique. Cette plateforme offre aussi la possibilité aux chercheurs de télécharger de très larges corpus de documents, accompagnés de services permettant de traiter efficacement les données par le biais de la fouille de textes.
Ces projets permettent de connecter une archive ouverte à d'autres répertoires, et rendre ses publications plus visibles, tout en assurant une meilleure diffusion des connaissances, grâce à ce croisement entre données.

Grâce à l'ouverture des données et des moyens numériques permettant de les lier les unes autres, beaucoup d'idées deviennent possibles.
L'INRIA a, dans ce sens, en 2016, commencé à recueillir les codes sources de tous les logiciels libres dans le monde entier à travers son projet Software Heritage. Ces codes deviennent désormais des données déposables sur une archive et peuvent être mis en relation avec d'autres. Hal recueille ainsi depuis septembre 2018 les codes source des logiciels libres, en partenariat avec l'INRIA. Ces codes peuvent ensuite être liés aux publications dont ils sont le produit.
Des concepts jusqu'alors abstraits et irréalisables deviennent possibles, grâce aux avancées technologiques et aux institutions qui prônent une science ouverte.

[51] http://www.bibliothequescientifiquenumerique.fr/wp-content/uploads/2018/03/Guide_analyse_Cadre_Juridique_Ouverture_donnees_Recherche_V2_licenceOuverte_prefaceDGRI.pdf
[52] Voir entretien n°7

Cette partie montre que la gestion d'une archive ouverte est très liée aux problématiques actuelles de l'IST. Ses acteurs mènent actuellement un travail conséquent pour se tenir à jour de ces évolutions, et nouveaux modèles, tout en composant avec les modèles économiques déjà existants dont il faut tenir compte. Leur mission est de proposer des solutions alternatives capables de concilier le monde d'hier et celui d'aujourd'hui.
Il en résulte un monde de la connaissance scientifique de plus en plus interconnecté, dont les bases et réservoirs peuvent être liés entre eux, diffusant cette connaissance partout dans le monde, et jusqu'aux pays dont l'accès à internet et aux savoirs scientifiques en particulier reste encore trop compliqué par manque de moyens suffisants.

Les éditeurs de logiciels de gestion d'archives ouvertes doivent eux aussi prendre en compte les nouvelles exigences liées à ces questionnements pour proposer des outils capables de les gérer, ou capables d'évoluer.
En effet, les avancées sont si rapides et nombreuses que même un outil capable de tout gérer aujourd'hui doit pouvoir être adaptable aux progrès et nouveautés de demain. Les éditeurs doivent proposer des logiciels qui, s'ils ne proposent pas déjà ces fonctionnalités, pourront les implémenter dans le futur.
Ce sujet est abordé dans la partie suivante, qui va tenter de présenter l'ensemble des fonctionnalités d'un logiciel de gestion d'une AOI qui permettront de répondre aux problématiques et enjeux que nous venons d'exposer.

Partie 3 - Etude de cas : l'archive ouverte institutionnelle de l'IRD, évolutions possibles dans le contexte de l'IST actuel

Dans le cadre de son projet de refonte du logiciel hébergeant son archive ouverte, nous avons réalisé pour L'IRD une étude de marché sur les solutions proposées et adoptées aujourd'hui par les instituts de recherches et universités.

Nous commencerons par exposer la méthodologie que nous avons suivie.
Puis nous présenterons les résultats de cette étude sous la forme d'un tableau récapitulatif de tous les entretiens que nos interlocuteurs nous ont accordés.
A partir de ce tableau et des retours des utilisateurs interrogés, nous avons établi cinq scénarios possibles pour l'IRD.
Enfin, nous établirons en fin de partie une liste des préconisations que devrait suivre l'IRD lorsque l'institut fera son choix final.

Pour commencer ce benchmark, nous nous sommes penchées sur les études déjà existantes, ainsi que sur les SIGB proposés sur le marché.
L'offre des logiciels destinés à la gestion d'une bibliothèque s'est en effet multipliée dans les dernières années, et est beaucoup plus complète aujourd'hui. Grâce aux API et aux webservices, les produits proposés sont interconnectables et répondent mieux aux besoins des utilisateurs.
L'informatisation des catalogues de bibliothèques n'est toutefois pas nouvelle et remonte aux années 70, bien avant l'avènement du web. Ces catalogues fonctionnaient déjà de pair avec des bases de données relationnelles, les SGBD (système de gestion de bases de données) qui ont connu de grandes avancées depuis les années 60.
Ce n'est qu'à la fin des années 80 qu'apparaissent les premiers SIGB: l'idée consistait à saisir une seule fois les notices pour plusieurs usages.

On exige des SIGB aujourd'hui qu'ils répondent aux nouveaux besoins des utilisateurs : réalisation de recherches multicritères, filtration des résultats fournis par le moteur de recherche et en téléchargement des contenus, mise en place d'un flux RSS des dernières publications, offre personnalisée en fonction de l'historique de recherche, partage des paniers avec d'autres utilisateurs, possibilité de s'interconnecter avec des outils de découverte, des catalogues en ligne, des moteurs de recherche extrêmement puissants, et d'autres encore.
Même si le fonctionnement des bibliothèques classiques diffère de celui des bibliothèques des instituts de recherche, et en particulier de celui de l'IRD qui connait la particularité de gérer à la fois un catalogue et une AOI, les nouvelles bibliothèques numériques permettent de plus en plus de gérer aussi bien l'un que l'autre, et les fonctionnalités principales demandées, que nous développerons dans cette partie, restent les mêmes.
L'offre technologique est donc très intéressante et permet d'envisager catalogues et AOI sous un nouvel angle.

On attend également des SIGB qu'ils répondent aux exigences particulières du monde de la recherche, comme de pouvoir exporter des notices bibliographiques en masse et dans différents formats, gérer les affiliations aux différents laboratoires, et donc intégrer des référentiels, mais aussi aux exigences de libre accès, comme gérer des embargos imposés par les éditeurs, assurer un accès vers le texte intégral, fournir des métadonnées riches, et d'autres fonctionnalités que nous listerons dans la partie 3.5: conclusion et préconisations.
En outre, la recherche d'un nouveau SIGB impose de se poser des questions essentielles avant de se mettre en recherche :
Le client souhaite-t-il une utilisation en mode SAAS ? faut-il acheter une licence, ou s'abonner ? Héberger la solution en interne ou chez l'éditeur ? Ou tout verser sur une plateforme centralisatrice qui prendra en charge la maintenance ? Faut-il préférer un logiciel propriétaire à un logiciel open source ?
La communauté autour des logiciels libres s'est largement développée au cours des vingt dernières années, et la communauté scientifique y participe aussi désormais, souvent encouragée par les gouvernements. Le modèle économique en résultant s'est montré viable.

D’autre part, les modèles open source ont prouvé leur aspect sécuritaire, par une communauté active qui peut réagir en temps réel en cas de problème technique.
Enfin, les solutions open source demeurent plus accessibles aux chercheurs des pays émergents, lorsque le code reste raisonnablement crypté. Or, ceci est une contrainte que doit observer l’IRD.
On remarque en outre, en parcourant la littérature académique anglosaxonne, que les concepts de libre accès et d’open source sont étroitement liés. Le mouvement de l’open source est né au tournant des années 2000, au moment où émergeait le mouvement de l’open access. (Awre, Green, 2017)[45]
Un produit open source, comme un document disponible en libre accès, peut être réutilisé, modifié, diffusé, sans compensation financière, par et pour tout le monde, pour différents objectifs.
L’open source est surtout encouragé pour procurer une vraie pérennité aux ressources exposées en libre accès. Les formats de logiciels propriétaires sont certainement plus compliqués à migrer. (Corrado, 2005)[46]
Nous allons tenter de montrer les avantages et inconvénients de chaque solution.

3.1 Méthodologie suivie pour notre benchmark

Nous avons sondé six instituts de recherche et une bibliothèque universitaire.
Le travail a commencé par une immersion dans le SID de l'IRD gérant l'AOI Horizon, afin d'en comprendre son fonctionnement, ses forces et ses lacunes.
Nous avons ensuite déterminé une liste d'instituts, en fonction des outils utilisés par ces derniers. Nous souhaitions interviewer un panel constitué de systèmes utilisant des logiciels open source, des logiciels propriétaires, et des solutions de mutualisation, même si tous n'utilisaient pas ces outils pour mettre en avant des archives ouvertes.
Ainsi, l'INSERM a mis en place la solution Dspace pour valoriser ses collections éditoriales, mais nous souhaitions avant tout un retour sur ce logiciel.

Nous voulions idéalement comparer Horizon à d'autres AOI, mais d'une part, tous les instituts de recherche ne disposent pas d'une AOI, et d'autre part, la particularité d'Horizon est, comme nous l'avons vu dans la première partie de ce mémoire, de gérer un catalogue classique de bibliothèque et des articles et autres publications parues dans les revues scientifiques. Or, peu d'instituts de recherche gèrent cette particularité.
On trouve beaucoup d'AOI dans les universités, mais leur fonctionnement interne se rattache à l'enseignement supérieur et diffère de celui de la recherche.
Nous avons donc fait le choix de nous entretenir avec des instituts de recherche dont les problématiques sont plus proches de celles de l'IRD.

D'autre part, l'IRD vient de mettre à disposition de son personnel le système de GED : Alfresco, pour la gestion de leurs documents internes. Nous avons dans un premier temps envisager d'inclure cette GED dans notre benchmark, puisqu'il est possible d'exposer sa production documentaire via ce système. Toutefois, Alfresco, comme d'autres GED, ne permet pas de gestion d'OPAC, et ne permet pas de gérer des métadonnées assez riches. Ses fonctionnalités étant trop limitées, nous avons d'office écarté cette solution.

Avant de commencer les entretiens, nous avons établi plusieurs trames de questionnaires, selon que nous interrogions des utilisateurs ou des éditeurs de logiciels.
Ces trames ont été revues au fur et à mesure que les entretiens étaient menés. Ces entretiens se sont tenus par téléphone pour tous, à part pour l'INED, où nous nous sommes rendues sur place pour une démonstration en préproduction de leur nouvel outil, Polaris OS.
Les entretiens sont disponibles dans leur intégralité en annexe 3. Nous présenterons dans cette partie un tableau comparatif récapitulatif que nous avons rempli en fonction des réponses obtenues, c'est pourquoi toutes les cases n'ont pu être renseignées.
Les entretiens ont duré environ une heure chacun, et les personnes interviewées étaient soit les administrateurs du système, soit des informaticiens ayant mis en place le système documentaire.

Nous avons ensuite réalisé des recherches documentaires sur Google et Google Scholar, sur les thèmes suivants :

- Les SIGB et les archives ouvertes
- La mutualisation des archives ouvertes
- Les benchmarks entre les différents logiciels de gestion des archives ouvertes
- Les différents projets IST menés par les institutions
- La bibliométrie et l'évaluation des chercheurs
- Les problématiques d'édition scientifique et la gestion des droits de diffusion des documents publiés,
- Les différents modèles économiques sur lesquels reposent les sujets abordés : les logiciels open source, l'édition scientifique
- L'urbanisation et l'architecture d'un système documentaire
- Le web sémantique et les questions d'interopérabilité

Nous avons cherché dans les ressources suivantes :

- Sites et blogs de réflexion, de professionnels de la documentation dans le domaine de l'IST,
- Livres blancs,
- Thèses et mémoires,
- Sites de différentes organisation liées au développement du libre accès.

Ces recherches ont été menées avec des termes français et anglais.

3.2 Présentation des instituts sondés et des résultats de l'enquête

L'intégralité des entretiens se trouvant en annexe 1, nous ne présenterons que succinctement les systèmes documentaires des entités interrogées.

- **Les instituts ayant choisi la solution HAL**

  - **L'institut Pasteur**- https://hal-pasteur.archives-ouvertes.fr/

L'institut expose sa production scientifique sur un portail Hal.
Sa direction ayant signé la Déclaration de Berlin en 2004, il a été décidé de créer une plateforme d'archives ouverte pour répondre à la demande de la politique d'Open Access,
Pasteur verse donc ses publications sur son portail HAL-Pasteur depuis 2007.La politique de dépôt de l'institut Pasteur est incitative, mais pas coercitive.
Le portail Hal-pasteur compte aujourd'hui plus de 3000 notices avec texte intégral, et plus de 2 700 références bibliographiques.

  - **L'INRA**- https://Prodinra.inra.fr

L'institut national de la recherche agronomique a pris la décision de fermer son archive ouverte institutionnelle Prodinra) et de verser toute sa publication scientifique dans HAL.
Prodinra gère actuellement près de 266 000 références dont presque 18% avec texte intégral.
Cette archive était jusqu'alors gérée par un ECM, dont les nombreuses mises à jour, et la maintenance étaient relativement coûteuses.
L'outil est devenu obsolète, la structure a dû être repensée, et c'est dans ce contexte que la décision a été prise de migrer Prodinra vers Hal.
L'INRA dispose déjà d'un portail HAL. il n'y a pas d'obligation de double dépôt mais un export vers HAL existe pour les UMR et partenaires.

  - **L'IRD**- hal.ird.fr/

L'institut gère également un portail HAL en plus de son archive ouverte Horizon. Nous avons intégré les résultats de cet entretien dans notre enquête pour compléter l'expérience de Hal.
La politique de dépôt n'est pas obligatoire, et la liste des ressources que l'on trouve sur le portail Hal-IRD n'est pas exhaustive, contrairement à ce qui est exposé sur Horizon.

- **Solutions issues du libre**

  - **L'INED -** archined.ined.fr

L'institut national d'études démographiques n'avait pas encore son AOI, et a dû penser entièrement la façon dont il souhaitait exposer ses publications.

Ce choix s'est porté sur une solution issue du libre : Polaris OS. Cette solution vient de faire son apparition sur le marché. Nous avons également interrogé l'éditeur pour avoir une vue d'ensemble des fonctionnalités proposées, que nous avons par conséquent intégrées dans notre tableau.
Les publications de l'INED sont aussi versées sur Hal mais le portail HAL_INED n'est pas exhaustif et Archined est considéré par l'institut comme la vitrine de ses publications.

- **L'INSERM -** http://www.ipubli.inserm.fr/

L'Institut national de la santé et de la recherche médicale utilise Dspace pour exposer ses collections éditoriales. Le nom de cette plateforme est IPubli. Il ne s'agit pas des publications individuelles produites par les chercheurs, celles-ci étant disponibles sur Hal et sur PubMed. L'équipe HAL-Inserm aide les chercheurs à effectuer leurs dépôts, les vérifier et éventuellement les compléter.
Cet entretien nous a permis d'avoir un exemple d'utilisation de DSpace, un logiciel libre largement choisi par les universités pour leur AOI.

- **Solutions à plusieurs composantes**

  - **L'université de São Paulo -** http://bdpi.usp.br/

L'université utilise DSpace. Cette solution a été complétée par un outil de visualisation Kibbana (http://bdpi.usp.br/dashboard.php) et le moteur de recherche ElasticSearch.

  - **Ifremer -** https://archimer.ifremer.fr/

La superposition de plusieurs applications est une solution qui a aussi été choisie par Ifremer, pour son archive Archimer, avec une différence toutefois : l'institut a choisi Oracle, un logiciel propriétaire, puis l'a enrichi de développements maison sur une base MySQL. Le SID d'Archimer a été documenté et expliqué dans deux document consultable en bibliographie (Merceur, 2005)[50].
Une obligation de dépôt a été mise en place pour l'ensemble du personnel Ifremer depuis septembre 2010.

- **Les éditeurs :**

Nous avons aussi interrogé deux éditeurs de logiciels: **PMB**(https://www.sigb.net/) et MyScienceWork pour le lancement de sa récente solution **PolarisOS**(https://www.mysciencework.com/polaris-os).

Nous avons en outre recensé certaines des fonctionnalités offertes par: La solution **Dspace**, sur la base de sa documentation technique disponible en libre accès (https://wiki.duraspace.org/pages/viewpage.action?pageId=78163330)

Nous avons signalé dans notre tableau la solution libre **Folio**(https://www.folio.org/),éditée par **Ebsco**, qui sortira courant 2019 mais pour laquelle nous n'avons pas pu obtenir d'informations.

- Enfin, nous avons contacté **DataPersée**(http://data.persee.fr/)qui a transposé toutes ces données dans un triplestore pour répondre aux besoins de ses utilisateurs.

Cette solution ne figure pas dans notre tableau car les fonctionnalités proposées ne correspondent pas à celles que nous avons recensées. Nous avons tenu à présenter l'entretien en annexe 3, car l'approche de DataPersée est innovante et intéressante.

Pour aller plus loin, nous citons, dans la bibliographie, quelques benchmarks réalisés par différents cabinets.
Le cabinet Tosca Consultants[53] réalise notamment tous les ans une étude comparative des solutions logicielles choisies par les grandes bibliothèques françaises, rendue sous forme de plusieurs tableaux récapitulatifs. (Maisonneuve, 2018)[52]
Les différents benchmarks comparent surtout des logiciels open source, et Dspace est présent dans tous les comparatifs.

Ci-dessous se trouve notre tableau comparatif.

[53] Les résultats de cette étude sont consultables sur https://toscaconsultants.fr/.

Nom de l'entité interviewée		Editeur	Inserm sur iPubli	DSpace	L'Université de São Paulo sur BDPI (production scientifique des chercheurs de l'Université de Saõ Paulo)	Ined sur ArchIned	MyScienceWork	Ifremer sur Archimer	IRD sur Horizon	CCSD	INRA sur ProdInra	Institut Pasteur sur HAL - Pasteur	EBSCO
Nom de la solution logicielle		PMB	Dspace	DSpace	Dspace + Kibbana + Elastic Search	Polaris OS	Polaris OS	Oracle + PHP	Cadic + PHP	HAL	ECM	Hal	Folio
Type de la solution logicielle		SIGB open source	open source software f	open source	Mixte	SIGB open source	OS next generation rep	Mixte	Mixte	Mutualisée		Mutualisée	open source
Utilisation		Présentation par l'éditeur	Plateforme de publication numérique des collections éditoriales de l'Inserm	Présentation par l'éditeur	Gestion de l'archive ouverte	Gestion de l'archive ouverte	Présentation par l'éditeur	Gestion de l'archive ouverte	Gestion de l'archive ouverte	Gestion de l'archive ouverte	Gestion de l'archive ouverte	Portail Hal	A l'heure actuelle, une version beta de l'outil est en préparation et les premières implémentations de l'outil sont prévues pour Janvier 2019.
URL		www.sigb.net	http://www.ipubli.inserm.fr/	https://duraspace.org/dspace/	http://bdpi.usp.br/	Site non encore ouvert	https://www.myscienceork.com/polaris-os	http://archimer.ifremer.fr/	http://horizon.documentation.ird.fr	https://hal.archives-ouvertes.fr/	https://prodinra.inra.fr/?locale=fr#	https://hal-pasteur.archives-ouvertes.fr/	https://www.ebsco.com/e/fr-fr/prodotti-e-servizi/folio
OUTIL	**CATEGORIE**												
Hébergement du SI chez le client ou chez l'éditeur	Facilité de prise en main de l'outil	au choix	client	au choix		client	au choix	client	client				
Administration de l'interface par le documentaliste	Facilité de prise en main de l'outil	Non	au-desssus: changer pour hébergement chez un intermédiaire (INIST)	Non		Oui	Oui	Non	Non	Non	Oui	Oui	
Ajout de modules ad hoc par un informaticien en interne	Facilité de prise en main de l'outil	Oui	Non	Oui, mais difficile		Possible	Possible	Oui	Non	Non		Non	
Ajout de modules d'extension par un informaticien en interne (API propriétaire mais implémenté en interne)	Facilité de prise en main de l'outil	Oui	Non	Oui, mais difficile		Possible	Possible	Oui	Oui, briques ma	Non		Non	
Fréquence et facilité des mises à jour	Facilité de prise en main de l'outil	NSP	peu fréquent, mais diff	Oui, mais difficile		Outil trop jeune pour répondre	Outil trop jeune pour répondre	Oui par l'informatique	Montées de ver	Non	Fréquentes mises à jour	Peu de mises à jour	
Exposition des données dans des formats et protocoles standards (OAI-PMH, RDF, ...)	Interopérabilité	Oui	Oui	Oui		Oui	Oui	Oui	Oui	Oui		Oui	
Possibilité de créer un triplestore?	Interopérabilité	Oui	NSP	Oui		Oui	Oui	Non. Données déjà liées	Oui	Oui	Pas avec l'ECM	Oui	
Interopérabilité avec d'autres services ou bases de l'institution (annuaires, BDD cartographique, BDD projet, entrepôt de données -> entre Horizon et NumeriSud par exemple)	Interopérabilité	Oui	Non	Oui		Oui	Oui	Oui	Oui	Pas d'intérêt ici	NSP	Oui	
Interopérabilité avec le CRIS (Current Research Information System)	Interopérabilité	NSP	Non	Oui		NSP	NSP	NSP	Oui	Non	NSP	NSP	
PRODUCTION													
Saisie (semi) automatique ou manuelle ?	Saisie		Oui	Oui		Oui	Oui	Semi-automatique	Les deux	Les deux	NSP	Les deux	
L'outil offre-ti-il une solution d'archivage sécurisée et pérenne?	Archivage	Oui	NSP	NSP		NSP	NSP	PDF	Non	Oui	NSP	L	
Y a-t-il une communauté d'utilisateurs ? de développeurs	Communauté	Oui	Oui	Oui		Outil trop jeune pour répondre	Outil trop jeune pour répondre	Non	Oui	Oui	NSP	Oui	
Panier	Fidélisation		Non	Oui		NSP	NSP	Oui	Oui	Non	Oui	Oui	
Import de notices à partir de l' identifiant ORCID	Format d'import	Non	A venir	Oui		Oui	Oui	Oui	Non	Non	NSP	NSP	
Import de notices à partir de l' identifiant DOI	Format d'import	Oui	Non			Oui	Oui	Oui	Non	Oui	NSP	Oui	
Import de notices à partir de l' identifiant (PUBMED) PMID	Format d'import	Non	Oui	Oui		NSP	NSP		Non	Oui	NSP	Oui	
Import de notices à partir de l' identifiant ISBN	Format d'import	Non	Non	Non		NSP	NSP		Non	Non	NSP	Non	
Import de notices à partir de l' identifiant arXiv	Format d'import	Non	Non			NSP	NSP		Non	Non	NSP	Non	
Import de notices à partir de l' identifiant du WOS Clé UT	Format d'import	Non	Non	Oui		NSP	NSP		Non	Non	NSP	Non	
Intégration de l'API Sherpa/RoMEO	Open access	Non	Non	Oui		Oui	Oui	Non, pas d'intérêt ici	Non	Non	NSP	Non	
Gestion automatique des levées d'embargo	Open access	Oui	Non	Oui		Non	Non	Non	Non	Oui	NSP	Oui	
Possibilité de stocker en accès réservé	Open access	NSP	Non	Oui		Oui	Oui	Oui	Oui	Oui	Oui	Oui	
Gestion des workflows	Workflow		Oui	Oui		Oui	Oui	Oui - 2 niveaux	Oui	Oui	Oui	Oui	

	Nom de l'entité interviewée	Editeur	Inserm sur iPubli	DSpace	L'Université de São Paulo sur BDPI (production scientifique des chercheurs de l'Université de Saõ Paulo)	Ined sur ArchIned	MyScienceWork	Ifremer sur Archimer	IRD sur Horizon	CCSD	INRA sur ProdInra	Institut Pasteur sur HAL - Pasteur	EBSCO
Gestion automatique du dédoublonnage	Dépôt	Oui	NSP	Oui		En cours	En cours	Non	Oui	Oui	Oui	Oui	
Possibilité d'édition d'un dépôt par tous les auteurs mentionnés	Dépôt	Oui	Non			NSP	NSP	Oui	Non	Oui	NSP	Oui	
Versement automatique dans Hal	Connexion autres plateformes OA	Oui	Non	Oui		Oui	Oui	Non	Non		NSP		
Versement automatique dans PubMed	Connexion autres plateformes OA	Oui	Non	Oui		NSP	NSP	Non	Non	Oui	NSP	Oui	
Versement automatique dans Repec	Connexion autres plateformes OA	NSP	Non	oui, mais à vérifier		Oui	Oui	Non	Non	Oui	NSP	Oui	
Service SSO / LDAP	Authentification	Oui	Non	Oui		Oui	Oui	Oui	Non	Oui	Oui	Oui	
Accès différenciés (public, identifié)	Authentification	Oui	Oui	Oui		Oui	Oui	Oui	Oui	Oui	Oui	Oui	
Gestion des autorités auteurs	Référentiels	NSP	Non	Oui		Oui	Oui	Non	Non	Oui	NSP	Oui	
Gestion automatique des affiliations	Référentiels	NSP	Non			Oui	Oui	Non	Non	Oui	NSP	Oui	
Autres référentiels: - Equipe, projet, enquête, autre…	Référentiels	Oui	Non	NSP		Oui	Oui	Non	Non	Oui	Oui	Oui	
Passerelles entre les référentiels maison et référentiels externes (avec HAL par exemple ou ABES IDRef)	Référentiels	Oui	Non	NSP		Oui	Oui	Non	Non	Oui	Oui	Oui	
Export de notices en format Bibtex (donc Zotero)	Format d'export	Oui	Oui	Oui		Oui	Oui		Oui	Oui	NSP	Oui	
RIS (donc Zotero)	Format d'export	Oui	Oui	Oui		Oui	Oui	Oui	Oui	Oui	NSP	Oui	
Export de notices en format TXT	Format d'export	NSP				Oui	Oui	Oui	Oui	Oui	NSP	Oui	
Export de notices en format RTF	Format d'export	NSP				Oui	Oui	Oui	Non	Oui	NSP	Oui	
Export de notices en format Excel; CSV	Format d'export	NSP	Oui	Oui		Oui	Oui	Oui	Non	Oui	NSP	Oui	
Export de notices en format Word	Format d'export	NSP				Oui	Oui	Oui	Non	Non	NSP	Non	
Export de notices en format MODS	Format d'export	NSP				Oui	Oui		Oui	Non	NSP	Non	
Export de notices en format Dublin Core	Format d'export	NSP	Oui	Oui		Oui	Oui		Oui	Oui	NSP	Oui	
Export de notices en format EndNote	Format d'export	NSP	Oui	Oui		Oui	Oui	Oui	Oui	Oui	NSP	Oui	
Export de notices en format EndNote XML	Format d'export	NSP	Non			Oui	Oui		Oui	(XML TEI)	NSP	(XML TEI)	
Export de notices en format de reporting	Format d'export	NSP	Non			Oui	Oui	Oui	Non	Non	NSP	Non	
RECHERCHE													
Quel moteur ?	Moteur de recherche	NSP	SolR	SolR		Elasticsearch	Elasticsearch	SolR mais sera bientôt remplacé par ElasticSearch	Fullcrum	SolR??	NSP	SolR??	
Facettes	Moteur de recherche	NSP	Oui	Oui		Oui	Oui	Oui	Non	Non	Oui	Non	
Recherche avancée	Moteur de recherche		Oui	Oui		Oui	Oui	Non	Oui	Oui	Oui	Oui	
Recherche en plein texte	Moteur de recherche		Non	Oui		Pas explicite	pas explicite	A partir de Google mais pas d'Archimer	Oui	Oui	NSP	Oui	
DIFFUSION													
Veille via flux RSS ciblé sur listes individuelles de publications, ou les listes des UMR, ou les listes thématiques (pour les commissions d'évaluation, par exemple)	Fidélisation	Oui	Non			Oui	Oui	Flux RSS mais sans ciblage	Flux RSS mais sans ciblage	Flux RSS mais sans ciblage	NSP	Flux RSS mais sans ciblage	
Flux RSS			Non			Oui	Oui						
Abonnement à des requêtes personnalisées (thématique, géographique…)	Fidélisation	Oui	Non	Non		NSP	NSP	Non	Non	Non	NSP	Non	
Envoi de mail automatique selon thématiques de nouvelles publications	Fidélisation	NSP	Non	Oui		NSP	NSP	Non	Non	Non	NSP	Non	
Production de rapports bibliographiques	Listes dynamiques	Oui	Non			Oui	Oui	Oui	Oui	Oui	NSP	Oui	
production de CV en ligne	Listes dynamiques	Oui	Non			Oui	Oui	Oui	Oui	Oui	NSP	Oui	
Permalien sur les PDF	URL	NSP	Non			NSP	NSP	Oui	Oui	Oui	NSP	Oui	
Permalien sur les notices	URL	Oui	Oui (Handle)			Oui (Handle ou ARK)	Oui (Handle)	Oui	Oui	Oui	NSP	Oui	
Faire des urls embarquées qui pourraient s'intégrer dynamiquement dans les pages personnelles des chercheurs (pour les listes individuelles de publications, ou les listes des UMR, ou les listes thématiques)	URL	NSP	Non	NSP		Oui	Oui	Oui	Non	Oui	NSP	Oui	
Protocole Sitemap	Indexation	Oui	NSP	Oui		Oui	Oui	Oui	Oui	NSP	NSP	NSP	
Indexation des listes dynamiques	Indexation	Oui	Non			Oui	Oui	Non	Oui		NSP		
Référencement Google	Indexation	Oui	Oui	Oui		Oui	Oui	Oui	Oui	Oui	NSP	Oui	
VALORISATION													
Lien avec le SI général de l'institution : base bibliométrique, demandes de moyens, autre…	Bibliométrie	Oui	Non			Oui	Oui	Oui	Oui	Non	NSP		
Indicateur de citation	Bibliométrie	NSP	Non			Oui	Oui	Oui	Non	Non	Oui	Non	
Indicateur de téléchargement en PDF	Bibliométrie	NSP	Non			Oui	Oui	Oui	Oui	Oui	NSP	Oui	
Indicateur de consultation en html	Bibliométrie	Oui	Non			Oui	Oui	Non	Oui	Oui	NSP	Oui	
Bulletin de veille (WOS entre autres)	Bibliométrie	NSP	Non			Oui	Oui	Non	Oui	Non	NSP	Non	

3.3 Scénarios possibles

Quels sont les scénarios possibles pour l'IRD?
Aucune application ne permettant de répondre à tous les besoins, les utilisateurs ont recours à une combinaison de solutions logicielles mises en place en interne, ou par un prestataire externe.
Le choix qui parait vraiment important est la façon dont sera organisée la gestion de l'archive et à qui elle sera confiée, l'importance qui lui sera accordée, et les moyens que la gouvernance décidera de mettre en œuvre. Cette réflexion autour d'une stratégie parait incontournable.
D'autres questions doivent être débattues bien en amont dans la vie du projet.
La gestion de l'AOI sera-t-elle confiée à un prestataire externe ou au service informatique?
Quel partenariat faut-il mettre en œuvre entre le service informatique et le service documentation? Et quelle implication pour le service informatique?
Le logiciel choisi doit-il être libre ou propriétaire?
Faut-il absolument viser une solution tout-en-un ou une superposition de briques ?
Le SID mis en place pour gérer l'AOI de l'IRD aujourd'hui repose sur un SIGB propriétaire, enrichi d'une base MySQL.

3.3.1 Maintien du système

L'IRD peut d'abord choisir de maintenir ce système mais à certaines conditions:
La force de ce système repose tout d'abord sur la prise en charge de la maintenance et des montées de versions du SIGB Cadic Intégrale. Les potentiels problèmes sont pris en charge par les techniciens. Le service informatique de l'IRD est ainsi déchargé de la maintenance d'un système complexe.
Sa valeur s'appuie aussi et surtout sur des développements internes sur-mesure implémentés par l'ancien directeur du service IST, qui a pu proposer des solutions répondant exactement aux besoins des administrateurs du SID, que des personnes alimentant la base Horizon, et des utilisateurs.
Enfin, cette solution permet d'assumer la spécificité de l'archive Horizon, qui consiste en un catalogue de fonds papier, et une AOI.

Toutefois, cette stabilité ne concerne que l'application CADIC Intégrale.
La base MySQL ne bénéficie pas d'une équipe technique consacrée à sa maintenance. La faiblesse de ce système est indéniablement l'absence de documentation technique relative à cette application, et l'opacité relative à son fonctionnement. En cas de défaillance du système, les documentalistes et les informaticiens ne disposent pas de références suffisantes pour la solution d'un problème.
En effet, les flux circulants entre les deux applications, sur la base de scripts ne sont parfois pas visibles, difficiles à situer au niveau des serveurs.
Ce système ne peut donc être conservé qu'à la condition de mandater un prestataire externe pour réaliser un audit technique complet de rétro-ingénierie, qui aura pour but d'analyser le système reposant sur la base MySQL pour en identifier les composants et leurs relations, et représenter le tout de manière compréhensible.
Cette solution présenterait néanmoins l'avantage d'imposer un budget similaire à celui dépensé jusqu'à présent. Le coût de l'audit ne serait que ponctuel, si le service informatique prenait ensuite en charge la maintenance de la base MySQL.
La maintenance de cette base pourrait bien sûr être confiée à un prestataire, mais cette option aurait pour conséquence d'alourdir le budget de maintenance globale, puisque le coût de maintenance serait doublé.
En revanche cette solution permettrait-elle de suivre les évolutions que vit le monde de l'IST actuellement?
Les SIGB classiques n'offrent pas toutes les fonctionnalités requises pour répondre aux nouveaux enjeux de l'IST, et même si la rétro-ingénierie permettait une plus grande maitrise

du système, une superposition de briques, si elles n'étaient pas toutes gérées par la même équipe de maintenance rendrait le système plus fragile.
D'autre part, le marché offre aujourd'hui des produits aux caractéristiques propres aux exigences de l'Open Access.
Enfin, les projets IST prônent l'utilisation de logiciels open source, et le partage de développements qui pourraient profiter à la communauté des AOI toute entière, ce qui implique nécessairement une forte implication de l'équipe informatique au moins dans la gestion de la base MySQL et la proposition de nouveaux scripts.
L'utilisation de solutions propriétaire présente elle aussi des risques, les éditeurs étant exposés aux aléas de cession d'activité, fusions ou autres opérations économiques pouvant faire disparaitre une solution dont le code est fermé et donc non réutilisable.
Le SID pourrait par conséquent être conservé tel quel. Toutefois, il existe une réelle volonté de le faire évoluer, ce qui fera l'objet d'une seconde hypothèse.

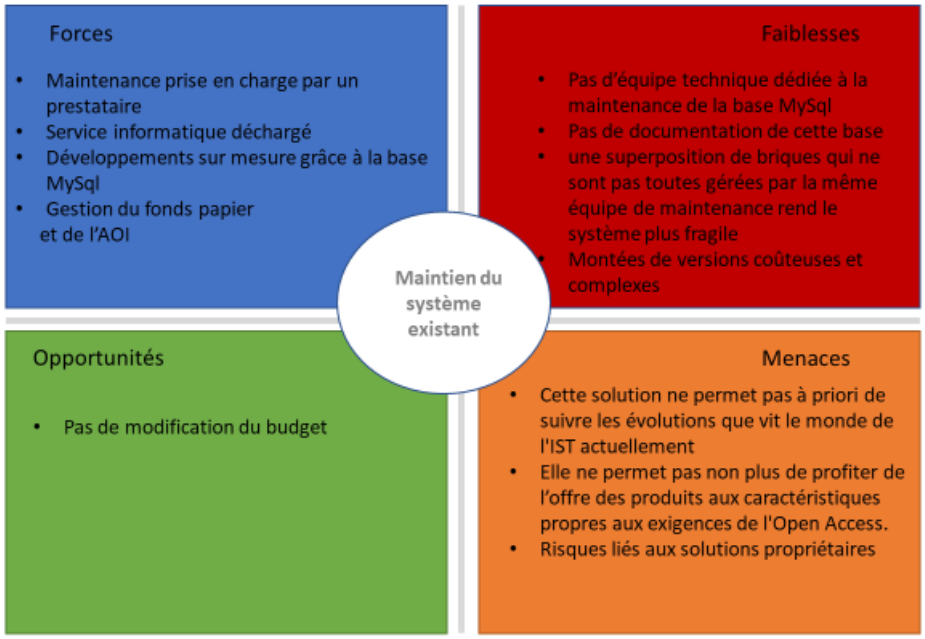


Figure 7 : SWOT du cas N°1

3.3.2 Enrichissement du système

L'IRD peut choisir de maintenir ce système et de l'enrichir:
En considérant qu'une rétro-ingénierie soit réalisée, le SID existant, qui permet déjà de gérer une AOI et un catalogue de fonds papier, mais aussi de produire des statistiques de consultation et de bibliométrie, pourrait être enrichi de modules supplémentaires afin d'apporter de nouvelles fonctionnalités.
Le système gère déjà un entrepôt OAI qui moissonne toute la base avec des métadonnées exposées en Dublin Core.
C'est ce qu'ont réalisé Ifremer, et la bibliothèque de l'université de São Paulo.
Les modules pouvant être implémentés sont (liste non exhaustive):

- **Un moteur de recherche**, tel SolR ou ElasticSearch, qui permet une recherche à facettes, la gestion de synonymes, des filtres, des recherches avancées, la présentation des résultats par pertinence, dates, ou autres. Ces logiciels permettent de référencer et d'interroger plusieurs bases. L'Ifremer par exemple a choisi le moteur SolR (avant d'opter récemment pour ElasticSearch) pour améliorer les fonctions de recherche documentaire, car le moteur proposé par Oracle, le logiciel

gérant Archimer, ne parvenait pas à répondre aux besoins d'une recherche très avancée.
L'université de São Paulo quant à elle a choisi d'emblée Elasticsearch car ce moteur a la capacité d'intégrer Kibbana, décrit ci-dessous.

- Un**outil de visualisation**. Kibbana, choisi par l'université de São Paulo en est un exemple. C'est un plugin compatible avec Elasticsearch fournissant des fonctions avancées de visualisation pour rendre compréhensibles des flux de données de taille conséquente et complexes à travers une représentation graphique.

- **Un outil de découverte** (discovery tool) permettant une recherche en ligne à la fois sur le catalogue de la bibliothèque, et sur les métadonnées des ressources numériques acquises par la bibliothèque.

Cette solution présente l'avantage d'apporter des développements sur mesure, grâce à des outils performants. Ces outils peuvent en effet s'appuyer sur l'entrepôt OAI déjà disponible pour la base Horizon, et proposer ainsi des fonctionnalités très intéressantes.
Les moteurs de recherche disponibles aujourd'hui, ainsi que les outils de découverte offrent aux utilisateurs des recherches façon Google.
Les outils de visualisation tels Kibbana procurent des statistiques de consultation pointues et permettent de générer des rapports qui répondent exactement aux besoins de l'établissement, comme l'a fait l'université de São Paulo.
Les besoins des utilisateurs et des administrateurs du système sont assurés grâce à des compétences en interne qui non seulement connaissent les nouveautés offertes par la technologie, mais savent en plus les mettre en place.

Toutefois, ceci suppose des compétences pointues en interne, au mieux au niveau du service IST, mais dans tous les cas indispensables au niveau de la DSI.
D'autre part, les montées de versions de systèmes très personnalisés peuvent se montrer complexes et coûteuses.
Enfin, de tels systèmes doivent être de préférence le résultat d'un travail collectif, dans le meilleur des cas. S'ils sont le résultat d'une seule personne, il faut impérativement que cette personne transfère ses compétences. En effet, le danger de ce type de système, construit sur mesure, se rencontre en cas de départ de la personne auteure, comme le vit l'IRD actuellement.

Cette solution présente en revanche l'opportunité pour l'IRD, par la puissance de ces outils, de participer pleinement à la communauté des AOI en proposant des solutions technologiques pour assurer une meilleure diffusion des connaissances scientifiques.
Nous avons rencontré ce cas lorsque nous avons interrogé l'administrateur de la bibliothèque numérique de l'université de São Paulo. Cette personne a été employée en tant que documentaliste mais est également dotée de grandes compétences informatiques, et a su mettre en place une plateforme riche et performante.
L'université a commencé la numérisation et l'exposition de ses archives dès 1985, à peu près en même temps que l'IRD. L'administrateur du SID permet aujourd'hui à l'université de rester dans la position de pionnière en matière de gestion d'AOI.
Nous avons aussi rencontré cette solution avec Archimer. Il s'agit en l'occurrence d'un logiciel propriétaire, enrichi par des développements maison, grâce à l'intervention d'un informaticien travaillant en interne. Nous nous trouvons ici dans une situation très similaire à celle de l'IRD. Le système gérant Archimer a, en revanche, été documenté.
Ces solutions, lorsqu'elles sont correctement documentées, sont des références précieuses pour la communauté des AOI.
L'IRD s'est aussi montrée précurseur dans la création des archives ouvertes et est considérée comme référence. La gouvernance décidera-t-elle de suivre ce chemin ?

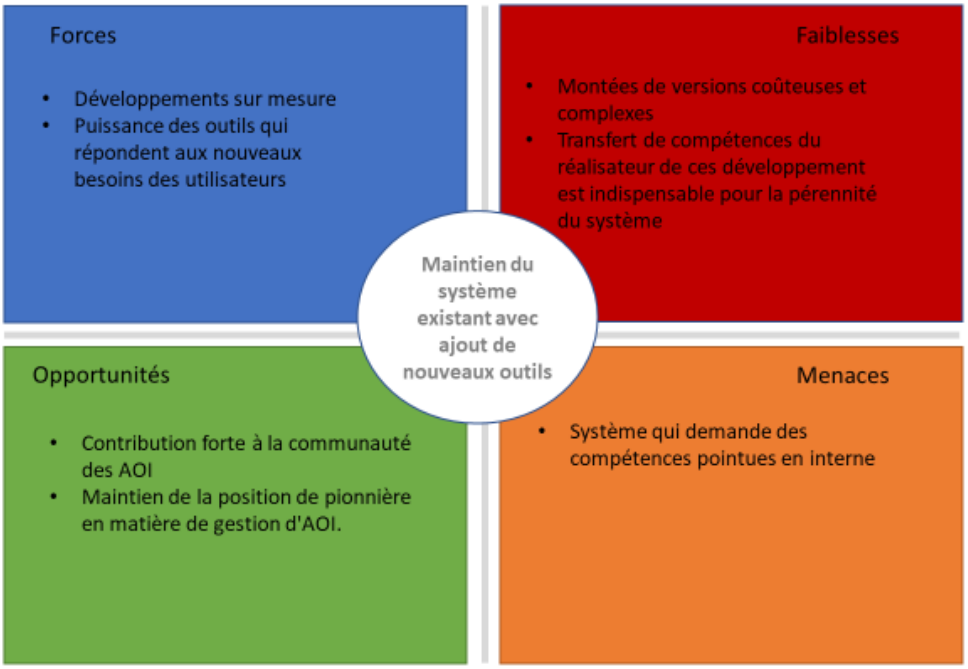


Figure 8 : SWOT du cas N°2

3.3.3 Adoption d'une solution open source reconnue

L'IRD peut choisir d'abandonner CADIC Intégrale pour adopter une solution open source.
L'IRD peut en effet choisir de se rapprocher des choix de la communauté des AOI en se dirigeant vers des logiciels déjà populaires, tel Dspace, Eprints, Fedora…
Nous n'avons pas trouvé d'utilisateurs de Eprints ou Fedora parmi les organisations que nous avons interviewées, cette analyse s'applique donc plus à la solution logicielle proposée par Dspace.
Lors de la journée des Archives Ouvertes Institutionnelles organisée par Sciences Po en juin 2018, nous avons pu constater que de nombreuses universités avaient adopté cette solution pour gérer leur répertoire d'archives ouvertes.
Nous avons interrogé l'INSERM qui utilise Dspace pour exposer ses collections éditoriales. L'usage n'est donc pas destiné à présenter ses publications, qui sont mises sur HAL, mais leur retour nous a permis de réaliser l'existence d'une grande communauté d'utilisateurs.
La communauté autour de cette application est donc forte, offrant un véritable appui pour les utilisateurs, et un vrai travail autour des solutions offertes à la communauté des AOI, ce qui va dans le sens d'une volonté de fédérer de l'activité destinée à trouver des outils de gestion d'AOI.

Toutefois, comme nous l'avons vu, open source n'est pas synonyme de gratuité, et les coûts d'installation puis de maintenance de l'application sont proches de ceux de SIGB propriétaires comme CADIC Intégrale. Dans un cas, le travail est assuré par un prestataire externe, et dans l'autre par la société éditrice, mais le budget à prévoir reste sensiblement égal.
Les montées de versions s'avèrent également compliquées et coûteuses, notamment lorsque des fonctionnalités spécifiques au fonctionnement de l'établissement ont été apportées. L'université de São Paulo nous a rapporté que la maintenance sur la partie personnalisée était très compliquée.
De plus, Dspace ne permet pas de gérer le catalogue d'un fonds papier. Si l'IRD choisit ce système, il faudra en plus installer un SIGB. De manière générale, nous constatons que les

logiciels propriétaires maîtrisent mieux les aspects liés aux fonctionnalités propres aux SIGB, comme la gestion d'un OPAC, des prêts ou encore des acquisitions.

Cette solution s'avère donc risquée en termes financiers car elle implique un double coût.
Enfin, le code de certains logiciels open source, en dépit de leur ouverture, reste relativement verrouillée et difficilement déchiffrable, même par un service informatique compétent.
C'est le cas pour Dspace notamment. La plateforme PLUME, présentée dans la partie précédente, fait état d'un « *logiciel puissant relativement difficile à maîtriser, à mettre en œuvre et à maintenir. Différentes sociétés de service spécialisées et des développeurs freelance offrent leurs services pour seconder les services informatiques*[54] ».
L'utilisation de logiciels open source dont le code est difficilement maîtrisable représente un risque de dépendance au prestataire, ce qui pourrait représenter des difficultés supplémentaires pour les partenaires des pays émergeants, disposant de peu de moyens. L'université de São Paulo qui utilise Dspace, a énuméré les limites de ce logiciel, et en particulier un code dont la structure est très complexe.
Certains clients potentiels peuvent également se montrer prudents face à une solution disposant d'un certain monopole du fait de son succès.

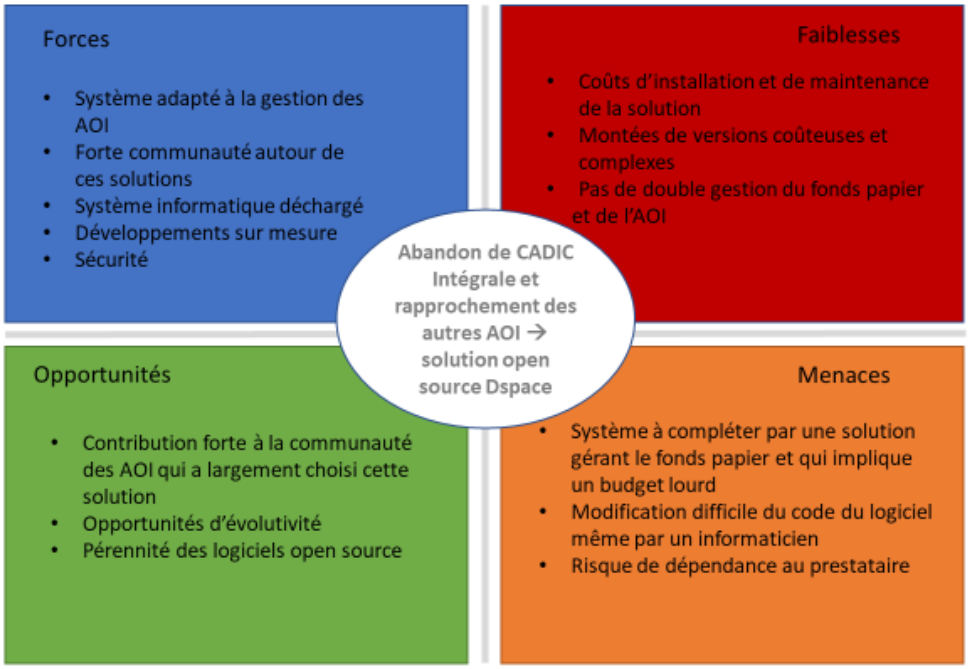


Figure 9 : SWOT du cas N°3

3.3.4 Adoption d'un système open source alternatif

L'IRD peut choisir de s'orienter vers d'autres types de logiciels open source
En effet, certaines applications permettent de gérer des AOI, mais aussi un fonds papier.
C'est le cas pour deux éditeurs que nous avons interrogés : PMB et POLARIS OS.
La société EBSCO est sur le point de présenter une solution open source présentant des fonctionnalités très intéressantes. Nous ne présenterons pas cette solution ici car nous n'avons pas obtenu suffisamment d'informations. La solution devrait sortir courant 2019.

[54] https://www.projet-plume.org/fiche/dspace

Nous avons pu nous entretenir avec le service commercial des outils proposés par PMB et POLARIS OS.
PMB a été lancé en 2001 et est largement utilisé par les bibliothèques. Sa communauté d'utilisateurs est donc forte et établie. D'autre part, l'éditeur a récemment pris la route du web sémantique en se plaçant parmi les premiers dans la FRBérisation de son logiciel. La question d'interopérabilité est donc prise en compte.
POLARIS OS de son côté est une solution développée par MyScienceWork, basée sur la donnée en proposant des modèles de données entièrement flexibles. Le logiciel s'inscrit également dans le web sémantique grâce à son interopérabilité, et se présente comme une solution nouvelle génération, qui rend les archives interopérables.
La gestion des formats de données est une raison importante qui a conduit L'INED à choisir l'application POLARIS OS.
Comme toute jeune solution, POLARIS OS ne bénéficie pas encore d'une large communauté d'utilisateurs, mais est actif pour se faire une place dans la communauté des AOI.
Ces deux solutions se décrivent ouvertes, afin de permettre ainsi à un service informatique ou même un administrateur de système d'avoir la main sur certains développements.
La fiche Plume dédiée au logiciel PMB décrit une application qui « *même si son installation ne nécessite pas particulièrement de connaissances pointues en informatiques, mieux vaut cependant travailler en collaboration avec un informaticien pour éviter des écueils au démarrage du projet* ».

La force des logiciels open source se situe dans l'association entre communauté d'utilisateurs et éditeurs qui permet d'enrichir ces applications offrant par la même une réelle pérennité à ces systèmes, et des opportunités régulières d'évolutivité.
Lorsque l'équipe informatique en interne a la capacité de développer de nouvelles fonctionnalités, l'institut peut participer aux groupes de travail liés aux progrès technologiques permettant aux archives ouvertes une meilleure visibilité, garantissant une voix à l'institut et une plus large diffusion de sa production scientifique.
Contrairement au marché des logiciels propriétaires, le marché de l'open source est en plus à l'abri d'une cession d'activité ou autre changement qui surviendrait dans l'entreprise, le code étant ouvert, l'application peut continuer d'être utilisée et enrichie par la communauté.
Enfin, les logiciels open source présentent un réel gage de sécurité, car lorsqu'un problème technique survient, la communauté d'utilisateurs travaille d'emblée à identifier et corriger le bug en temps réel.
L'avantage d'un logiciel récent est aussi d'intégrer les problématiques actuelles. POLARIS OS intègre ainsi une solution bibliométrique qui permet à l'INED de n'avoir à gérer qu'un seul outil pour manier à la fois les rapports bibliométriques et l'AOI.

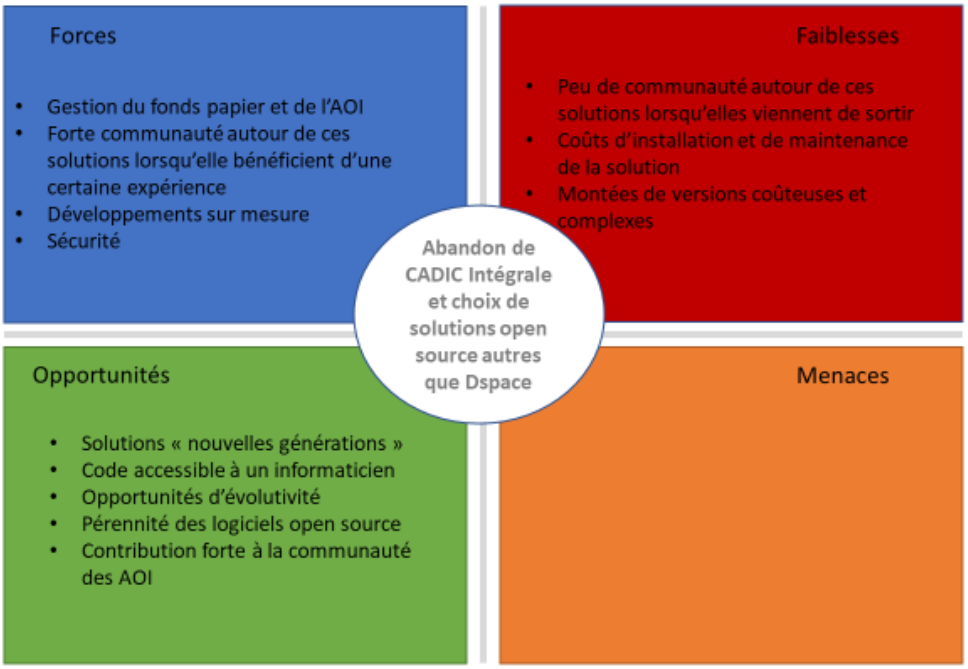


Figure 10 : SWOT du cas N°4

3.3.5 Le choix de la mutualisation

L'IRD peut enfin faire le choix de verser l'intégralité de sa production scientifique sur HAL et fermer Horizon.
HAL compte aujourd'hui 143 portails. L'IRD en fait partie et dépose ses publications dans HAL, mais les publications qui y sont exposées ne sont pas exhaustives, et l'institut n'impose pas d'obligation de dépôt à ses chercheurs.
Nous avons interrogé l'INRA, qui a pris la décision de fermer son archive ouverte Prodinra pour basculer l'intégralité de ses publications sur HAL.
D'autre instituts ont adopté cette solution dès la mise en place de leur archive ouverte

La décision de verser l'intégralité des publications scientifiques de l'institut dans HAL est d'abord politique. Elle résulte du choix de la gouvernance de l'institut de s'en remettre aux outils de mutualisation proposés par le CCSD créé par le CNRS en 2000et de participer à l'archive nationale, comme l'INRA, mais aussi de créer une plate-forme d'archive ouverte répondant à la politique de l'open access, comme c'est le cas à l'institut Pasteur.
La tendance actuelle est à la mutualisation, nous l'avons vu dans la deuxième partie avec des initiatives qui se développent de part et d'autre et notamment le recours à une plateforme qui centralise l'ensemble des publications.

La deuxième raison qui motive les instituts ou universités à centraliser leur production scientifique sur HAL est indéniablement la réduction des coûts engendrés par la gestion d'une archive qui repose sur un logiciel dont l'installation et la maintenance nécessitent un réel budget.
Tous les coûts ne disparaissent pas totalement pour autant, notamment dans le cas de l'IRD qui aura besoin dans tous les cas de s'équiper d'un SIGB pour gérer ses acquisitions.
Chaque institut dispose d'un portail propre, et peut personnaliser, dans une certaine mesure, sa présentation.

De plus, toutes les fonctionnalités liées à la gestion d'une archive ouverte sont possibles, de façon plus ou moins satisfaisante, notamment la levée d'embargo automatique et la gestion des affiliations.
En outre, le service informatique est déchargé de cette tâche, les problèmes techniques étant entièrement pris en charge par le CCSD.
En revanche, la communauté d'utilisateurs est très active dans le développement de nouvelles fonctionnalités. Les laboratoires, universités, sont producteurs de nouveaux développements destinés à améliorer l'utilisation de la plateforme. On trouve les résultats de ces travaux sur une page wiki mise en place par Hal[55].
Les informaticiens peuvent par conséquent être sollicités pour participer à ce travail collectif. Cette décision appartient à l'organisme utilisateur, et aucune obligation n'existe, mais dans un contexte de mutualisation d'efforts collectifs, cette participation n'est-elle pas nécessaire, surtout pour un institut, comme l'IRD, pionnier en matière d'exposition de sa production scientifique?
D'autre part, certains partenariats qui se jouent avec le CCSD peuvent être conclus sur la base d'un "don de main d'œuvre", comme ce qui a été entrepris à l'INRA, qui a mis l'un de ses informaticiens à disposition du CCSD. Cette mise à disposition d'un personnel INRA à temps plein dans les équipes du CCSD s'ajoute à la participation directe de cette institution dans la gouvernance du centre pour la communication scientifique directe.

Enfin, cette solution ne fait pas l'économie d'une équipe entière de documentalistes chargés d'alimenter la plateforme, avec tout le travail que cela suppose. L'INRA, comme l'institut Pasteur qui a toujours en charge la vérification des affiliations, des métadonnées du dépôt, du texte intégral, de la licence, ainsi que de la gestion des laboratoires de l'institution dans le référentiel AureHAL. L'INRA de son côté garde le contrôle la validation de chaque dépôt.
Or, certaines tâches peuvent être plus fastidieuses du fait de lacunes liées à certaines fonctionnalités. La gestion des affiliations notamment représente un travail considérable, et pourtant incontournable. Le module gérant les affiliations ne fonctionne pas parfaitement, et est source d'erreurs qui doivent être ensuite corrigées manuellement.
Le contrôle des métadonnées représente de même un temps de travail humain non négligeable et repose complètement sur les travailleurs locaux. Cette tâche est difficilement contournable, surtout à l'heure des archives nouvelles générations qui exige des métadonnées riches et de très bonne qualité pour être interopérables. Un contrôle qualité est donc indispensable en amont, comme le fait l'IRD, dont les documentalistes vérifient chaque notice avant de la verser dans HAL.

La force de la plateforme HAL ne se situe donc pas dans sa technologie qui est loin de dépasser celle des logiciels de gestion d'archives ouvertes cités plus haut.
Certaines fonctionnalités ne sont pas encore développées, et dans ce cas doivent être redéveloppées en interne, ce qui représente un coût ;
Les workflows, par exemple, ne sont possibles que sur deux niveaux, alors que les systèmes propres aux instituts de recherche proposent plus de niveaux, comme c'est le cas à l'INRA, qui va devoir adapter sa procédure. Prodinra permet en effet trois niveaux de workflows et ce système permet de rendre la notice accessible en interne pour plusieurs milliers d'utilisateurs dans un délai très court, même si son statut est encore en cours de validation, alors qu'une notice sur HAL ne sera disponible qu'après validation finale et définitive.
De même, les indicateurs bibliométriques proposés par Hal sont assez sommaires, même si le CCSD est en train de retravailler sur le module de statistiques, cela suffira-t-il ? . L'INRA va ainsi développer une solution en interne pour la production d'indicateurs bibliométriques.

[55]Fiche consultable sur la page Wiki du CCSD

Dans le système actuel, l'outil OCDHAL permet d'exporter des données à partir du portail qui peuvent ensuite être retravaillées dans des tableaux croisé dynamiques, ce qui suppose une main d'œuvre en interne spécialisée en bibliométrie, une compétence pointue et coûteuse.
Si HAL permet de récupérer les dépôts pour être exploités par les systèmes d'information propres aux institutions (site web, catalogue des bibliothèques, logiciel de bibliométrie), c'est à la condition que celles-ci disposent d'un budget pour mettre en place un SID en interne. Or, la principale motivation des institutions à verser leur production scientifique dans HAL plutôt que de choisir leur propre AOI est économique.
Il reste néanmoins possible pour les instituts d'articuler Hal avec leur propre AOI, et notamment d'intégrer leur annuaire LDAP et les référentiels RH internes afin que les utilisateurs puissent s'authentifier et accéder à un document en fonction des droits d'accès attribués à ce document.

Hal constitue malgré tout un bon outil de valorisation de la production scientifique, en centralisant les publications des chercheurs, permettant à un utilisateur de trouver plusieurs réservoirs sur une même plateforme, même s'il n'avait pas prévu de les consulter au départ.
Une limite est à prendre en compte toutefois, les utilisateurs ont bien conscience que Hal est surtout un réservoir de notices, et que l'on ne trouve pas forcément le texte intégral. Un chercheur qui souhaite accéder directement au texte intégral ira-t-il consulter la plateforme avant de se rendre directement sur l'AOI d'une institution ?
Notons toutefois que le plan pour la science ouverte annoncé par le ministère cet été a annoncé la mise en œuvre de moyens pour améliorer Hal. Des moyens humains seront-ils développés ?
Le temps de traitement d'un dépôt sur Hal prend en effet plus de temps et sa mise en ligne est pour l'instant plus lente que lors d'un dépôt sur Horizon.
La solution Hal reste par conséquent une option, à condition d'être clair sur ce qu'un institut qui possède déjà sa propre archive est prêt à abandonner en termes de fonctionnalités et de contrôle qualité.

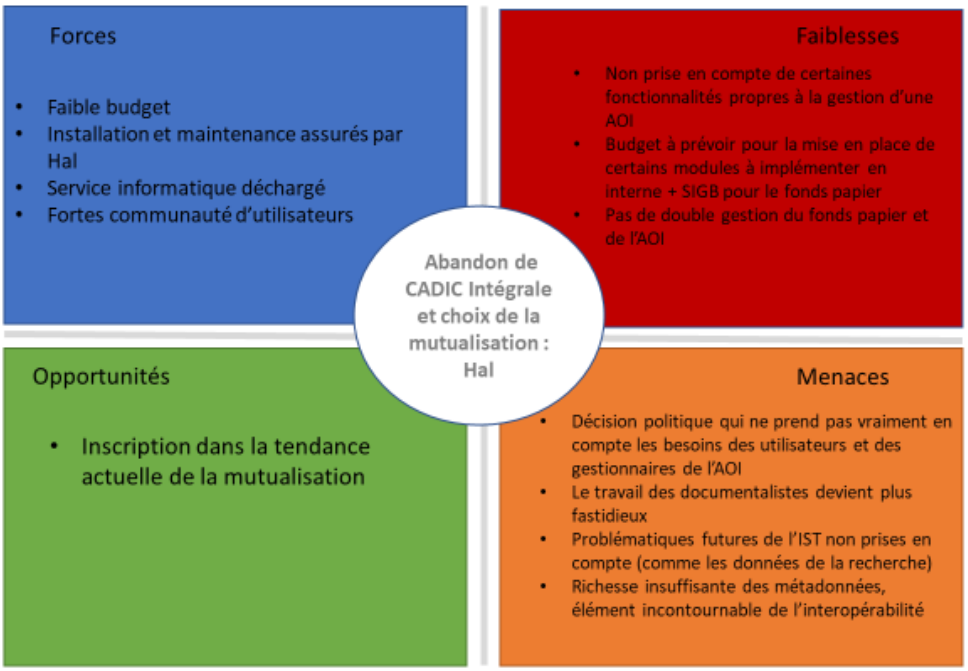


Figure 11 : SWOT du cas N°5

3.4: Préconisations

Ce comparatif entre les différentes solutions choisies par les différents instituts de recherche et universités nous montre que malgré la diversité des solutions, toutes présentent un dénominateur commun de fonctionnalités.

Ces fonctionnalités sont plus ou moins approfondies selon chaque solution. Chacune propose un système de workflows, avec des niveaux différents selon les applications. Les flux RSS sont également possibles partout, les plus basiques permettent de faire apparaitre les dernières publications, les plus avancés permettront de choisir les thématiques des dernières publications.

Les fonctionnalités communes à toutes les solutions que nous avons relevées lors de notre étude :

- Toutes les solutions peuvent être **hébergées** au choix chez le client ou chez l'éditeur qui propose également des solutions d'hébergement.
- Les solutions étudiées permettent toutes d'être moissonnées en **OAI-PMH**
- Toutes les solutions permettent de mettre en place des **workflows** plus ou moins élaborés.
- Toutes présentent la possibilité d'intégrer le système Handle d'**identifiant unique**
- La plupart des éditeurs peuvent implémenter des **modules personnalisés** par rapport aux besoins du client.
- Les profils d'utilisateurs peuvent aussi être différenciés selon qu'il s'agit d'un **accès public ou privé.**
- La plupart des solutions étudiées proposent des **flux RSS** pour faire apparaitre les dernières publications.
- Nous avons pu noter que les solutions adoptées sont souvent complétées par d'autres outils, notamment l'outil de découverte, et la grande majorité est interfaçable avec une **base MySQL interrogeable par des scripts PHP**.
- Le **référencement** est assuré dans toutes les solutions. L'utilisation du protocole Sitemap notamment permet aux notices et aux documents d'être correctement référencés par Google Scholar.

Nous avons relevé que les mêmes fonctionnalités présentent des difficultés communes aux organismes.

Ci-dessous se trouve une liste, non exhaustive, de fonctionnalités auxquelles il faut porter une attention toute particulière lors du choix de la mise en place d'un nouveau SID.

3.4.1 Recommandations techniques

- ✓ **L'alimentation d'une AOI**

Les moyens d'alimenter l'AOI doivent permettre au déposant de gagner en temps, la nouvelle solution devra apporter une réelle aide à la récupération de données par des fonctionnalités comme la saisie automatique de la notice, la possibilité de copier des notices et d'en modifier certains champs.
Le temps passé par les documentalistes à l'enrichissement des notices est important, et la nouvelle solution devra permettre d'automatiser certaines de ces opérations.
Le SID devrait laisser la possibilité aux chercheurs de déposer eux-mêmes leurs publications sur une plateforme. Ce dépôt doit être soumis à un workflow pour vérifier que les champs renseignés l'ont été correctement, et compléter le cas échéant.

- ✓ **Les connecteurs**

Les connecteurs avec les plateformes de l'information scientifique et technique doivent être pris en compte.
Aussi, la nouvelle solution devra permettre des passerelles vers d'autres plateformes. Ces connecteurs sont nombreux, nous citerons les plus importants, c'est-à-dire les connexions à partir du numéro DOI de la publication et le compte ORCID du chercheur.
Le DOI (digital object identifier) signifie «identifiant numérique d'objet ». C'est un mécanisme d'identification de ressources.
ORCID est un code alphanumérique qui permet d'identifier de manière unique les chercheurs et auteurs de contributions académiques et scientifiques.
Il serait souhaitable de mettre en place un module de connexion ORCID qui permette une synchronisation automatique du compte ORCID du chercheur vers la base Horizon et inversement. Ceci implique que l'IRD soit membre du consortium ORCID pour pouvoir bénéficier de cette API.
Ces options sont possibles techniquement avec la base MySQL. Cette base comporte un annuaire dont un des champs consisterait en un lien qui redirigerait vers la page du chercheur.
Les partenariats avec les instances de l'IST permettent ces connexions, et nous avons vu en étudiant le SID Archimer qu'une icône ORCID renvoie vers le compte ORCID du chercheur.
En outre, si la solution ne le permet pas dans ses fonctionnalités propres, elle devra laisser l'option pour un développeur d'implémenter une API « Crossref REST » qui permettra de récupérer les métadonnées d'une publication à partir de son DOI.
L'IRD attribue également un numéro appelé FDI (fonds documentaire) à chacune de ses publications, la nouvelle solution devra permettre d'implémenter un module d'export automatique d'Horizon vers Hal à partir de ce numéro.
D'autres moyens technologiques existent pour alimenter une AOI, comme les protocoles, notamment le Z39-50 qui permet de récupérer des notices cataloguées dans le SUDOC[56].
Ceci n'est pas une liste exhaustive, les connecteurs sont multiples, et avant de choisir une solution, il convient d'établir, en amont du projet de refonte du SID, une liste des plateformes IST avec lesquelles l'IRD souhaite connecter son AOI.

- ✓ **Les affiliations et référentiels**

Toutes les entités interrogées rencontrent les mêmes difficultés, notamment au niveau de la gestion des affiliations, qui, même lorsqu'elles sont gérées par un référentiel intégré dans la solution logicielle, doivent dans tous les cas être contrôlées manuellement, notamment lors des imports en provenance du WOS. Cette tâche représente un travail humain qui demande beaucoup de temps. La façon de renseigner les affiliations par les chercheurs n'est pas uniformisée, et présente beaucoup d'erreurs. Nous avons vu notamment que la gestion des affiliations par HAL, du fait de l'auto-archivage par les chercheurs, contient des erreurs qui doivent ensuite être corrigées par les documentalistes.
Plusieurs solutions existent pour éviter toute confusion, la première consiste à former les chercheurs à renseigner correctement leur affiliation.
L'IRD a mis en place une charte de nommage des UMR affiliés ou non, il est essentiel de la rendre visible pour les chercheurs en lui accordant une place privilégiée sur l'intranet par exemple.
L'ABES et ORCID ont en 2016 conclu un protocole d'entente[57] visant principalement à encourager l'utilisation d'identifiants pérennes pour les chercheurs et leurs organisations. Il devrait résulter de cette entente le développement d'API pour d'accroitre l'interopérabilité en matière de communication scientifique.

---

[56] SUDOC est le catalogue collectif des bibliothèques de l'enseignement supérieur
[57] Memoranding of Understanding consultable sur https://fil.abes.fr/wp-content/uploads/sites/6/2016/07/mou.pdf

Les possibilités techniques existent aussi, les outils que l'on trouve actuellement sur le marché gèrent de façon plus ou moins avancée la question des affiliations, et cette fonctionnalité sera l'une des plus importantes à prendre en compte lors du choix final de la solution logicielle.
L'INED dans sa présentation de son AOI nous a montré comment les référentiels comme les projets de recherche avaient pu être intégrés. A chaque référentiel correspond un champ avec une liste déroulante. Le travail pour intégrer les listes dans les menus déroulants a été très long, mais le résultat très satisfaisant.
Les affiliations reposent sur des référentiels structurés. Or, ces référentiels sont très nombreux aujourd'hui et ceci génère une certaine confusion.
C'est le cas par exemple pour les affiliations. Certains pays, comme le Brésil ont mis en place un système *unique* d'identification des chercheurs, leur permettant ainsi de rassembler plus efficacement leurs publications.
En France, le projet Conditor[58] devrait à terme permettre d'avoir un système centralisé pour gérer les affiliations et des publications des chercheurs français.

Un autre problème se situe au niveau des passerelles entre les référentiels internes et les référentiels des plateformes, ou encore des institutions. Les liens ne sont pas évidents, comme par exemple le lien entre l'annuaire interne de l'institution, et le référentiel des affiliations.
Certains logiciels offrent la possibilité de dérouler de listes dans le champ des affiliations, comme l'a fait l'INED. Le travail en amont est conséquent, mais il facilite le dépôt.
Il est aussi possible de récupérer l'annuaire maison et de l'intégrer complètement dans la base plutôt que de le connecter.
Certains logiciels peuvent aussi aller récupérer l'information dans le LDAP qui réunirait tous les auteurs.
Le laboratoire de recherche Deuxième Labo préconise la diffusion de la norme VIVO pour rendre interopérables les annuaires de recherche en France et construire un méta-annuaire national comme le portail inter-institutionnel VIVOsearch, réunissant à titre expérimental des organismes de recherche. (Blanchard, Sabuncu, 2015)[44]
Ces choix sont à faire en amont du choix de la solution logicielle et sont primordiaux, car les indicateurs statistiques peuvent être faussés si les affiliations contiennent des erreurs.

- ✓ **La gestion des doublons**

La gestion des doublons représente aussi du temps de travail humain. La plupart des solutions offrent un module de dédoublonnage, plus ou moins performant et qui, s'il ne repose pas sur de bons référentiels, nécessitera un travail de vérification.
Même un système efficace comme celui présenté par l'Ifremer l'import est semi-automatique et le repérage des doublons est incontournable, et nécessite des corrections.
Ce module est en cours de développement par POLARIS OS, qui promet un dédoublonnage dans les publications, mais aussi au niveau des référentiels.
L'institut Pasteur nous a expliqué que HAL propose également un module, utilisable uniquement par l'administrateur de portail et envoie une alerte au déposant et au modérateur lorsqu'un doublon est repéré.

[58] Développé par la Direction de l'information scientifique et technique, le projet Conditor a pour objectif de recenser l'ensemble de la production de la communauté de l'Enseignement supérieur et de la Recherche.

- ✓ **Les fonctions de recherche**

Le moteur de recherche est un élément clé du SID et les institutions doivent parfois implémenter un moteur pour améliorer les fonctions de recherche documentaire car ceux proposés par les bases de données ou les SIGB ne permettent pas des recherches très poussées.
Les moteurs classiques permettent en effet de lancer des recherches sur les mots-clés avec lesquels le déposant a indexé la publication.
Ces moteurs indexent tous les champs de la notice, ainsi que le texte intégral du document. L'administrateur décide ensuite quels champs seront rendus interrogeables dans son interface.
Mais les moteurs offrent aujourd'hui de puissantes fonctionnalités de recherche, outre des délais de réponse performants, et des facettes permettant de filtrer les résultats sur différents champs, ils peuvent, aussi être associés à des outils de visualisation des résultats via des graphes ou encore de statistiques sur les recherches des utilisateurs, comme nous l'a montré l'université de São Paulo.
Les moteurs présentent également des fonctionnalités de fidélisation, en proposant un historique des recherches faites.
Dans le même registre, il est désormais possible d'envoyer une notification à l'utilisateur lorsqu'une nouvelle publication correspond à une expression de recherche enregistrée.
Il faut veiller à ce que le moteur choisi puisse être interfaçable avec une base MySQL, comme le moteur StopWords, si l'IRD choisit de garder le système dual d'un SIGB enrichi d'une base de données supplémentaire.

Les organismes interrogés utilisent la plupart du temps l'un des deux principaux moteurs de recherche proposés sur le marché: SolR ou ElasticSearch. Ce dernier a l'avantage d'intégrer l'outil Kibana, qui permet de faciliter la production de rapports, et d'agréger les données sous forme de tableaux de bords bien renseignés. De plus, la maintenance de ce moteur est relativement simple pour un informaticien.
Idéalement, le moteur à facette proposera des options de filtrage par: - type de ressource - type d'articles (par exemple, articles référencés dans le Web of Science, ou Direction d'ouvrages ou de revues) - sujet - auteur - UMR - date de publication - projet de recherche - langue - nom de la revue.
Il proposera également un historique des expressions de recherche, et accompagnera les résultats d'un graphe présentant à minima les résultat de la recherche et les statistiques sur les recherches des autres utilisateurs.

- ✓ **L'archivage**

L'archivage est souvent organisé en interne par l'institution. Les solutions logicielles proposent une sauvegarde, mais pas d'archivage sécurisé et pérenne.
Les ressources sont souvent archivées au format PDF/A, car il est raisonnable de penser que ce format, largement utilisé dans le monde, sera encore lisible dans de nombreuses années.
D'autres optent pour le format XML qui peut être lisible par n'importe quel appareil, et ne nécessite pas d'installer un logiciel spécifique pour la lecture du document, comme à l'INSERM.
Les services du CINES proposent des serveurs spécifiques pour l'archivage des publications. Lors de la refonte du SID, il faudra se poser la question d'un accord avec cette institution pour mettre en place un archivage pérenne des documents d'Horizon afin de garantir une conservation sécurisée de la production scientifique dans son ensemble.

- ✓ **La bibliométrie**

Les institutions présentent chacune leurs indicateurs, mais des problèmes reviennent dans chaque cas. Les instituts de recherche n'évaluent pas non plus les chercheurs de la même façon. Certains vont utiliser des indicateurs chiffrés issus de la bibliométrie alors que d'autres

ne vont prendre en compte que des indicateurs qualitatifs, comme des appréciations par les directeurs d'unités, comme ce sera le cas à l'INRA prochainement.
L'Ifremer fonctionne déjà de cette façon carles chercheurs ne sont pas évalués sur la base d'indicateurs bibliométriques. La bibliométrie proposée a plus une visée d'appui à la recherche à travers des cartographies thématiques, des comparaisons thématiques ou institutionnelles. La bibliométrie vient donc en appui au pilotage de la recherche au niveau des unités de recherche et de l'institution.

Concernant les indicateurs quantitatifs, ce sont les systèmes d'information documentaires qui vont permettre d'extraire des chiffres, le plus compliqué consistant à isoler un corpus de documents et en tirer des informations.
Il est à noter toutefois que les institutions n'utilisent pas toutes leur système de gestion documentaire pour faire de la bibliométrie, même si c'est techniquement réalisable.
La nouvelle solution proposera un format d'export CSV pour pouvoir ensuite travailler sur les tables et sortir des statistiques.
En tenant compte d'une des spécificités de l'IRD, la pluridisciplinarité, la solution choisie devra aussi permettre de faire remonter les indicateurs métriques et bibliométriques des chercheurs en SHS de l'IRD via les outils proposés par SQL.

- ✓ **La gestion des droits de diffusion des documents**

La levée automatique d'embargo sur un fichier n'est pas toujours possible, mais certaines solutions le permettent, comme HAL ou POLARIS OS.
Les fichiers éditeurs sont stockés en accès réservé, mais le fichier auteur peut être déposé. Or, les chercheurs déposants ne sont pas toujours au fait de ces nuances, et, même s'ils en font la demande, l'éditeur ne donne pas forcement le bon fichier.
Un travail de médiation/formation est nécessaire à ce point.
La nouvelle solution devra impérativement intégrer une API Sherpa Romeo qui permettra d'attribuer une date d'embargo qui pourra être levée automatiquement.

- ✓ **Les communautés d'utilisateurs et de développeurs**

Les solutions éditrices sont plutôt actives dans le développement d'une communauté d'utilisateurs. Selon leur ancienneté sur le marché, cette communauté sera bien sûr plus ou moins importante, mais selon les pays également. Certaines communautés communiquent plus et sont plus productives d'un pays, ou d'un continent à l'autre.
Ce paramètre sera également à prendre en compte.

- ✓ **Possibilités d'évolution**

La possibilité de faire évoluer le SID en implémentant de nouveaux modules ou en greffant de nouveaux modules, développés soit en interne par les informaticiens, soit par un prestataire, ou l'éditeur lui-même est une possibilité que devra autoriser la nouvelle solution.
Dans un système composé de plusieurs briques, il est en effet important d'utiliser des outils qui peuvent facilement s'interfacer les uns avec les autres.
L'uniformisation d'un tel SID est importante. Beaucoup reposent sur des bases miroirs requérables par des scripts PHP. Idéalement les scripts doivent être automatisables et intégrables dans une solution logicielle unique afin de pouvoir les exporter lors du passage vers la prochaine solution, si la base MySQL est conservée. En intégrant une API les scripts peuvent être migrés sans avoir à être réécrits.

A l'inverse d'évolution, le nouveau SID devra aussi être réversible et les ressources devront pouvoir être reprises en cas d'expérience d'utilisation non concluante.
Enfin, il serait souhaitable que le SID puisse être testé sur quelques utilisateurs.
La force d'un SID réside dans les possibilités d'enrichissement de sa base. Certains organismes souhaitent pouvoir avoir la possibilité d'intégrer un fonds iconographique, cartographique ou autre, ou tout au moins pouvoir connecter son SID à d'autres SID de l'institution, comme l'a fait Archimer.

Il est en effet intéressant de pouvoir présenter derrière la même vitrine tous les fonds dont dispose un organisme, ou de lier les bases entre elles, comme c'est le cas à l'Ifremer.
L'IRD dispose notamment d'un fonds cartographique qui pourrait être mieux exposé s'il était intégré à Horizon.

- ✓ **Les options de fidélisation de l'utilisateur**

Les nouvelles bibliothèques numériques offrent des services personnalisés aux utilisateurs, afin de le fidéliser. Ces services sont précieux et permettent de faciliter les recherches.
La nouvelle solution devrait par exemple permettre aux chercheurs de partager des paniers avec d'autres utilisateurs lors de recherches collectives, de pouvoir contacter un documentaliste via un formulaire de contact, s'abonner aux listes individuelles de publications, d'UMR ou aux listes thématiques, permettre au chercheur d'inclure des balises ou des critiques dans ses références et marquer des publications comme favoris (avec mise en place d'un panier).
Les flux RSS sont un moyen de veiller encore très vivant et notamment dans la communauté IST. La nouvelle solution devra permettre de produire des flux RSS Horizons sur les pages d'accueil intranet et internet avec la possibilité de décliner les flux par département ou thématique.
L'utilisateur pourra également se voir envoyer une notification lorsqu'une nouvelle publication correspond à une expression de recherche enregistrée.
L'outil choisi devra pouvoir générer des URL embarquées qui iront s'intégrer dynamiquement dans les pages personnelles des chercheurs et mettre à jour leurs CV.
Cette liste n'est pas exhaustive, de nombreux services sont possibles et peuvent être imaginés.
La solution mise en place par l'université de São Paulo offre de nombreux services à ses utilisateurs, rendant leur expérience de navigation ergonomique et riche.

- ✓ **La maintenance**

La maintenance pour les solutions très personnalisées peut s'avérer très coûteuse, car elle est trop contraignante pour être confiée à un informaticien en interne. Ceci est encore plus vrai lorsque de nombreux développements ont été réalisés pour personnaliser le système documentaire.
Les organismes interrogés souhaitent en majorité des outils open source, certes, mais dont le code reste facile d'accès, et qui permettrait de personnaliser facilement la solution.
Avant de s'orienter vers une solution, l'IRD devra se décider sur le niveau de dépendance qu'il est prêt à avoir par rapport à une solution, et par extension au prestataire chargé de l'installation, de la maintenance, et des futurs développements potentiels.

- ✓ **La reprise des données**

C'est une opération complexe qui demande de réelles compétences. Un organisme souhaitant migrer son SID a tout intérêt à employer les services d'un consultant spécialisé en la matière. L'INED a rapporté avoir consacré énormément de temps et de travail à cette tâche.

- ✓ **Les appareils de connexion des utilisateurs**

Selon les pays ou les continents, les utilisateurs ne se connectent pas tous à partir du même appareil, selon leurs préférences ou les contraintes dues aux problèmes de connexion rencontrés dans certains pays, et notamment les pays émergents, partenaires de l'IRD. La nouvelle solution devra être « ATAWAD » (anytime, anywhere, anydevice) c'est-à-dire accessible de partout dans le monde, autant sur portables, téléphones mobiles que sur des PC de bureau.

- ✓ **Les formats**

Horizon gère déjà les formats classiques bibliographiques, garantit un format de données compatible avec le moissonnage. La nouvelle solution devra absolument conserver ces

spécifications et proposer en plus aux chercheurs des listes dynamiques exportables sous différents formats au moyen de web services.

- ✓ **Identification des ressources**

La nouvelle solution devra permettre de générer des URL pérennes pour les notices afin qu'elles deviennent un objet et non le résultat d'une requête.
La nouvelle solution devra garantir un système d'identifiant unique. Nous pensons à la spécification technique Handle qui permet d'attribuer, gérer et résoudre des identifiants persistants attribués à des objets numériques et à d'autres ressources Internet.

- ✓ **Recommandations de la COAR**

Enfin, nous avons consulté le site de la COAR (Confederation of Open Access Repositories) qui donne des recommandations quant aux entrepôts d'archives ouvertes.
La COAR est une association internationale comptant une centaine de membres et partenaires à travers le monde représentant les bibliothèques, universités, instituts de recherche, financeurs publics…
Sa mission consiste à réunir la communauté des archives ouvertes dans le but de rassembler des moyens, d'harmoniser les politiques et les pratiques autour de la gestion des archives ouvertes, et d'agir comme porte-parole de cette communauté. Son rôle est d'accroître la visibilité et l'application des résultats de la recherche grâce à un réseau mondial d'archives en libre accès fondés sur la collaboration internationale et l'interopérabilité.
La COAR a publié cette année un rapport présentant les fonctionnalités et les recommandations techniques pour les archives ouvertes de prochaine génération (next generation repositories). (COAR, 2017)[40]
Ce rapport préconise l'interopérabilité et la distribution des archives ouvertes.
Un système distribué ne signifie pas forcément d'aller contre un système centralisé des ressources, mais de bien définir qui est propriétaire des ressources, pour ne pas créer d'ambiguïté. Tant que la propriété d'une ressource est clairement définie et qu'un cadre légal lui est octroyé , leur concentration ne posera pas de problème.

L'interopérabilité quant à elle ne situe pas tant au niveau technique qu'au niveau de la qualité des données et des métadonnées. Avant même d'être interopérable et interconnectable, un entrepôt doit être riche.
La COAR considère que le modèle OAI-DC, utilisé par défaut, et qui garantit l'interopérabilité aujourd'hui, n'est plus assez poussé. Or, définir un profil de métadonnée plus détaillé est possible, et la technologie pour le faire existe déjà.
L'important est d'identifier les ressources en se servant de la technologie "global persistent identifier" (DOI, ORCID...).
Un format plus riche est donc en cours de création, par la COAR: le format de métadonnées RIOXX[59], déjà utilisé dans de nombreuses universités anglaises.
Ce format possèdera plus de champ que le format Dublin Core, et est recommandé par le United Kingdom Council of Research Repositories[60].
Un autre moyen de garantir l'interopérabilité est d'encourager l'utilisation d'un vocabulaire contrôlé.
Beaucoup de thesaurus sont créés, notamment avec l'essor de l'open science. Nombreux sont ceux qui décrivent un domaine très spécifique, et ceux-là sont utiles.

[59] http://rioxx.net/
[60] https://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2018/open-access-repositories-report-and-recommendations.pdf#search=open%20source

Il faut néanmoins uniformiser le vocabulaire décrivant les concepts communs au plus grand nombre. Par exemple, le concept d'accès au document, qui indique si le document est en Open Access ou en accès réservé n'est jamais nommé de la même façon selon les institutions. Ce concept répond à trop d'appellations différentes, et la COAR recommande de les uniformiser. Il faut se servir des mêmes pilotes.
Aujourd'hui, la COAR encourage à utiliser le vocabulaire SKOS, une recommandation du W3C publiée en 2009 pour représenter des thésaurus, classifications ou d'autres types de vocabulaires contrôlés ou de langages documentaires.
S'appuyant sur le modèle de données RDF, son principal objectif est de permettre la publication facile de vocabulaires structurés pour leur utilisation dans le cadre du Web sémantique.

L'utilisation de profils d'application partagés est une autre condition pour une organisation des connaissances et de la description claire des notices bibliographiques.
Selon les recommandations de la COAR, le nouveau SID devra aussi permettre d'implémenter les technologies de signposting, qui consiste en une page web contenant des métadonnées et un lien vers la ressource. Le lien hypertexte n'est en effet reconnaissable que par l'humain, le fait de le "signposter" le rend reconnaissable par les machines.
La technologie du signposting permet notamment de relier la ressource avec toutes les informations qui lui sont attachées, comme les identifiants (DOI).
La COAR préconise également d'adopter la technologie Resource Sync, qui est à rapprocher de la technologie OAI-PMH. Resource Sync permet de moissonner les ressources, et pas seulement les métadonnées, comme le fait OAI-PMH.

La COAR recommande également une solution qui permette de déclarer des licences Creative Commons au niveau de la ressource elle-même, afin que les conditions d'utilisation et/ou de réutilisation de la ressource apparaissent sur la page de la ressource.
La COAR incite aussi à suivre le plan de la science ouverte mis en œuvre par la commission européenne.
Il faut enfin prendre en compte le rôle des pays émergents dans un futur proche. L'Amérique latine s'est inscrite dans le mouvement du libre accès il y a déjà quelques années, et possède aujourd'hui des outils bien élaborés. L'Asie commence aussi à développer des outils et stratégies intéressants.
La nouvelle solution devra donc se rapprocher des standards préconisés par la COAR et utiliser les standards les plus internationaux.

3.2.2 Les recommandations organisationnelles

Les aspects logistique et financier sont aussi importants que les aspects techniques.

Voici les éléments qui nous ont semblé mériter une attention toute particulière lorsqu'il faudra faire le choix d'un nouveau SID.

- Le **coût** des solutions: la question budgétaire doit être systématiquement soulevée. Toutes les solutions, qu'elles soient issues du monde du logiciel libre ou de sociétés éditrices, présentent un coût d'installation et de maintenance conséquent. Plus un SID est personnalisé pour coller aux besoins précis d'un organisme, et plus sa maintenance, et notamment les montées de version, seront onéreuses.

- La **durée des projets** de mise en place ou de refonte d'un système est généralement longue qu'il s'agisse de migrer l'archive institutionnelle sur HAL (projet planifié sur deux ans pour l'INRA par exemple) ou de changer de solution technique, comme c'est le cas à l'INED pour qui le projet a demandé deux années de travail, un an de travail avec le prestataire pour rédiger le cahier des charges (48 pages) et un an de travail en interne.

- **L'organisation des ressources humaines**, en revanche, est très différente d'un organisme à l'autre. Tout dépend de la gouvernance et de la place accordée au service de la documentation, mais aussi de la disposition géographique des services les uns par rapport aux autres.

Nous avons rencontré tous les cas de figure:

- Un informaticien faisant partie de l'équipe des documentalistes et qui peut ainsi créer des développements sur mesure, comme ce fut le cas à l'IRD

- Un informaticien dédié au projet documentaire de l'archive ouverte, et travaillant au sein d'une équipe d'informaticiens dédié à d'autres projet. Cette organisation a permis de mettre en place un écosystème composé de tous ces projets connexes, comme à l'Ifremer.

- Une mutualisation de plusieurs services de documentation de diverses institutions, comme ce sera le cas notamment dans le cadre du futur campus Condorcet.

- Des personnels spécialisés en informatique documentaire, qui permettent de développer des fonctionnalités assez complexes et réalisées sur mesure pour les besoins de leur service, comme à l'université de São Paulo qui dispose également d'un service de huit informaticiens dédiés à la maintenance du catalogue commun aux 48 bibliothèques.

- Les éditeurs de logiciels communiquent la plupart du temps directement avec le service documentation. La maintenance est confiée aux directions des systèmes d'informations.

- Nous avons aussi rencontré le cas de services informatiques et documentation travaillant sur le même plateau, facilitant ainsi la communication des deux services et leur permettant de travailler vers le même objectif. DataPersée est un exemple positif d'une collaboration réussie entre un service informatique qui travaille au quotidien avec le service des documentalistes pour proposer de nouveaux services aux chercheurs.

- Dans tous les cas rencontrés, au moins un ETP était dédié à la réalisation du projet de création, de refonte, ou de migration du SID, opérationnel jusqu'à la mise en route du système.

- Des sessions de médiation sont systématiquement organisées pour former les utilisateurs au nouveau système. Ce volet pédagogique est dans tous les cas pris en charge par les documentalistes.

- Enfin, des actions de revalorisation et de visibilité des services mis à disposition des chercheurs doivent être entrepris au moyen d'un vrai projet de communication. De réels moyens doivent être mis en œuvre.

- Les projets couronnés de succès associent aussi les utilisateurs (en particulier les chercheurs) à leurs réunions de travail.

D'une manière générale, le temps consacré à l'alimentation d'un SID ne doit pas être négligé. Il est donc recommandé de consacrer :

- 1 ETP dédié à la refonte du SDI le temps du projet jusqu'à sa mise en route,
- Des journées de formation qui seront annoncées sur la page d'accueil de l'intranet,
- Solliciter le département communication pour mettre en place un plan de communication autour du projet de refonte, ainsi qu'une campagne de revalorisation et de visibilité du service et des services aux chercheurs.

Dans tous les cas, nous avons pu noter une très forte implication des documentalistes dans le fonctionnement et la compréhension du SID dont ils ont la charge et qui sont souvent à l'origine du projet de refonte ou d'enrichissement. Il en découle des SID avec des développements très personnalisés pour répondre aux demandes des utilisateurs du SID.
Les diverses enquêtes réalisées auprès des utilisateurs rapportent aussi que les utilisateurs apprécient le contact humain avec les documentalistes.

Nous avons toutefois noté l'absence du service juridique. La question n'a pas été formellement posée, il est donc possible que celui-ci intervienne, sans qu'il nous ait été mentionné.
Sa contribution parait indispensable pour traiter de sujets relatifs à la gestion des droits d'auteur pour les publications des chercheurs, comme veiller à la bonne application du RGPD pour les bases de données concernant les utilisateurs, ou les statistiques de consultation, et s'assurer du bon encadrement légal de la propriété des données.

Pour conclure cette sous-partie, nous répétons que cette liste n'est pas exhaustive, les fonctionnalités sont nombreuses, mais pour chacune, il convient d'en questionner finement l'utilité, et d'être averti sur leur avenir : seront-elles encore utiles dans un an, ou deux ?
L'idée d'une solution offrant une large panoplie de fonctionnalités est attirante, mais la gestion d'un tel outil peut s'avérer compliquée.
Nous avons réalisé avec cette étude comparative que le marché offre une panoplie complète et intéressante de solutions logicielles de gestion d'AOI, propriétaires comme libres.
Une solution tout-en-un ne permettra pas de gérer l'AOI à elle seule, et l'ajout de briques logicielles, ou au moins de modules supplémentaires sera nécessaire.
Dans tous les cas, l'aspect technologique ne suffira pas à lui seul, un travail d'équipe humain est indispensable pour la gestion d'une AOI. Les plateformes à elles-seules ne peuvent gérer le système complexe d'une AOI, les efforts sont également assurés par les institutions, notamment à travers le travail des services documentalistes et des informaticiens pour contribuer à la valorisation et à la diffusion des résultats de la recherche.
Les documentalistes notamment sont une source d'information essentielle pour les usagers.

A cet effort collectif sont invités à se joindre les services de communication pour assurer une réelle valorisation du travail des professionnels de la documentation.
Il est donc essentiel qu'une réelle collaboration entre ces services se mette en place pour le choix, mais aussi pour la gestion d'un SID, jusqu'à son évolution.

La mutualisation est une tendance que les institutions ne peuvent ignorer, par mutualisation nous entendons mise en commun de ressources sur des plateformes, mais aussi mise en commun des efforts intellectuels des équipes.
Les séminaires et journées d'étude ne manquent pas pour réunir les documentalistes et personnels IST des différentes institutions, mais nous avons vu qu'il existe aussi une collaboration entre les services informatiques et les développeurs qui travaillent ensemble pour faire évoluer les systèmes exposant la production scientifique mondiale.
Une journée d'étude rassemblant des informaticiens *et* les personnels IST pourrait être constructif, et permettre d'aider à formuler les besoins en termes de SID et d'échanger sur les bonnes pratiques de collaboration.

L'IRD, un des premiers instituts à avoir monté son AOI dans les années 80 en utilisant au mieux les moyens technologiques encore limités de l'époque a certainement un rôle à jouer dans cette entreprise.
Le United Kingdom Council of Research Repositories recommande particulièrement aux instances de l'éducation supérieure britannique, dans son dernier rapport de 2018[61] de faire des relations avec les fournisseurs de services technologiques (éditeurs ou libres) une priorité.
Or les interlocuteurs des fournisseurs sont les personnels IST tout autant que les informaticiens eux-mêmes.
Enfin, les recommandations au niveau international vont toutes dans le sens de l'interconnexion, et pour aboutir à cette cohérence, ceci implique l'adoption de politiques, normes, pratiques et technologies appropriées.
Il faudrait que la grande majorité des entrepôts ou archives ouvertes adoptent les technologies recommandées par la COAR pour que les échanges puissent réellement s'opérer.

[61] Article 6.4.1 du rapport consultable sur le site Universities UK[49]

Conclusion

Cette étude a permis de mettre en perspective les différents aspects de l'IST et leurs objectifs dans le cadre de la recherche d'un nouvel outil de gestion documentaire.
Elle apporte des éléments de réponse, mais nécessite en parallèle une réflexion interne à l'établissement, couplée à des objectifs clairs définis par la gouvernance. L'institution a tout intérêt, avant de s'engager, à définir précisément la politique dans laquelle elle souhaite s'engager afin de porter un projet réalisable et qui pourra s'inscrire sans réserve dans cette époque riche en évolution.
Il en ressort que les mutations que connait l'IST aujourd'hui ne peuvent être ignorées dans le choix d'un nouvel outil, les aspects informationnels et technologiques étant étroitement liés. Chaque innovation que connait l'IST est concrètement reprise par la technologie pour être mise en œuvre concrètement à travers les fonctionnalités des logiciels aujourd'hui disponibles. Ces fonctionnalités sont nombreuses et de nouvelles sont développées en permanence, mais les outils mis à disposition par les éditeurs de logiciels suffiront-ils pour se projeter dans le monde interconnecté et interopérable des bases de données aujourd'hui ?
Les résultats de l'étude que nous avons menée apportent sur ce point des réponses techniques, tout en soulevant certaines questions, notamment organisationnelles. Les deux aspects sont très liés et une collaboration entre le service IST et le service informatique sera indispensable pour mener avec succès le projet de refonte de l'outil de gestion documentaire d'Horizon.
L'étude a en effet montré que les projets les plus aboutis étaient le fruit d'une contribution de ces équipes. L'apport des équipes IST dans le monde de la recherche est considérable, et doit être valorisé.
De plus, dans un esprit de mutualisation cher à l'enseignement supérieur et au monde de la recherche, cette réflexion s'étend aujourd'hui à toute la communauté. La mise en commun des efforts des développeurs et informaticiens des EPST est encouragée par le monde des acteurs de l'IST, mais aussi par les gouvernements à travers le monde.
L'IRD, nous l'avons vu dans la première partie, est l'un des premiers établissements à avoir ouvert sa production scientifique, et sa position de pionnier est évidente et reconnue. La réflexion menée dans le cadre de son projet de refonte, associée à l'époque de changements actuelle est l'opportunité d'assoir cette position et d'avoir une voie au milieu des débats actuels.
Sa mission d'aide aux pays du Sud s'en trouvera renforcée. La diffusion des connaissances scientifiques, et une meilleure accessibilité à ces connaissances pour ces pays font partie non seulement des enjeux de l'IST, mais aussi de l'objectif d'une archive ouverte et du projet en cours.
Enfin, il ne faut pas oublier le chercheur, qui est au cœur de ces réflexions, et dont les habitudes se trouvent bouleversées. Les actions de formation et de médiation menées par le service IST de l'IRD sont nombreuses. La coopération entre l'équipe et les chercheurs est active et riche. De même, les actions de la part de l'ensemble des acteurs de l'IST sont innombrables, et collectives. Les forums et séminaires sont fréquents, permettant une mise en commun et une communication des idées.
A l'heure de ces mutations, la mutualisation est donc incontournable pour apporter des solutions créatives et innovantes.
L'IST bénéficie désormais de nouveaux modes de valorisation, grâce à des moyens technologiques efficaces.
C'est donc un moment opportun pour un institut de recherche qui cherche à faire évoluer son archive ouverte et l'opportunité de prendre le virage stratégique que connait l'histoire de l'IST aujourd'hui.
Les conséquences seront bénéfiques, tant au niveau de ses missions internes de partenariat avec les pays du Sud, qu'à un niveau plus global de circulation des savoirs à travers le monde.

Comme le rappelle le CNRS « *Mieux partager les connaissances, c'est d'abord mieux partager l'information scientifique et technique, être à l'écoute des besoins présents des chercheurs et de la société*[62] ».

[62] http://www.cnrs.fr/dist/z-outils/documents/STRATEGIE.pdf

Bibliographie

La bibliographie qui suit est classée par ordre thématique et est arrêtée au 3 décembre 2018. Les documents de cette bibliographie ne sont pas tous cités dans le texte, mais ont permis d'alimenter une réflexion.

- **Science ouverte :**

[1] CHARTRON Ghislaine. L'Open science au prisme de la Commission européenne. Education et sociétés. 2018/1 (n°41), p.177-193. Disponible en ligne :[consulté le 3 décembre 2018]<https://hal.archives-ouvertes.fr/hal-01888841/document>.

Cet article récapitule les recommandations de la Commission européenne en matière d'Open science et aborde les changements et les risques induits par ce concept, tout en proposant des pistes pour accompagner ces transformations que connait la communauté scientifique.

[2] DIRECTION DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE DU CNRS. « Mieux partager les connaissances ». Décembre 2014. Disponible en ligne :[consulté le 3 décembre 2018]<http://www.cnrs.fr/dist/z-outils/documents/STRATEGIE.pdf>.

Ce rapport présente quatre plans d'actions mis en place par le collège de Direction du CNRS dans le cadre sa stratégie pour mettre en œuvre le plan de la Science publique ouverte.

[3] MINISTERE DE L'ENSEIGNEMENT SUPERIEUR, DE LA RECHERCHE, ET DE L'INNOVATION. Plan national pour la science ouverte. 4 juillet 2018. Disponible en ligne :[consulté le 3 décembre 2018]<http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf>.

Le plan pour la science ouverte présente les quatre que la France s'engage à suivre pour développer sa stratégie.

- **Historique de l'information scientifique et technique :**

[4] CHARTRON Ghislaine. Aux sources de l'Information Scientifique et Technique. 2001. Disponible en ligne :[consulté le 3 décembre 2018]<https://archivesic.ccsd.cnrs.fr/sic_00804152/document>.

Cet article retrace l'histoire de l'IST de 1960 à nos jours, en tire les principales caractéristiques, et met en relief les liens de l'IST avec les questions politiques et industrielles et universitaires.

[5] UNESCO. Rapport de l'UNESCO sur la science, vers 2030 [en ligne]. Paris : Editions UNESCO, 2015 [consulté le 3 décembre 2018]<http://unesdoc.unesco.org/images/0024/002464/246417f.pdf>.

L'UNESCO sort tous les cinq ans un rapport rédigé par des experts internationaux présentant les tendances notées dans le monde de la recherche, de la technologie et de l'innovation.

- **Aspects légaux :**

[6] COMMISSION EUROPEENNE. RECOMMANDATION (UE) 2018/790 DE LA COMMISSION du 25 avril 2018 relative à l'accès aux informations scientifiques et à leur conservation. 25 avril 2018. Disponible en ligne :[consulté le 3 décembre 2018]<https://eur-lex.europa.eu/legal-content/FR/TXT/?qid=1543944073594&uri=CELEX:32018H0790>.

Par cette recommandation, la Commission européenne fait un bilan et met à jour le paquet adopté en 2012 sur l'accès aux informations scientifiques dans le but d'atteindre les objectifs d'Open access alors fixés.

[7] CONSEIL DE L'UNION EUROPEENNE. Conclusions du Conseil sur l'accélération de la circulation des connaissances dans l'UE. 29 mai 2018. Disponible en ligne :[consulté le 3 décembre 2018]<http://data.consilium.europa.eu/doc/document/ST-9507-2018-INIT/fr/pdf>.

Le Conseil dans ces conclusions rappelle les rôles de l'Union et de la Commission dans le transfert de connaissances et leurs objectifs de diffusion de la science dans le but de favoriser la recherche et l'innovation européennes.

[8] DIRECTION DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE DU CNRS. Livre blanc - Une Science ouverte dans une République numérique. 21 mars 2016. Disponible en ligne :[consulté le 3 décembre 2018]<http://www.cnrs.fr/dist/z-outils/documents/2016%2003%2024%20Livre%20blanc%20Open%20Science.pdf>.

Ce livre blanc écrit par la DIST fait partie de la Consultation nationale sur la loi numérique et aborde les thèmes de valorisation et de partage de l'information scientifique à l'heure du numérique, et comment ils s'inscrivent dans les objectifs poursuivis par la Science ouverte.

[9] DIRECTION DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE DU CNRS. Genèse, portée et application de la loi numérique sur l'information scientifique. 2016. Disponible en ligne :[consulté le 3 décembre 2018]<http://www.cnrs.fr/dist/z-outils/documents/loi-numerique/Loi-numerique-genese-portee-application-final.pdf>.

Ce rapport relate les grandes lignes de la loi pour une République numérique et en étudie les applications pour l'information scientifique et technique.

[10] FABRE Renaud, de la direction de l'information scientifique et technique du CNRS. L'application de la loi "pour une République numérique" - Un "guide partagé pour le travail de

la science". 31 mars 2017. Disponible en ligne :[consulté le 3 décembre 2018]<http://www.cnrs.fr/dist/z-outils/documents/analyse-systemique_5points-alerte.pdf>.

Cet article est une analyse de la loi pour une République numérique et présente une synthèse claire particulièrement intéressante pour le domaine du text and data mining. Il tente aussi d'apporter des éléments de définitions des données de la recherche.

[11] OCDE. Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics. 2007. Disponible en ligne :[consulté le 3 décembre 2018]<http://www.oecd.org/fr/science/inno/38500823.pdf>.

Ce rapport énonce les grands principes des données de la recherche, en donne une définition et replace ce sujet dans le contexte actuel.

- **Les missions de l'IRD :**

[12] INSTITUT DE RECHERCHE POUR LE DEVELOPPEMENT – [en ligne] [consulté le 3 décembre 2018] <http://www.ird.fr/l-ird/presentation>.

Site institutionnel de l'IRD sur lequel j'ai pu collecter des données et chiffres sur l'institut.

[13] ROSSI Pier Luigi. Numérisation, bibliothèques électroniques et libre accès : entre renforcement de capacités et perspectives en Afrique francophone. In West and Central African Research and Education Network. WACREN 2018, 12-16 mars 2018, Lomé, Togo. Bondy, 2018. [consulté le 3 décembre 2018] <http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers18-03/010072585.pdf#search=%22numerisation%22>.

L'auteur dans cet article relate les projets de transfert de compétences en numérisation vers les partenaires africains de l'IRD et la mise en place du serveur BEEP et dresse un état des lieux des archives institutionnelles des pays africains.

[14] ROSSI Pier Luigi. Free access to scientific publications for developing countries : the research archive of the French National Institute for Sustainable Development (IRD). 2017. Disponible en ligne :[consulté le 3 décembre 2018]<http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers18-09/010073827.pdf>.

Cette communication de conférence explique le projet de numérisation du fonds papier existant à l'IRD, son socle technique et expose les informations relatives à la consultation des documents, ainsi que la façon dont elles sont exposées.

[15] SCHÖPFEL Joachim. Open Access to Scientific Information in Emerging Countries. D-*Lib* Magazine. March/April 2017, Volume 23, Number 3/4. Disponible en ligne :[consulté le 3 décembre 2018]<http://www.dlib.org/dlib/march17/schopfel/03schopfel.html>.

Cet article aborde les problématiques liées à l'accès aux connaissances des pays émergents, le rôle conséquent de l'Open access, pour le Brésil, la Russie, l'Inde, la Chine et l'Afrique du Sud.

[16] UNESCO. Proclamation du 28 septembre «journée internationale de l'accès universel à l'information». 2015. Disponible en ligne :[consulté le 3 décembre 2018]<http://unesdoc.unesco.org/images/0023/002352/235297f.pdf>.

Par cette proclamation l'UNESCO rappelle les grands principes liés au droit à l'information. Ce document contient des références intéressantes à d'autres déclarations liées aux droits de savoir et de liberté d'expression.

[17] UNESCO. La Plate-forme africaine sur l'accès à l'information adoptée à la conférence du Cap. 23 septembre 2011. Disponible en ligne :[consulté le 3 décembre 2018]<http://www.unesco.org/new/fr/member-states/single-view/news/african_platform_on_access_to_information_adopted_at_confere/>.

Cet article fait une synthèse des évolutions et difficultés liées à l'accès à l'information dans les pays africains.

[18] VICART Benoît. Concevoir un portail documentaire pour les chercheurs et étudiants du Sud [en ligne]. Mémoire, CNAM-INTD, 2015 [consulté le 3 décembre 2018]. 131 p. <http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers16-05/010067103.pdf#search=%22vicart%22>.

Ce mémoire réalisé à l'IRD aborde le sujet des missions de l'institut en Afrique dans le cadre de la conception du portail NumeriSud. Il a été un bon point de départ dans ma recherche sur ce thème.

- **Système d'information documentaire :**

[19] ASCHER Judith, DESBROSSES Arnaud, RIBET Fabrice. Comprendre enfin les systèmes d'information. Paris, La documentation française, 2009. 120 p. ISBN 978-2110072207.

Cet ouvrage m'a apporté des éléments de compréhension pour appréhender les architectures technique et fonctionnelle du système derrière Horizon, en particulier les chapitres sur la structuration de la démarche d'un système d'information, ses composants, son architecture, ses enjeux et ses objectifs (chapitres 3, 4 et 6).

- **Libre accès, Open Access :**

[20] CHARTRON Ghislaine. Stratégie, politique et reformulation de l'open access. Revue Française des Sciences de l'information et de la communication. Août 2016. Disponible en ligne :[consulté le 3 décembre 2018]<https://journals.openedition.org/rfsic/1836>.

Cet article retrace l'histoire de l'Open access, les évolutions politiques du mouvement, les stratégies éditoriales, et les caractéristiques liées au domaine des SHS, replaçant le sujet dans le contexte français à l'heure de la loi pour une République numérique.

[21] LAWRENCE Steve. Free online availability substantially increases a paper's impact. Nature 411, 521, 31 mai 2001. Disponible en ligne :[consulté le 3 décembre 2018]<https://www.nature.com/articles/35079151>.

Par une analyse sur un corpus d'article dans le domaine informatique, cet article s'attache à démontrer qu'un article disponible en ligne est plus susceptible d'être cité.

[22] SALAUN Jean-Michel. Libre accès aux ressources scientifiques et place des bibliothèques. Bulletin des bibliothèques de France (BBF). Novembre 2004, n°4. Disponible en ligne :[consulté le 3 décembre 2018]<http://bbf.enssib.fr/consulter/bbf-2004-06-0020-003>.

Cet article traite des orientations du mouvement du libre accès et des différents rôles de ses acteurs et comment leurs missions s'articulent dans le monde scientifique à l'heure de la révolution numérique.

[23] TENNANT Jonathan P.,Waldner François, COLLISTER Lauren B. , HARTGERINK Chris. H.J. The academic, economic and societal impacts of Open Access: an evidence-based review. 21 septembre 2016. Disponible en ligne :[consulté le 3 décembre 2018]<https://f1000research.com/articles/5-632/v3>.

Cet article replace l'open access dans le contexte actuel de l'open Science, en analysant les impacts économiques, sociétaux et académiques.

[24] CHARTRON Ghislaine, SCHÖPFEL Joachim. Open access et Open science en débat. Revue Française des Sciences de l'information et de la communication. 2017, n°11. Disponible en ligne :[consulté le 3 décembre 2018]<https://journals.openedition.org/rfsic/3331>.

Cet article aborde les aspects historiques, économiques, les pratiques de dépôt des chercheurs et la nécessité d'accompagner les chercheurs dans ces pratiques.

- **Edition scientifique :**

[25] CHARTRON Ghislaine. Open access et SHS : Controverses. Revue européenne des sciences sociales. 2014, n° 52-1. Disponible en ligne :[consulté le 3 décembre 2018]<https://journals.openedition.org/ress/2658>.

Cet article aborde les particularités liées aux publications en SHS dans le contexte de l'Open access, les difficultés, les risques et contradictions que peut engendrer ce mouvement.

[26] CHARTRON Ghislaine, Scénarios prospectifs pour l'édition scientifique Ghislaine Chartron. Hermès, La Revue. 2010/2 (n°57), p. 123-129. Disponible en ligne :[consulté le 3 décembre 2018]<https://archivesic.ccsd.cnrs.fr/sic_00558746/document>.

Cet article traite de l'évolution de l'édition scientifique dans le contexte du développement numérique et du libre accès. Il apporte une vision différente de ce domaine et des pistes de réflexions pour accompagner la transformation du marché de l'édition.

[27] CHARTRON Ghislaine. Evolution de l'édition scientifique, 15 ans après Ghislaine Chartron. in groupe européen et interdisciplinaire sur les enjeux et usages des tic. EUTIC 2007, 8-10 novembre 2007, Athènes, Grèce. France, 2007. [consulté le 3 décembre 2018]<https://archivesic.ccsd.cnrs.fr/sic_00186675/document>.

Cette communication fait un compte rendu des évolutions des 15 dernières années dans le monde de l'édition scientifique et à partir de ce constat présente des scénarios possibles pour le futur.

[28] CHARTRON Ghislaine. Une économie renouvelée de la publication scientifique. Perspectives documentaires en éducation. 2006, n° 062/2006. Disponible en ligne :[consulté le 3 décembre 2018]<https://archivesic.ccsd.cnrs.fr/sic_00117798/document>.

Cet article retrace les développements de l'édition scientifique des dix dernières années, liés aux nouveaux modes de communication scientifique, et les incidences sur les comportements des utilisateurs finaux de la masse d'informations produites dans ce contexte.

[29] ELSE Holly. Europe's open-access drive escalates as university stand-offs spread. Nature. 2018, n°557, p. 479-480. . Disponible en ligne :[consulté le 3 décembre 2018]<https://www.nature.com/articles/d41586-018-05191-0>.

Cet article évoque les nouveaux types de contrats signés entre éditeurs scientifiques et bibliothèques, des stratégies de ces dernières pour mener les négociations en cours actuellement en Europe.

[30] VANHOLSBEECK Marc. La notion de Science Ouverte dans l'Espace européen de la recherche. Revue Française des Sciences de l'information et de la communication. 2017, n°11. Disponible en ligne :[consulté le 3 décembre 2018]<https://journals.openedition.org/rfsic/3241>.

Cet article met en perspectives les contradictions que fait ressortir l'open Science qui d'un côté encourage la diversité des canaux de diffusion de la production scientifique tout en incitant les institutions à se fonder sur les indicateurs des revues internationales les plus réputées, et les pistes possibles qu'apportent les modèles de publication en Open Access.

[31] VERLAET Lise. Modèles d'affaire de l'Open access. Réflexions autour du projet NumeRev. Les cahiers de la SFIC. Février 2017, n°13, p.39-50.

A partir d'un constat sur les impacts de la révolution numérique sur l'édition scientifique, cet article recense les différents modèles de publications scientifiques, questionne sur la pertinence de chacun, leurs incidences sur les pratiques de dépôt des chercheurs et propose un nouveau modèle d'affaire pour le projet NumeRev.

- **La mutualisation**

[32] BAUDE Catherine, BARDET Florence, MARGUERIN Stéphane. Mutualisations. État des lieux et enseignements. i2D. 2015, n°3, volume 52, p.28-29.

Cet article passe en revue des actions de mutualisation en cours dans le domaine de la documentation et questionne sur l'avenir de ces nouvelles pratiques.

[33] GIRARD Chloé. Les mécanismes de centralisation des données de la recherche. Revue Française des Sciences de l'information et de la communication. 2017, n°11. Disponible en ligne :[consulté le 3 décembre 2018]<https://journals.openedition.org/rfsic/3255>.

Cet article évoque la question de la centralisation de l'hébergement des ressources, les limites de cette solution, et apporte des pistes possibles pour un système basé sur la distribution, en adéquation avec les objectifs de la science ouverte.

[34] TEXIER Bruno. Bibliothèques : quand les SIGB passent à la mutualisation. Archimag. Juillet 2018.

Cet article expose les raisons qui conduisent les bibliothèques à choisir un système mutualisé, et prend en exemple le cas du SGBM de l'ABES.

[35] VALLUY Jérôme. Libre accès aux savoirs et accès ouvert aux publications. Revue Française des Sciences de l'information et de la communication. 2017, n°11. Disponible en ligne :[consulté le 3 décembre 2018]<https://journals.openedition.org/rfsic/3194>.

Cet article aborde les questions liées à la centralisation des ressources dans le domaine des SHS et des pratiques managériales qui en découlent.

[36] Les mutations du marché des SGB et les questions qu'elles soulèvent. Ar(abes)ques. Juin 2018, n°89, p.6-7.

Cet article s'adresse aux professionnels des bibliothèques qui sont dans une démarche de ré informatiser leur système et revient sur les mutations du marché des progiciels du secteur des dernières décennies et ses impacts sur le marché actuel.

- **Bibliométrie**

[37] LARIVIERE Vincent, SUGIMOTO Cassidy R. Mesurer la science [en ligne]. Québec : Les Presses de l'Université de Montréal, 2018 [consulté le 3 décembre 2018]<https://www.pum.umontreal.ca/fichiers/livres_fichiers/9782760639522.pdf>.

Cet ouvrage fait une synthèse complète de l'histoire de la bibliométrie, de ses indicateurs et des enjeux liés à la mesure de la science.

- **Les nouvelles interopérabilités :**

[38] CONFEDERATION OF OPEN ACCESS REPOSITORIES. Que faire pour une meilleure interopérabilité. 2018. Disponible en ligne :[consulté le 3 décembre 2018]<https://visiarchives.sciencesconf.org/data/pages/coar_ngr_couperin_final_2018.pdf>.

Ce rapport dresse un bilan de la communication académique et présente sa vision consistant à développer un réseau global et interopérable pour lequel il présente ses recommandations techniques.

[39] CONFEDERATION OF OPEN ACCESS REPOSITORIES. Recommandations techniques d'interopérabilité de la COAR. 2018. Disponible en ligne :[consulté le 3 décembre 2018]<https://github.com/coar-repositories/ngr/tree/master/webroot/content/behaviour>.

Cette page est une liste des recommandations techniques de la COAR, leur signification et les moyens technologiques pour les mettre en œuvre.

[40] CONFEDERATION OF OPEN ACCESS REPOSITORIES. Les archives ouvertes de prochaine génération. 28 novembre 2017. Disponible en ligne :[consulté le 3 décembre 2018]<https://www.coar-repositories.org/files/NGR-Final-Formatted-Report_french-version.pdf>.

Ce rapport présente les nouvelles fonctionnalités, les technologies, les normes et protocoles qui permettront la construction d'un vaste réseau de ressources scientifique sur le principe de la distribution.

[41] CONFEDERATION OF OPEN ACCESS REPOSITORIES. The Current State of Open Access Repository Interoperability. 2012. Disponible en ligne :[consulté le 3 décembre 2018]<https://www.coar-repositories.org/files/COAR-The-Current-State-of-Open-Access-Repository-Interoperability.pdf>.

Ce rapport dresse la liste des initiatives lancées et des technologies associées et leurs applications possibles, dans le but de développer un système interopérable d'archives ouvertes à travers le monde.

[42] INSTITUT DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE. Rapport d'activité 2017. 2017. Disponible en ligne :[consulté le 3 décembre 2018]<https://www.inist.fr/wp-content/uploads/2018/08/inist_rapport_d_activite_2017.pdf>.

Ce rapport de l'INIST évoque dans son deuxième axe « valorisation des données de la recherche » des projets mis en place qui illustrent la notion d'interopérabilité.

[43] MELHEM Hiba. Usages et applications du web sémantique en bibliothèques numériques [en ligne]. Thèse de doctorat, Sciences de l'Information et de la Communication, laboratoire GRESEC (Groupe de recherche sur les enjeux de la communication), 2016 [consulté le 3 décembre 2018]. 220p. <https://tel.archives-ouvertes.fr/tel-01742957>.

Cette thèse explique de façon claire et compréhensive les différentes couches composant le web sémantique, ses applications et ses enjeux.

- **Logiciels libres :**

[44] AWRE Christopher, GREEN Richard. From Hydra to Samvera: an open source community journey. Insights. 2017, 30(3), p. 82-88. Disponible en ligne :[consulté le 3 décembre 2018]<https://insights.uksg.org/articles/10.1629/uksg.383/>.

Cet article présente une histoire récente du monde du logiciel libre au Royaume-Uni, et les raisons de son essor.

[45] CORRADO Edward M. The Importance of Open Access, Open Source, and Open Standards for Libraries. Science and Technology Librarianship. 2005. Disponible en ligne :[consulté le 3 décembre 2018]<http://www.istl.org/05-spring/article2.html>.

Cet article permet de comprendre les avantages que présentent les logiciels libres pour les bibliothèques, et l'importance que ce constat revêt.

[46] DURASPACE. The DSpace Community Annual Report 2017. Juillet 2018. Disponible en ligne :[consulté le 3 décembre 2018]<https://duraspace.org/wp-content/uploads/2018/09/DSpace-Annual-Report-F.pdf>.

Ce rapport permet d'une part d'illustrer l'importance de la communauté d'utilisateurs et leurs profils, d'autre part de présenter les prochaines évolutions techniques que promet le logiciel.

[47] TORRE Dominique. Le modèle économique du logiciel Open Source : viabilité et compétitivité. Revue d'économie industrielle, 2011, n°136, p. 11-16. Disponible en ligne :[consulté le 3 décembre 2018]<https://journals.openedition.org/rei/5163>.

Cet article décrit le modèle économique sur lequel reposent les logiciels Open Source, et interroge sur la pérennité de ce système.

[48] UNIVERSITIES UK OPEN ACCESS COORDINATION GROUP. Open access repositories: report and recommendations. Juillet 2018. Disponible en ligne :[consulté le 3 décembre 2018]<https://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2018/open-access-repositories-report-and-recommendations.pdf#search=open%20source>.

Ce rapport établit une liste des recommandations à suivre pour la gestion d'une archive ouverte au Royaume-Uni, un constat des pratiques, les tendances et les débats qui animent cette communauté. Il permet de recouper les informations obtenues dans les rapports de la COAR et d'en observer les applications qui en sont faites.

- **Exemple de systèmes d'information documentaires :**

[49] MERCEUR Frédéric. Archimer, ou la mise en place d'une Archive Institutionnelle à l'Ifremer. 24 novembre 2005. Disponible en ligne :[consulté le 3 décembre 2018]<https://archimer.ifremer.fr/doc/2005/rapport-657.pdf>.

Cet article écrit par le maître d'œuvre d'Archimer, l'archive ouverte d'Ifremer, documente de façon très claire le système pourtant complexe mis en place et permet d'en comprendre l'architecture technique et la façon dont il est connecté avec d'autres bases de données.

- **Etudes comparatives**

[50] CASTAGNÉ Michel. Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra. 14 août 2013. Disponible en ligne :[consulté le 3 décembre 2018]<https://open.library.ubc.ca/cIRcle/collections/graduateresearch/42591/items/1.0075768>.

Cette étude compare différents logiciels de gestion d'archives ouvertes institutionnelles, et liste les principaux critères, m'aidant ainsi à comprendre quels aspects retenir pour l'étude comparative menée dans le cadre de mon stage.

[51] MAISONNEUVE Marc. L'enquête annuelle 2018, analyse des chiffres. Ar(abes)ques. Juin 2018, n°89, p.9.

Cet article est une synthèse de l'étude annuelle menée par le Tosca Consultants sur le marché français des logiciels métiers destinés aux bibliothèques.

[52] NOORMAN MASREK Mohamad, HAKIMJAVADI Hesamedin. Evaluation of Three Open Source Software in Terms of Managing Repositories of Electronic Theses and Dissertations: A Comparison Study. Journal of Basic and Applied Scientific Research. 2012. Disponible en ligne :[consulté le 3 décembre 2018]<https://pdfs.semanticscholar.org/4f20/fa7f78e82ed10b57ec0eb6e54dd29e23d0f1.pdf>

Cette étude, comme la précédente, est intéressante dans la recherche d'indicateurs pour le cas de l'IRD.

[53] SINGH Siddharth Kumar, WITT Michael, SALO Dorothea. A Comparative Analysis of Institutional Repository Software. Janvier 2010. Disponible en ligne :[consulté le 3 décembre 2018]<http://biecoll.ub.uni-bielefeld.de/volltexte/2011/5076/pdf/abs_singh_comparative.pdf>.

Cette étude, encore une fois, est intéressante pour les nombreux critères d'évaluation qu'elle présente et dont j'ai pu m'inspirer pour réaliser le comparatif.

Annexes

Annexe 1 : Description du système Horizon par Dominique Cavet

Organisation des serveurs et architecture du système d'information Horizon

Le système d'information Horizon est composé de trois bases principales :

- **la base Horizon Cadic**
- **la base Horizon SQL**
- **les PDFs**

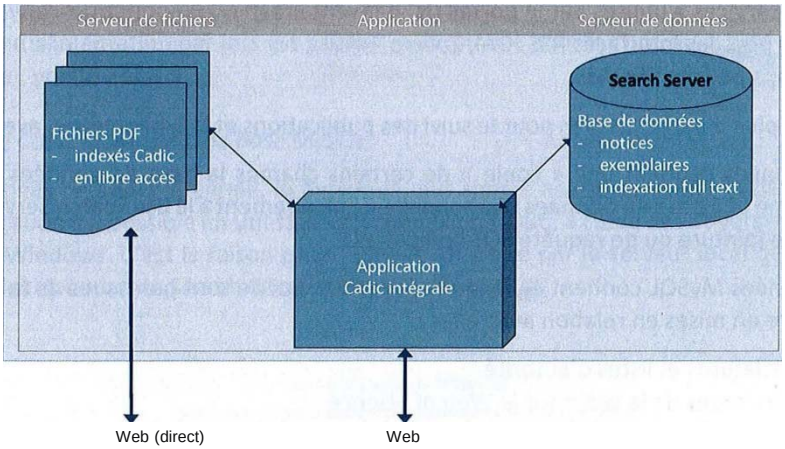
La base Horizon Cadic est le fondement de tout le système d'information, et est alimentée et gérée via le logiciel *Cadic Intégrale* (version 2011), système intégré de gestion de bibliothèques (SIGB) de la société Cadic.

C'est la base Horizon Cadic qui est exportée quotidiennement dans la base Horizon MySQL qui permet ensuite l'exploitation d'Horizon via des services en lignes.

Le document ci-dessous décrit dans un premier temps les serveurs et les applications qu'ils hébergent et dans un deuxième temps les flux d'alimentation des différentes briques.

Les serveurs et leur contenu

1 Serveur *Cadic Intégrale*



L'application Cadic Intégrale est un SIGB, elle est spécifique et fonctionne en PHP[1], avec des interfaces web (portail utilisateur + portail de gestion authentifié). Elle est dans le domaine http://horizon.documentation.ird.fr.

[1] Il s'agit d'une version adaptée par Cadic du php standard.

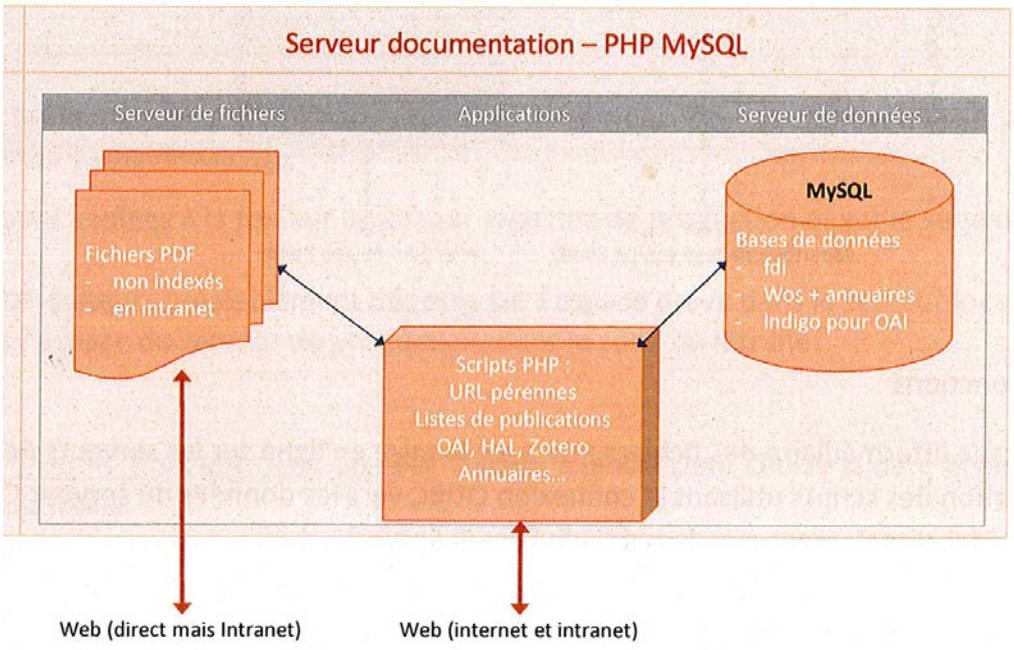
Elle est hébergée sur le serveur *censvrap0006* à l'adresse IP 10.7.1.56[2]. La DDUNI est chargée de l'exploitation de ce serveur selon une documentation qui a été remise à l'issue de la précédente montée de version (voir document en annexe).

Les données sont stockées dans une base de données **Search Server (SearchServer_3.7)** gérée par l'application mais également directement accessible avec une interface ODBC (RMQ : l'alimentation de la base MySQL se fait via l'ODBC).

Pour les fichiers PDF, on utilise partiellement les fonctionnalités GED de Cadic Intégrale :

- L'indexation *full text* des fichiers PDF est gérée par l'application (au travers de fichiers .txt générés par l'utilitaire pdftotext, voir plus bas)
- L'accès web aux fichiers pdf se fait directement via le serveur Apache du serveur, dans le domaine http://horizon.documentation.ird.fr mais en dehors de l'application Cadic, les URL des fichiers PDF étant stockées dans les notices de la base.

2 Serveur web de la doc



Ce serveur est en fait composé de deux serveurs (virtuels) : un pour le Web et les scripts PHP et l'autre pour la base MySQL.

Le serveur a deux fonctions :

1 permettre des extensions aux fonctionnalités de l'application Cadic Intégrale (flux Rss, serveur OAI…)
2 gérer d'autres données nécessaires aux services en ligne de l'IST : bibliométrie, suivi des publications…

[Rmq : les pdfs en intranet ne sont pas du tout indexés, ni par Cadic Intégrale, ni par Google]

[2] Sur ce serveur sont hébergés quatre environnements de l'application : production (port 8080), test (8081), qualification et formation – voir document d'exploitation en annexe.

Pour Horizon MySQL, il n'y a pas d'application à proprement parler, mais une série de scripts accessibles sous formes d'URLs dans le domaine www.documentation.ird.fr:

- URLs reprises par le portail public Cadic Intégrale (tableaux, etc.)
- URLs diffusées aux chercheurs (listes dynamiques de publications, etc.)
- « web services » utilisés par le portail (n° d'inventaires)
- Scripts pour les interfaces RSS, OAI, Zotero, HAL…
- Sitemaps pour le site Web

Le serveur a de plus d'autres usages pour le suivi des publications et la bibliométrie, avec des scripts en intranet.

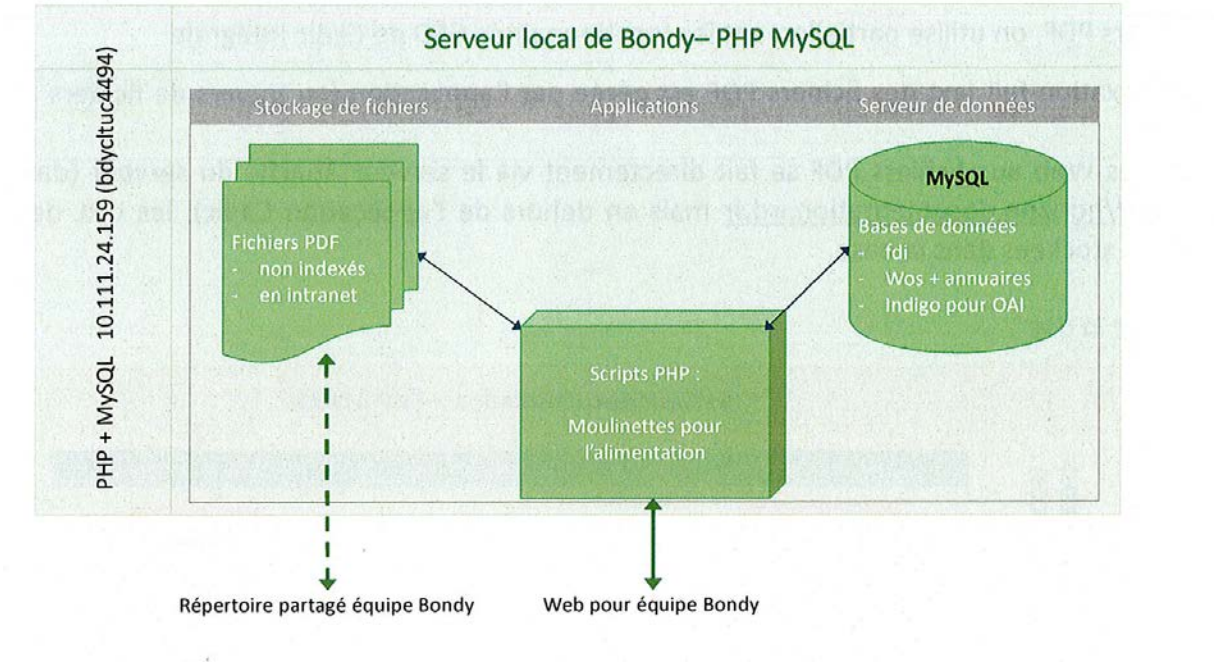
La base de données contient une « copie » de certains champs de la base de données *Cadic intégrale*. Elle est organisée comme une base de données relationnelle, contrairement à la base Search Server qui est plutôt orientée Full Text (pas de jointure ou de requête SQL complexe).

La base de données MySQL contient également des tables qui ne sont pas issues de la base Horizon Cadic mais qui peuvent être mise en relation avec elles :

- Nomenclatures et listes d'autorité
- Données issues de la veille sur le Web of Science
- Données sur les chercheurs et sur les unités

Comme il n'était pas possible de mettre certains fichiers PDF en accès réservé sur le serveur Cadic Intégrale, un espace en intranet est présent dans le serveur Web de la doc. Il contient les fichiers PDF venant des sites des éditeurs, accessibles seulement en intranet pour les personnels IRD. L'URL de ces fichiers, dans le domaine www.documentation.ird.fr, est stockée dans un champ de la base Cadic.

3 Serveur local[3] à Bondy



Il a plusieurs fonctions :

[3] Environnement WAMP (Php 5.2.11 ; MySQL 5.1.36 ; Apache 2.2.11 ; Windows 8)

- Stockage immédiat des fichiers PDF avec la mise en ligne sur les serveurs de production
- Utilisation des scripts utilisant la connexion ODBC vers les données du serveur Cadic (seul moyen d'accéder directement aux données de Search Server)
- Archivage des fichiers PDF

Flux pour l'alimentation

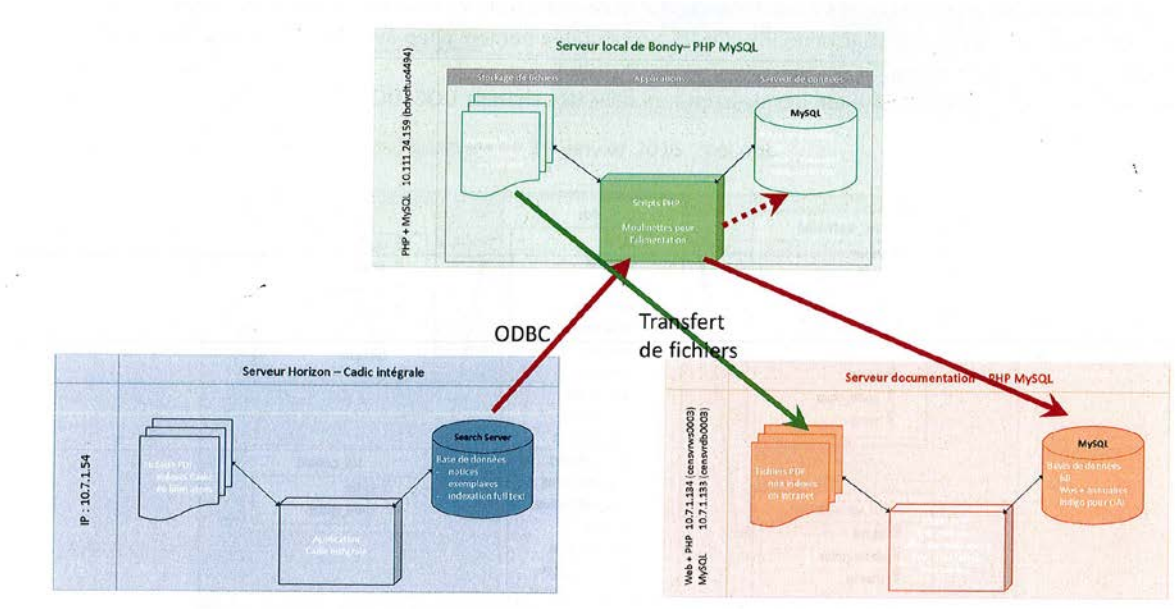
1 Alimentation courante d'Horizon avec *Cadic Intégrale*

Elle se fait par l'interface web de l'application Cadic Intégrale, en mode authentifié. Cependant :

- on n'utilise pas (pour l'instant) l'insertion automatique des fichiers PDF dans l'interface Cadic, ce qui nécessite d'utiliser des scripts décrits ci-dessous dans le point 4
- la mise à jour les données du serveur web de la doc (base MySQL) se fait au moyen de la procédure de synchronisation du point 2
- une procédure d'alimentation par lots est réalisée, notamment pour les notices issues de la veille sur le Web of Science, voir le point 3

2 Synchronisation entre *Cadic Intégrale* et la base MySQL

Pour récupérer les données de Cadic Intégrale, le moyen le plus simple est de consulter la base Search Serveur directement. Cette opération est possible en utilisant une connexion ODBC, à l'aide des drivers installés par Cadic sur les « clients » sous Windows. C'est la raison pour laquelle on passe par le serveur local de Bondy pour ces opérations de recopie des données de Search Server dans la base MySQL.



La synchronisation se fait donc par des scripts tournant sur le serveur local de Bondy. Elle se fait uniquement dans le sens Search Server Cadic > MySQL.

- les données sont copiées à la fois sur le serveur web doc de production et sur le serveur local (base de travail)
- les fichiers PDF éditeurs préalablement déposés sur l'espace prévu dans le serveur local de Bondy sont recopiés dans l'espace du serveur de production, dans la zone en intranet.

Scripts mis en œuvre :

- http://bdycltuc4494/outils_horizon/menu.php est le script à lancer pour choisir la date de lancer une mise à jour incrémentale ou totale,
- http://bdycltuc4494/outils_horizon/maj-horizon.php fait une mise à jour incrémentale, pour les données modifiées ou ajoutées depuis la date indiquée dans le paramètre « depuis »,
- http://bdycltuc4494/outils_horizon/raz-horizon.php régénère complètement les tables MySQL utilisées pour Horizon.

6 tables **fdi** de la base MySQL sont concernées par cette synchronisation (fdi_auteur, fdi_centre, fdi_chap, etc....), et contiennent les principaux champs de la table **ILS_DOC** de Cadic Intégrale ; quelques champs de gestion sont également ajoutés lors de cette opération (id_fdi, clé primaire qui reprend l'identifiant Cadic, code_type qui génère le type de document, etc...).

3 Alimentation par lots

Elle est utilisée tous les mois pour insérer dans Cadic Intégrale les notices issues de la veille effectuée sur le Web of Science.
La source de données est le fichier End Note au format XML produit tous les mois par E Ambert.
Les fichiers PDF éditeurs en accès réservé doivent auparavant être copiés en FTP sur l'espace adapté du serveur web de la doc. Les fichiers PDF éditeurs en libre accès doivent être auparavant copiés sur l'espace dédié dans le serveur partagé X: avec les autres PDFs en libre accès, avant d'être traités par L. Rossi puis intégrés dans les notices lors de l'opération quotidienne décrite en point 2 (synchronisation entre Cadic Intégrale et la base MySQL).
Un module de Cadic Intégrale, fonctionnant en fullweb (module ExImport) permet d'importer dans Cadic intégrale chaque notice du fichier End Note xml.
A l'issue de la procédure d'insertion dans Cadic Intégrale, il faut lancer la procédure de synchronisation décrite au point 2.

4 Traitement des fichiers PDF

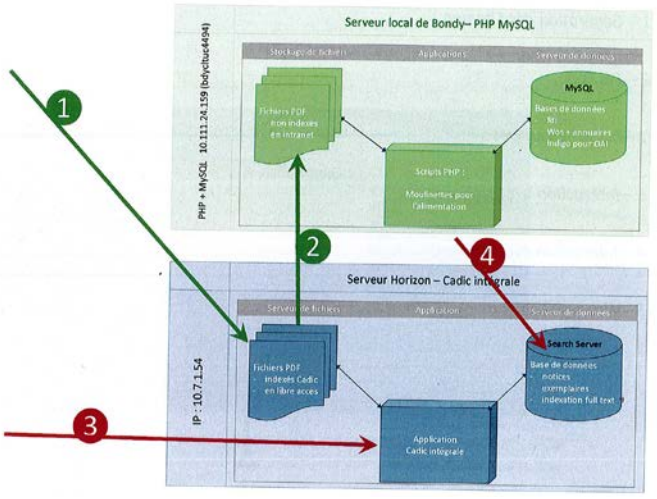
1 – copie des fichiers PDF préparés, au fur et à mesure de leur production

Puis, **tous les mois** :

2 – copie du dossier mensuel sur le serveur local de Bondy

3 – Indexation des fichiers PDF pour Cadic, sur le serveur Cadic

4 – Insertion des données relatives à la GED dans la base Search Server



Autres utilisation des serveurs

1. Génération des sitemaps
2. Flus RSS
3. Exposition en OAI
4. Interactions avec HAL
5. Interaction avec Zotero
6. Webservice pour l'interface Cadic (numéros d'inventaire)

Annexe 2 : Liste des personnes interviewées

- Les instituts ayant choisi la solution HAL :

  - **Entretien n° 1** : Institut Pasteur - https://hal-pasteur.archives-ouvertes.fr/
    Personne interrogée : responsable du Centre de Ressources en Information Scientifique (CeRIS)
    Date : 31 août 2018

  - **Entretien n° 2** : INRA - https://Prodinra.inra.fr
    Personne interrogée : responsable de l'archive ouverte Prodinra de l'INRA
    Date : 14 août 2018

  - **Entretien n° 3** : IRD - hal.ird.fr/
    Personne interrogée : responsable des abonnements et de la coordination du réseau documentaire et de la production de la base Horizon Pleins Textes
    Date : 15 septembre 2018

- Solutions issues du libre :

  - **Entretien n° 4** : INED - archined.ined.fr
    Personne interrogée :cheffe de projet Archive ouverte
    Date : 17 septembre 2018

  - **Entretien n° 5** : INSERM - http://www.ipubli.inserm.fr/
    Personne interrogée :ingénieure d'études | Chargée de ressources documentaires
    Date : 1er août 2018

- Solutions à plusieurs composantes

  - **Entretien n° 6** : Université de São Paulo - http://bdpi.usp.br/
    Personne interrogée :Chef de la Division de la gestion de l'information du Système intégré de bibliothèque (SIBi) de l'Université de São Paulo
    Date : échanges de mail août 2018

  - **Entretien n° 7** : Ifremer - https://archimer.ifremer.fr/
    Personnes interrogées :
    - Directrice de la bibliothèque La Pérouse
    - Administratrice d'Archimer - Bibliothèque La Pérouse
    - Maître d'œuvre d'Archimer - Service Ingénierie des systèmes d'information de l'Ifremer

    Date :10 septembre 2018

- Les éditeurs :

  - **Entretien n° 8** : PMB - https://www.sigb.net/
    Date : 3 septembre 2018
    Personne interrogée : Responsable commerciale

- **Entretien n° 9** : PolarisOS - https://www.mysciencework.com/polaris-os
  Personne interrogée : Directeur de l'innovation et du développement
  Date :7 août 2018

- Triplestore :
  - **Entretien n° 10** : DataPersée : http://data.persee.fr/
    Personne interrogée : Responsable pôle informatique

    Date : 31 août 2018

Annexe 3 : Entretiens

Entretien n°1: Institut Pasteur - https://hal-pasteur.archives-ouvertes.fr/

Entretien réalisé le 31 août 2018 avec la responsable du Centre de Ressources en Information Scientifique (CeRIS).

Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger.

- **Présentation**

Le portail HAL-Pasteur a été créé suite à la présentation de l'INSERM de son nouveau portail HAL-INSERM en 2006. Ce portail a été dupliqué et sa mise en place relativement simple.

- **Les ressources gérées**

Le portail Hal-pasteur compte aujourd'hui 3000 notices avec texte intégral, et plus de 2600 références bibliographiques
On y trouve essentiellement des articles de chercheurs, des thèses (initialement déposées sur le portail HAL de tel par L'ABES), des chapitres d'ouvrages, des actes de congres, et quelques prépublications.

- **Qu'est ce qui a motivé votre choix ?**

Tout est géré par CCSD, et ceci est un argument en faveur de la décision de confier son archive a HAL. Le paramétrage et la configuration de la page d'accueil sont gérés par l'administrateur du portail de l'institut au sein du CeRIS - bibliothèque. Le paramétrage est relativement simple à gérer.
Il est également possible de paramétrer des portails de façon différente, ce qui donne des portails HAL avec des spécificités propres à chacun.

- **D'où partiez-vous ?**

Une base de publications a été créée en 2002, qui, bien qu'assez exhaustive, ne comprenait que des notices (pas de texte intégral). Notices importées du WOS et de Pubmed (de HAL par la suite). Cette base a été arrêtée en 2016.
L'objectif, en ouvrant un portail HAL, n'était pas de construire une base de publications, mais de créer une plateforme d'archives ouverte pour répondre à la demande de la politique d'Open Access, l'Institut Pasteur ayant signé la Déclaration de Berlin en 2004.

- **Versez-vous vos publications sur une autre plateforme ? Pubmed? Repec?**

Un lien vers HAL est créé dans Pubmed sous forme d'une icône HAL, quand il y a un fichier auteur dans HAL (il ne s'agit pas d'un versement).

- **Quelle est la politique de versement dans HAL ?**

La politique de dépôt de l'institut Pasteur est incitative, mais pas coercitive, donc le portail HAL n'est pas exhaustif.
L'institut a mis en place des mesures incitatives en septembre 2016. Ainsi, les chercheurs qui demandent des promotions doivent avoir déposé toutes leurs publications depuis 2010 avec le texte intégral. C'est une mesure qui marche mais qui concerne 10 ou 15 chercheurs par an.

- **Quelles sont les facilites de paramétrage et d'évolution**

Tout est géré par le CCSD, et ceci est un autre argument en faveur de la décision de confier son archive à HAL. Le paramétrage et la configuration de la page d'accueil sont gérés par

l'administrateur du portail de l'institut au sein du CeRIS (Centre de ressource en information scientifique). Le paramétrage est relativement simple à gérer.

- **Workflow**

les workflows ne sont pas possibles sur HAL. Il y a un historique, qui permet de savoir qui a travaillé sur le dépôt et à quel moment, et qui permet de savoir qui a fait le dépôt, la date de modération (avec certaines nuances car on ne sait pas qui a modèré). Le tamponnage aussi est visible grâce à l'historique.

- **Alimentation**

La récupération des métadonnées est automatique à partir du PDF, à condition qu'il soit bien structuré, et à partir des articles référencés grâce au numéro PubMed.

- **Import de références aux formats bibliographiques standards (bib tex, ris, ...) Quels formats de md supporte Hal?**

Il est difficile de réaliser de gros imports mais il est possible d'importer des formats différents grâce à des outils développés par certaines universités ou instituts de recherche.
Ainsi l'outil Bib2HAL développé par l'INRIA, permet de déposer un lot de publications dans HAL à partir d'un fichier BibTeX. Les exports de notices BibTex sont facilement exploitables par Zotero et les fichiers en PDF.
L'application OCdHAL (Outil de Contrôle des données de HAL) développée par Grenoble, permet de consulter, modifier et synchroniser l'ensemble des articles d'une collection HAL, présentés sous la forme d'un tableau dans lequel les articles s'affichent en ligne et leurs métadonnées en colonnes.
Il est prévu que le CCSD intégre ces outils dans HAL. Ils sont d'ores et déjà mis à contribution de la communauté, via le wiki de HAL

- **La gestion des droits d'auteur**

HAL permet de bien gérer les droits d'auteur. L'administrateur peut indiquer la date d'embargo de son choix et le fichier s'ouvrira automatiquement à la date indiquée. De plus, il est possible d attribuer une licence Creative Commons à l'article.
Un article sur lequel un éditeur dispose de droits ne sera visible que par le déposant etles administrateurs. Toutefois tout le monde pourra en faire la demande via le bouton "Request-a-print".
La politique est donc la même que ce soit pour un chercheur de l'Institut ou toute personne extérieure. Il n'y a pas de site intranet dans HAL. Le bouton « request-a-print » est accessible à tout le monde à condition d'avoir un compte dans HAL et d'être connecté. (ce qui est possible pour tout le monde).
La visibilité des fichiers sous embargo infinis (fichiers qui n'ont pas vocation à être diffusés en dehors de l'institution) par les chercheurs de l'institut a fait l'objet de plusieurs demandes d'établissement.
Pour le moment, quand un article est en open Access et sous licence CC, l'institut rajoute le fichier éditeur et dans le cas échéant contraire (article sous copyright), c'est le fichier auteur qui sera dépose. Un travail de médiation/formation est nécessaire a ce point car souvent les chercheurs comprennent mal la différence, et même s'ils en font la demande, l'éditeur ne donne pas forcement le bon fichier.
La communauté académique considère les postprint comme des fichiers éditeurs (après peer-reviewing et avant publication (c'est-à-dire sans la mise en page de l'éditeur) et les preprint comme fichiers de travail (première version du fichier envoyé à l'éditeur soit avant peer-reviewing).
La loi Lemaire a par ailleurs donné le droit aux auteurs de déposer leurs fichiers auteurs en autoarchivage à condition qu'ils aient été financés a moitié par des fonds publics.
A l'institut Pasteur, les financements sont répartis entre le public et le secteur privé de façon très complexe, il est donc en général considéré qu'au moins la moitié du financement est public.

- **Est-il possible de faire un dépôt par tous les auteurs mentionnés?**

Les affiliations générées automatiquement par HAL contiennent des erreurs, il faut donc revérifier toutes les affiliations si on veut avoir des dépôts propres. même si depuis la version simplifiée, on n'a pas obligation d'affilier tous les auteurs
Des progrès devraient être faits dans la gestion des affiliations automatiques, pour l'instant, le travail humain reste indispensable.
Il faut notamment prendre en compte la date de publication de l'article et non la dernière affiliation réalisée dans Hal.

- **Gestion du dédoublonnage**

Le dédoublonnage se fait au niveau du dépôt et de la modération, HAL alerte respectivement le déposant et le modérateur.
Il existe un module de dédoublonnage pour gérer manuelle (pour l'administrateur de portail uniquement).

- **Interopérabilité**

Les données sont exposées dans des formats et protocoles standards (OAI-PMH, RDF, ...). HAL a développé un entrepôt RDF il y a un an.
En revanche, il est difficile de savoir par qui le portail est moissonné (à part Openaire) or cette information est utile à savoir.

- **Connecteurs IST**

Les connecteurs IST sont opérationnels et marchent correctement.

- **Les référentiels**

Le CCSD a développé un référentiel adossé à HAL, qui comprend les auteurs, structures, revues, domaines scientifiques, projets européens et ANR. AUREHAL est en lien avec d'autres référentiels IST par l'intermédiaire des identifiants, comme le RNSR, IdREF, ISNI. En revanche il n'est pas en lien avec ScanR.
Les institutions exposant leurs publications sur un portaitl HAL peuvent utiliser ce rérérentiel sans avoir recours à leurs référentiels maison.

- **Les affiliations**

Les affiliations générées automatiquement dans HAL contiennent des erreurs et le travail humain reste indispensable.
A noter toutefois, depuis la version simplifiée, il n'y a plus d'obligation d'affilier tous les auteurs.
Des progrès devraient être faits dans la gestion des affiliations automatiques (notamment pour prendre en compte la date de publication de l'article et non la dernière affiliation réalisée dans Hal) mais pour l'instant, le travail humain reste indispensable.

- **Possibilité de générer des permaliens**

Chaque dépôt dans HAL a un permalien du type : URL du portail Hal, suivi de l'identifiant du dépôt .
Exemple: https://hal-pasteur.archives-ouvertes.fr/pasteur-01858477
Pour les listes de publication, des outils externes marchent très bien; on peut donner un lien une URL au chercheur qu'il pourra intégrer dans sa page web qui se mettra à jour au fur et à mesure des dépôts.
HAL gère aussi les CV des chercheurs.

- **L'utilisation de HAL par les chercheurs**

Les chercheurs ne s'en servent pas comme outil de recherche car il n'est pas exhaustif, à moins qu'ils n'aient été redirigés depuis PubMed ou Google.

Ils utilisent plus le portail générique s'ils cherchent des publications dans des domaines connexes.
L'institut Pasteur dispose d'un connecteur HAL dans leur outil de découverte EBSCO.

- **Bibliométrie**

Les principales statistiques existent (mais on trouve des erreurs pour le nombre de connexions au portail, dues à la façon d'utiliser les fichiers de log). Le CCSD est en train de refondre entièrement son module de statistiques.
A l'Institut, nous faisons des exports de notre portail Hal avec OCDHAL et travaillons avec des tableaux croisés dynamiques dans Excel pour sortir toutes sortes de statistiques.
Le CCSD est en train de refondre entièrement le module statistiques proposé par HAL.

- **Archivage sécurisé et pérenne**

L'archivage est garanti par le CINES.
Toutefois, on trouve aussi dans HAL des versions brouillons, or l'archivage ne devrait être proposé que pour les versions éditeurs.

- **Ressources Humaines**

Deux personnes et demi équivalent temps plein travaillent sur la modération. Leurs missions consistent principalement à :

- Vérifier les affiliations
- Rajouter le texte intégral
- Rajouter les projets de recherche : cette manipulation est importante car Openaire moissonne HAL et expose sur sa plateforme les articles associés a des projets de recherche européens identifiés par Openaire.

Dès le départ, Pasteur a été modérateur[63] de son portail (tout comme L'Inserm). Peu de portails sont modérateurs.
C'est l'INIST qui est en désormais en charge de la modération.
La modération se fait sur les notices avec texte intégral donc on trouve des notices sans texte intégral qui peuvent comporter des erreurs.
Depuis la mise en place de la version simplifiée, les affiliations d'auteur comprennent aussi beaucoup d'erreurs.
Du fait de la non vérification des affiliations, le tamponnage automatique est source d'erreurs. Il existe malgré tout la possibilité de détamponner, mais tout cela implique une vérification manuelle de toutes les affiliations.
A l'institut Pasteur le tamponnage est manuel : donc il faut aussi aller rechercher les dépôts faits dans les autres portails (grâce à la fonction « gérer mes collections ») où un co-auteur est affilié à l'IP.

- **Y a-t-il un informaticien dédié**

Il n'y en a pas besoin avec HAL.

- **Ce qui reste à la charge de l'institut**

La vérification des affiliations, la vérificiation de toutes les métadonnées du dépôt, le texte intégral, la licence, ainsi que la gestion des laboratoires de l'institution dans le référentiel AureHAL.

63 *Le rôle du modérateur dans HAL: https://doc.archives-ouvertes.fr/wp-content/uploads/2018/03/Charte-de-la-mod%C3%A9ration-dans-HAL.pdf

Entretien n°2: INRA - https://Prodinra.inra.fr

Entretien réalisé le 14 août 2018 *avec la responsable de l'archive ouverte Prodinra.*
Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger.

- **Présentation**

L'INRA va fermer Prodinra pour verser l'intégralité de sa production dans Hal.
Aujourd'hui Le fonds de l'INRA est géré par deux outils : un outil pour gèrer les bibliothèques et un ECM pour gèrer Prodinra. La maintenance et le développement de cet ECM sont sous-traités.

- **Les ressources gérées**

L'INRA compte 259668 références don't 17,29% avec texte intégral (indiquer la date).

- **Y a-t-il un fonds associé à ProdINRA?**

L'INRA ne gère pas de fonds papier. Il reste encore quelques bibliothèques papier mais qui sont dissociées de ce qui est déposé dans ProdINRA.

Les bibliothèques ont tendance à fermer et comme le budget pour la numérisation n'a pas été débloqué, il risque d'y avoir perte de connaissances. Certains fonds seront donnés aux écoles supérieures d'agronomie.

- **Sur quel outil fonctionne ProdINRA aujourd'hui ?**

ProdINRA fonctionne sur un ECM qui coute très cher, et qui est devenu obsolète.

- **Les raisons du choix**

Ce sont surtout des motivations politiques qui ont conduit l'INRA à verser sa production scientifique dans HAL, pour participer notamment à l'archive nationale, contribuer aux outils de mutualisation de l'Enseignement Supérieur et de la Recherche.
Derrière cette décision se trouvent aussi des raisons financières, car la gestion d'une archive ouverte coûte cher.

- **Obligation de versement dans HAL**

Il n'y a pas d'obligation de double dépôt. Un export vers HAL existe pour les UMR et partenaires.

- **Versez-vous vos publications sur une autre plateforme ? Pubmed? Repec…**

Non, le dépôt en tant que tel se fait sur ProdINRA et HAL, pas d'autres bases.
HAL et ProdInra exposent en OAI, ce qui permettra à d'autres bases de moissonner le portail INRA-HAL.

- **La migration vers HAL**

Tout sera migré. Les pièces jointes à diffusion interne ne seront pas diffusées. Les fichiers seront stockés sur un serveur INRA. L'usager sera redirigé par un lien avec authentification. Ils seront accessibles dans la migration.

- **Les formats**

Tous les formats d'exports qui n'existent pas dans HAL peuvent être développés Sauf ceux qui ne sont pas publics car HAL ne peut pas gérer les exports avec condition d'authentification.

- **Interopérabilité**

L'INRA a essayé de mettre en place un triplestore mais l'outil ECM ne le permettait pas. La tentative a été renouvelée En utilisant la solution Vivo (externe à ProdInra) mais le pilote n'a pas été satisfaisant.
La mise en place d'un triplestore n'est pas la priorité pour l'instant, et l'INRA va voir ce que propose HAL sur le sujet.

- **Les fonctionnalités gérées par Prodinraet qui ne le seront plus dans HAL**

Deux fonctionnalités seront abandonnées :

- La gestion du niveau de diffusion des pièces jointes: elles peuvent être publiques ou internes grâce au LDAP de l'INRA qui permet de demander l'authentification pour accéder au document, et un troisième niveau qui permet de gérer la notice du document même si le document n'est diffusé ni en interne ni en public ; donc la diffusion interne ne sera plus possible avec HAL qui ne permet que la diffusion publique.

- Les workflows à l'intérieur de Prodinra qui permettent d'enregistrer la notice à différentes étapes et lui attribuer différents statuts ne seront pas possibles non plus. L'outil de gestion de Prodinra permet aujourd'hui trois statuts : en cours, à relire et validés.
  Il est possible avec Prodinra de faire aller et venir la notice d'un statut à l'autre et tant que la notice n'est pas validée, elle est quand même disponible en interne (et donc disponible pour environ10 000 personnes).
  Dans HAL il n'y a qu'un statut de la notice et tant qu'elle n'est pas validée, aucun accès n'est possible.

L'INRA est donc en train de développer une base interne qui pourra être enrichie à partir d'un export de HAL. Ceci représente un coût supplémentaire, mais HAL n'ayant pas de solution de calculs d'indicateurs hors évaluation, c'est une dépense nécessaire.
Pour la production d'indicateurs bibliométriques l'INRA adoptera le modèle de l'INRIA depuis que l'institut a confié l'intégralité de ses archives ouvertes à HAL, à savoir une solution en interne.

- **Bibliométrie**

- Les indicateurs pour l'évaluation des chercheurs :

Avec Prodinra, l'évaluation est faite directement à partir de la base qui fournit l'export pour l'évaluation.
Les deux indicateurs, jusqu'à cette année étaient le taux de citation sur WOS et le facteur d'impact. Or depuis cette année, l'INRA a décidé de supprimer ces deux métriques pour évaluer les chercheurs.
Les chercheurs sont dorénavant évalués sur la base du qualitatif, à partir de la lecture d'articles, l'encadrement du nombre de doctorant...
En outre, HAL et l'INRA vont créer un développement spécifique en plus de celui de HAL.

- Les indicateurs bibliométriques :

La base Prodinra est liée aux bases RH et celles des laboratoires, ce qui permet de concaténer l'information.
Avec HAL, ces renseignements sont déclaratifs, et cela risque de poser un problème.
De plus HAL n'est pas sur des référentiels qui permettront de le relier aux bases RH.
L'INRA est donc en train de développer une base interne qui pourra être enrichie à partir d'un export de HAL. Ceci représente un coût supplémentaire, mais HAL n'ayant pas de solution de calculs d'indicateurs hors évaluation, c'est une dépense nécessaire.

Pour la production d'indicateurs bibliométriques l'INRA adoptera le modèle de l'INRIA depuis que l'institut a confié l'intégralité de ses archives ouvertes à HAL, à savoir une solution en interne.

- **Planning du projet**

Il faut compter deux ans depuis la décision de migrer dans HAL (janvier 2017). C'est un projet à long terme. La migration aura lieu en janvier 2020
Le projet a démarré avec une première phase d'études préalables.
La phase opérationnelle qui a suivi devrait durer entre 12 et 15 mois et concerne la correspondance des modèles de données pour la migration, le paramétrage du futur portail et les développements specifiques.
Enfin, la migration elle-même devrait prendre au maximum une semaine entre la fermeture de Prodinra et l'ouverture du portail HAL-INRA.

- **Les ressources humaines**

  - Pendant le projet :

Pour la partie fonctionnelle, l'INRA a recruté un chef de projet. Cette personne a dû être formée pendant plusieurs mois.
Pour la partie développement informatique, un poste de développeur a été créé.

  - Après le projet :

L'organisation actuelle du service ne changera pas : il faudra un responsable du portail HAL, et des documentalistes pour valider les dépôts et gérer les référentiels.
Au niveau des tâches à réaliser, elles resteront sensiblement les mêmes, à part quelques petites différences ; les tâches principales resteront à faire : les notices à valider, le contrôle qualité qui restera le même, le travail sur les référentiels …
Il n'y aura pas d'informaticien dédié, et développeur continuera à travailler pour le CCSD. C'est la participation de l'Inra en tant que tutelle du CCSD.

- **Ce qui reste à la charge de l'institut**

La politique de dépôt, et notamment la validation des dépôts.
Le développement informatique est désormais assuré par HAL.

Entretien n°3: IRD - hal.ird.fr/

Entretien réalisé le 4 août 2018 avec la responsable des abonnements et de la coordination du réseau documentaire et de la production de la base Horizon Pleins Textes.

Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger

- **Présentation**

Les portails sont des visualisations de Hal, ce ne sont pas des entrepôts à part. Tout ça va dans Hal, et ensuite on fait apparaitre dans des portails différents ce que l'on veut faire apparaitre.
Aujourd'hui, Hal a trois usages différents:
1. Archive ouverte (donner accès à du texte intégral pour les versions d'auteur et les versions libres de droit)
2. Être la représentation de toute la production scientifique française, établissement par établissement.
3. Un usage bibliométrique mais qui comprend aussi du déclaratif pur et simple. Le problème, c'est qu'on peut signaler un article soi-disant paru dans une revue, mais qui dans les faits n'est pas encore publié ou paru, ce qui place l'article dans la catégorie littérature grise. C'est du comptage d'activité, mais pas de la bibliométrie. Et il ne faut pas tout mélanger.

HAL est plus de la collecte que du dépôt.

- **Ressources Humaines**

C'est encore elle qui vérifie les politiques d'éditeur, la qualité du fichier, le contrôle des métadonnées;
Si l'article n'avait pas eu le droit d'être auto-archivé, alors Francine l'aurait retiré.
Ceci est le travail de Francine en tant que modérateur du portail IRD-Hal.
Elle corrige aussi les erreurs sur les noms d'auteur, ou les affiliations, et dans ce cas Francine envoie un mail à l'auteur avec une suggestion de modification.
Cette procédure permet aussi de dépister les doublons. Hal repère en effet les dépôts bien mieux qu'avant mais fait encore des erreurs.
Il y a aussi un travail d'intégrité et de qualité par rapport aux référentiels (md, auteurs, etc. °).
Francine fait aussi de l'information/ formation.

- **Alimentation**

L'articulation avec Hal n'est pas très automatisée.

Notre portail IRD HAL est alimenté automatiquement parce que le premier auteur est IRD et ça monte directement dans le portail HAL. Possibilité de détamponner.

Par la suite, une alerte a été mise en place qui signale toutes les entrées dans lesquelles on trouve la mention IRD (et toutes les formes pour désigner l'IRD).
Francine reçoit quotidiennement un mail automatique qui liste toutes les références dans Hal à l'IRD, puis prend un temps pour balayer cette alerte et voir ce qui peut être rattaché à Horizon.
Paramétrage fait pour que dès qu'on a une mention de Hal et du texte intégral, la publication aille s'afficher automatiquement dans Hal-IRD.

Entretien n°4: INED - https://archined.ined.fr

Entretien réalisé le 17 septembre 2018 avec la cheffe de projet Archive ouverte.

Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger

- **Présentation**

PolarisOS est une solution libre. Son code et la documentation qui l'accompagne sont disponibles sur GitHub.
Le requêtage se fait en Javascript.
POS n'est pas construit comme une base relationnelle. Tout est géré par des requêtes.

- **Les ressources gérées**

Article dans une revue, Pré-publication, Document de travail, Autre publication, Rapport,Thèse, Chapitre d'ouvrage, Direction d'ouvrage, Proceedings, Dossier, Poster.

- **Où hébergez-vous votre SI ?**

A l'INED

- **Facilité de prise en main**

L'utilisateur a la main sur l'interface. Les blocs peuvent être paramétrés. Tout est modulable.
Les icônes sont standards.
Il est possible d'intégrer la charte graphique de l'institut.
Possibilité d'intégrer la charte graphique.

- **Alimentation**

5 étapes de dépôt avec :
1 étape de métadonnées obligatoire (pour le chercheur)
1 étape de métadonnées optionnelles (pour le documentaliste)
On peut importer les métadonnées d'un PDF, et également copier une notice.
La société éditrice a une bonne connaissance des métadonnées

- **Les formats**

Les formats d'export sont très importants.

- **La gestion des droits d'auteur**

Tous les PDF sont en libre accès.
L'INED ne veut/peut pas tout publier, toutefois POS permet de diffuser ces documents avec 4 différents niveaux d'accès :

- sous embargo
- restreint
- confidentiel
- accès public

L'API Sherpa Romeo a été intégrée. De plus, toutes les licences CC ont été intégrées, ce qui permet d'atribuer une licence .

- **Intégration de l'API Sherpa/RoMEO**

Oui. POS permet aussi d'attribuer une licence car toutes les licences CC ont été intégrées.

- **Gestion du versioning**

Métadonnées sur les versions. La gestion du versioning est encore un peu compliquée à ce stade.
A la fin d'un embargo, on ne peut pas rajouter la version éditée par-dessus la version auteur : il faut supprimer le document et le remplacer. On peut fixer la date d'embargo, mais lorsque l'embargo est levé, il faut donc faire un nouveau dépôt.

- **Interopérabilité**

Les données sont exposables dans des formats et protocoles standards (OAI-PMH, RDF…).
Il est possible d'implémenter un triplestore.

- **Workflow**

Oui, POS permet de mettre en place des workflows de validation de dépôt.
Le workflow est intégré et seul le développeur a la main dessus.

- **Ergonomie**

Des bulles d'aide sont visibles et à disposition de l'utilisateur.
Format responsive pour une lecture sur mobile, tablette, PC, etc

- **HAL et autres connecteurs API**

Le reversement dans HAL et REPEC est automatique.
L'import par le DOI fonctionne aussi très bien.

- **Ressources Humaines**

Les RH engagées pour ce projet sont :

- 2 personnes de la documentation
- 1 éditrice en chef
- 1 personne en charge de l'analyse et du suivi de la production scientifique (pour l'évaluation).
- quelques chercheurs pour tester
- quelques documentalistes

C'est un projet qui a été initié par le service de documentation et qui a demandé deux années de travail : une avec le prestataire, pour rédiger le cahier des charges notamment (48 pages) et une autre de travail en interne.
Enormément de temps et de travail ont été consacrés à la reprise des données.

Puis, à la fin du projet, beaucoup de temps sera consacré à la formation des utilsateurs/chercheurs.
A l'issue du projet, une seule personne sera en charge de l'administration du SID.
L'administrateur aura la possibilité de paramétrer des rôles pour les utilisateurs.
L'informatique gèrera l'assistance technique et POS gèrera la maintenance.
Le développeur a conseillé de se former en interne pour permettre des développements internes.

- **POS gère-t-il des accès différenciés ?**

Oui, les droits d'utilisateurs peuvent être différenciés.

- **Les référentiels**

Possibilité d'intégrer tous les référentiels utiles, comme les projets de recherche. Cela permet ensuite à l'utilisateur de faire une recherche de tous les documents liés à tel projet de recherche (projet d'envergure européenne, projets internes, projets de l'ANR…).
A chaque référentiel correspond un champ avec liste déroulante.
Il est recommandé de récupérer l'annuaire maison et l'intégrer complètement dans la base au lieu de juste le connecter, mais on peut aussi aller chercher l'information dans le LDAP, permettant de réunir tous les auteurs.

Dans le cas d'ArchIned, le choix s'est porté sur un référentiel propre et géré par l'administrateur, donc le référentiel n'est plus adossé au LDAP.

- **Autres listes d'autorité :**

- Liste des noms de co-auteur :

Des typologies permettent des listes de co-auteur dès qu'on a les données, ce qui est beaucoup plus simple. On peut faire un export en MASAS. Le co-auteur peut aussi déposer la publication dans son espace.

- Thesaurus :

Le thesaurus d'origine a été intégré sous SKOS. possibilité de rajouter des mots-clés. L'INED l'a intégré sous XML (thesaurus traduit).

- **Les affiliations**

On peut ajouter une affiliation dans le profil du chercheur.

- **Identifiants**

Il est possible de générer des identifiants pérennes grâce à l'identifiant Handle (pour un coût raisonnable). Ces identifiants sont créés à mesure du dépôt.
Le code html peut être intégré n'importe où et permet donc d'embarquer une URL.

- **L'indexation des listes dynamiques**

POS permet de générer de la bibliographie dynamique et des rapports bibliographiques par auteur, année, type de document, projet, etc. et de les exporter sous plusieurs formats.
Le requêtage se fait en Javascript.
POS n'est pas construit comme une base relationnelle. Tout est géré par des requêtes.
L'Intégration dynamique d'URL dans les pages personnelles des chercheurs est également possible : le code html peut être intégré n'importe où et permet donc d'embarquer une URL.

- **Options de fidélisation**

Possibilité de mettre en place des flux RSS;
Possibilité d'intègrer un circuit mail pour informer le chercheur sur son dépôt, à savoir s'il a été validé ou pas.

- **Bibliométrie**

L'INED veut un référencement exhaustif car l'institut s'en sert pour les évaluations.
Les statistiques produites par l'INED sont déjà intégrées dans l'outil et n'ont pas fait l'objet de développements spécifiques.
On a plus qu'un seul outil pour gérer l'AOI et la bibliométrie.
Possibilité d'export en csv.

- **Archivage sécurisé et pérenne**

Les informations sont systématiquement sauvegardées.

- **Conclusion**

Polaris offre une solution basée sur la donnée. La gestion des référentiels est une vraie valeur ajoutée. De plus la solution présente des aspects bibliométriques intéressants.
Sa communauté est encore petite, car la solution est nouvelle, mais l'éditeur oeuvre pour son développement.

Entretien n°5: INSERM - http://www.ipubli.inserm.fr/

Entretien réalisé le 17 septembre 2018 avec l'ingénieure d'études et Chargée de ressources documentaires.

Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger

- **Présentation**

Depuis 1995, il existait un serveur ist.inserm.fr réunissant plusieurs collections documentaires sous Basis/Opentext. Ce serveur a été détruit en 2009. Le DISC/IST a entrepris alors les démarches pour trouver des solutions de dépôt et de diffusion.
- Le MeSH bilingue avec une application développée à part a été pris en charge par le CSSD.
- Pour les collections documentaires, la migration des données a été effectuée en partenariat avec l'Inist-CNRS. Le choix du logiciel DSpace est fait parce qu'à l'Inist, il était déjà utilisé pour les plateformes LARA et iRevues.

En 2008, la refonte de la plateforme a été faite avec la montée de version toujours sur DSpace.
Le logiciel DSpace est mis en place progressivement en 2010-2011 en interne. La plateforme a été mise en ligne avec les premières collections en juillet 2012.

Les archives institutionnelles de l'INSERM sont composées de deux plateformes distinctes: HAL-Inserm et iPubli
- HAL est un entrepôt où les chercheurs auteurs peuvent déposer d'eux-mêmes (en principe) leurs travaux avec leur autorisation de les mettre en libre accès,
- tandis que sur iPubli, l'administrateur assure également l'ensemble des démarches de « soumission ». DSpace fonctionne pour la pérennisation (archivage) et la publication d'ores et déjà en libre accès.
Mais DSpace est fait pour que le workflow de dépôt (soumission) – validation – embargo et autorisation d'accès soit assuré par les chercheurs et laboratoires (« communautés » de recherche) en respect de leur autonomie.
En temps normal, la plateforme est censé être gérée à la fois par l'administratrice et par l'informaticien. Elle est maintenue en collaboration tripartite entre Inserm, Inist-CNRS et le prestataire mandaté par DSpace.

- **Les ressources gérées**

iPUBLI présente essentiellement des publications scientifiques signées par l'Inserm (tandis que HAL-Inserm est destiné aux dépôts de chercheurs individuels) de types variés: article de revue scientifique, rapport, article journalistique.
iPubli a une vocation plus patrimoniale ayant pour but de mettre ces ressources à disposition des publics en libre accès, au travers du traitement documentaire à valeur ajoutée: la création de métadonnées, la consolidation de fichier, ou l'indexation via le thesaurus MeSH dont la traduction en français est assurée également par l'Inserm.

- **Ressources Humaines**

Une plateforme opérée par Dspace est censée être gérée par un administrateur et un informaticien. Le prestataire mandaté par l'éditeur assure aussi une grande partie de la maintenance et des développements demandés par le client.

Dspace peut organiser des sessions de formation afin de permettre au personnel interne de réaliser quelques opérations comme les tests nécessaires à la personnalisation de l'outil.
Le déploiement de Dspace demeure toutefois assez complexe, et seul un informaticien développeur possédant des connaissances très pointues peut le maintenir.

Cependant différents types de formations peuvent être mis en place (pour développeur/administrateur pour la mutualisation de connaissances.
Au moment de l'installation de Dspace-iPubli, en 2012, le personnel de l'IST a suivi une formation sur DSpace. Grâce à cette formation, une instance locale a été montée pour permettre des tests nécessaires en vue de la personnalisation de l'outil.

- **Maintenance**

La mise en œuvre et la maintenance sont assurées par une collaboration entre l'INSERM, et l'INIST-CNRS qui héberge les serveurs DSpace, mais aussi un prestataire mandaté par Dspace.
L'application a été développée dans un esprit collaboratif. Les membres de la société de prestation contribuent régulièrement au forum sur le wiki de DSpace, ce qui témoigne de ce mode de développement. Toute une communauté d'utilisateurs a été créée sur le plan mondial, en raison de la popularité de cet outil dans le milieu académique.

La maitrise de cet outil nécessite une haute technicité et une expertise informatique consolidée.
Lors des montées de version, c'est la DSI de l'Inist-CNRS qui déploie les codes, déposés dans un réservoir. La montée de version vers la 5.2 a représenté un grand saut technologique.
La création de modules, notamment, représente un travail de vérification considérable,
La configuration des outils associés doit être assurée par l'utilisateur.
L'exploitation de Dspace nécessite donc certaines compétences informatiques du coté utilisateur, qui doit être formé spécialement à l'utilisation de DSpace, en plus des services de maintenance assurés par le prestataire.

- **La gestion des droits d'auteur**

Le principe de iPubli destiné aux larges publics – DSpace fait que l'ensemble des publications diffusé est en libre accès au public. Pas de différence donc entre intranet et extranet.
En termes de DA, l'ensemble est détenu par l'Inserm et seul le droit d'utilisation privée est offert. Passage vers CC est envisageable mais pas sur l'ensemble des ressources, les images-photos en seront exclues.

- **Les formats**

L'outil accepte tous types de métadonnées et d'ontologies. L'outil PostgreSQL permet aux métadonnées d'être requêtables par SQL.
L'Inserm a mis en place, en collaboration avec le service numérique, un traitement intermédiaire pour calibrer les métadonnées. Lors de l'import, un fichier distinct est ainsi dédié aux métadonnées.
Dspace peut finalement accepter tout modèle de métadonnées. Il suffit d'indiquer cela sur l'interface, par exemple Marc.
L'outil accepte aussi tous types d'ontologies.
Il est possible d'appliquer d'autres formats de MD que DSpace supporte, tels MODS. En principe, il suffirait d'implémenter des autres formats pour pouvoir les utiliser. [Mais certainement il doit y avoir un développement supplémentaire et le travail de structuration de ressources dans ce sens – par ex, pour la saisie manuelle, pour le moment nous n'avons que des libellées de Dublin Core.

- **Interopérabilité**

L'entrepôt des notices est exposé au selon format Dublin Core, sur le serveur dédié et interrogeable via Dspace OAI (l'interface qui répond aux requêtes selon le protocole OAI-PMH).
Les notices sont aussi exposées en RDF. (tout comme selon d'autres formats tels MODS ou METS).

- **Workflow**

L'outil gère l'ensemble du workflow depuis le dépôt d'article vers la publication.

- **HAL et autres connecteurs API**

Les outils n'offrent pour l'instant pas de perméabilité entre les deux systèmes. Le protocole SWORD utilisé pour le reversement automatique vers HAL n'a pas été implémenté.
Les ressources déjà déposées sur PubMed sont récupérées sous forme de fichier source pour être diffusés dans iPubli.
Par ailleurs il est possible de récupérer des informations ORCID via un module à implémenter.
Pas de connecteur DOI. Se servent du Dublin Core identifier.
Dspace peut générer des identifiants HANDLE.

- **Options de fidélisation**

Un flux RSS sur la page d'accueil internet à partir de Dspace et des dernières publications est possible, de même que l'affichage des derniers dépôts et des widget d'affichage des derniers documents intégrés**.**

- **Bibliométrie**

Un module de bibliométrie peut être mis en place.
Depuis plusieurs années, il est possible d'enregistrer le trafic de l'interface utilisateur en activant l'enregistrement de données Google Analytics dans Dspace, et depuis la version 5, on peut même exposer les données Google Analytics dans Dspace.

- **Archivage sécurisé et pérenne**

Le format xml assure plus de pérennité du fichier car n'importe quelle machine peut et pourra le lire, alors que pour lire le format PDF, il faut un logiciel sur la machine.

Entretien n°6: Université de São Paulo - http://bdpi.usp.br/

Entretien réalisé échanges de courriers électroniques en août 2018 avec le chef de la Division de la gestion de l'information du Système intégré de bibliothèque (SIBi) de l'Université de São Paulo.

Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger.

- **Présentation**

Au Brésil notamment, le gouvernement brésilien a mis en place en 2012 un dispositif incitatif pour encourager l'utilisation de DSpace par IBICT (Instituto Brasileiro de Informação em Ciência et Tecnologia - http://www.ibict.br/) ce qui en a fait le logiciel le plus populaire pour les dépôts institutionnels du pays.
L'université de Saõ Paolo a intégré DSpace en 2012, après avoir hébergé un SIGB propriétaire pendant des années.
L'université utilisait depuis 1985 le SIGB Aleph (un logiciel propriétaire édité par le groupe Ex Libris) pour gérer son catalogue de notices commun à leurs 48 bibliothèques avec le SIGB.
Avant de lancer notre archive ouverte, l'université avait dans sa base 700.000 notices et 150.000 de thèses et mémoires.
Une enquête d'envergure nationale auprès de des institutions et organismes ayant une archive institutionnelle en 2013 a permis de montrer que 83.7% d'entre eux utilisaient DSpace (http://www.revistas.usp.br/incid/article/view/69327).
Les raisons de ce succès sont surtout d'ordre financier, Dspace étant un logiciel don't le code est en libre accès.

- **Les ressources gérées**

L'université de São Paulo utilise Dspace pour gérer son catalogue commun aux bibliothèques de l'université composé de plus de 700 000 notices, et plus de 150 000 thèses et mémoires.

- **La maintenance**

Au Brésil, l'IBICT, l'équivalent de l'INIST, assure l'aide au déploiement de Dspace dans les organismes utilisateurs.
Ailleurs, la maitenance est assurée par un prestataire mandaté par Dspace. Il faut savoir que DSpace est difficile à personnaliser, et la maintenance sur la partie personnalisée est également compliquée.

- **Les limites**

DSpace est difficile à personnaliser, et la maintenance sur la partie personnalisée est également compliquée.
La structure du code est compliquée (Communautés, sous-communautés et collections).
JSPUI et XMLUI sont obsolètes en tant qu'interfaces utilisateur (DSpace 7 va abandonner ces interfaces et implémenter REST et Angular par défaut)
La recherche d'informations n'est pas fiable.
L'interfaçage du catalogage est difficile à configurer et manque de vocabulaires contrôlés.

- **Existe-t-il un outil qui pourrait tout faire ? Les modules complémentaires sont-ils une nécessité ?**

Non, pour l'instant il n'existe pas d'outil qui puisse tout faire, mais j'observe le développement de Folio (https://www.folio.org/ - https://dev.folio.org/ - https://github.com/folio-org/) et je pense qu'il sera une bonne option quand il sera prêt à être utilisé. C'est un outil open source, modulaire, évolutif, et ce sera une bonne solution pour notre dépôt institutionnel et notre ILS. Et il est facile d'y ajouter un module.

- **Pourquoi préférez-vous ElasticSearch à SolR ? Et pourquoi n'avez-vous pas choisi VuFind, qui a servi d'inspiration pour construire votre système de documentation ?**

Elasticsearch et SOLR sont tous deux basés sur Lucene, mais Elasticsearch utilise JSON comme référentiel de documents pour créer l'index et a intégré Kibana. Kibana est un outil puissant qui facilite la production de rapports et expose les données agrégées aux utilisateurs finaux.
Vufind est un très bon logiciel, bien documenté. Mais nous avons des difficultés à personnaliser et à générer des rapports qui collent à nos besoins. En utilisant Kibana, nous pouvons faire un tableau de bord utile pour nos utilisateurs : http://bdpi.usp.br/dashboard.php
En outre, la maintenance d'Elasticsearch est plus facile pour moi. Je ne suis pas un développeur, je suis bibliothécaire.

- **Maintenant que vous avez construit ce système, qui est responsable de la maintenance ? Qui est responsable des nouveaux développements informatiques ?**

Au Brésil, nous traversons une crise financière et le personnel de l'Université est très limité. Aujourd'hui, le développement n'incombe qu'à notre personnel, avec tous les autres systèmes que nous utilisons. Ce n'est pas une situation idéale. Nous ne recommandons pas l'utilisation de notre logiciel, car nous ne pouvons pas bien documenter. Mais, nous essayons de créer de nouvelles fonctionnalités quand c'est possible.

- **Pouvez-vous m'en dire plus sur l'organisation du service de documentation ? Comment votre service interagit-il / collabore-t-il avec le service informatique ? Le partenariat est-il bon ?**

Nous sommes un grand réseau de bibliothèques (48 bibliothèques). Nous avons un département informatique dédié (8 développeurs) dans notre département, mais nous gérons de grands systèmes (Le plus grand catalogue sur le Brésil...). C'est très utile. Ils comprennent notre modèle d'affaires. Mais nous sommes dans une crise financière et nous ne pouvons pas contracter de nouveaux systèmes. C'est pourquoi nous devons créer des solutions et l'open source est aujourd'hui une possibilité.

- **L'Université de São Paulo semble être un pionnier des archives ouvertes au Brésil. Avez-vous inspiré d'autres universités ou instituts de recherche ? Développent-ils aussi leurs propres outils ?**

Après le lancement de notre archive par l'Université de São Paulo, de nombreuses autres universités au Brésil nous ont contacté pour obtenir plus d'informations et même après le lancement de leur propre archive. Aujourd'hui, presque toutes les universités brésiliennes disposent d'une archive institutionnelle, mais nous avons la culture du développement.

- **Avez-vous fait appel à un fournisseur de services pour intégrer DSpace ? Si oui, s'agit-il du même fournisseur qui intègre DSpace dans tout le Brésil ? Ou le marché est-il partagé ?**

La plupart des archives utilisent DSpace, mais c'est intéressant, car DSpace dans la plupart des universités est le premier logiciel open source adopté par les bibliothèques. C'est un changement culturel pour les bibliothèques car nous avons au Brésil un fournisseur structuré pour le support des logiciels open source. Le personnel informatique des universités a des difficultés à utiliser l'open source, le développement est plus difficile pour eux. L'IBICT fournit un soutien, mais limité.

- **La communauté des développeurs et des utilisateurs est-elle efficace ?**

Je pense que oui. Ce modèle de développement des outils open source est plus rapide que le modèle d'une entreprise qui vend un outil.

- **Et Polaris ? Avez-vous entendu parler de cette nouvelle solution open source ? (Il a été développé par des chercheurs français)**.

Polaris ne fournit pas de services au Brésil. Mais, je vais me renseigner à ce sujet dans la Marshall Breeding´s Library Technology Guides (https://librarytechnology.org/vendors/innovative/), je pense que c'est une option très intéressante.

Entretien n°7: Ifremer- https://archimer.ifremer.fr/

Entretien réalisé en deux fois en septembre2018avec d'abord la Directrice de la bibliothèque La Pérouse puis le maître d'œuvre d'Archimer du service ingénierie des systèmes d'information de l'Ifremer. Compléments apportés par écrit par l'administratrice d'Archimer et de la bibliothèque La Pérouse.

Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger.

- **Présentation:**

Archimer est une solution entièrement développée en interne, non libre et intégrant quelques briques libres.
Archimer repose sur une application composée d'une base de données Oracle sur laquelle sont stockées toutes les métadonnées, et d'un moteur de recherche.
Chaque nuit, Chaque nuit, le moteur de recherche indexe une partie de ces métadonnées.
Les développements internes sont faits en code java propriétaire, les scripts permettent de récupérer les métadonnées et de les mettre en forme sur le web via HTML.
Ni, CMS, ni outils de visualisation ne rentrent en jeu.

- **Les ressources gérées**

Actes de colloque, brevets, chapitres d'ouvrage, communications sans actes, diaporamas, expertises, jeux de données stockés sur Seanoe, mais visibles aussi depuis Archimer, ouvrages, posters, publications, rapports, rapports d'activité et thèse.

Ifremer n'est pas le seul producteur de données. Archimer est l'archive institutionnelle de l'Ifremer mais son périmètre est plus large : Archimer contient aussi les publications issues des campagnes à la mer de la flotte océanographique française et de TGIR comme Argo (flotteurs disséminés dans les océans et collectant des données physico-chimiques)
Seanoe est dédiée aux données de la recherche en sciences marines et héberge non seulement les données produites par l'Ifremer mais également d'autres données bancarisées à la demande d'autres instituts.

Archimer existe depuis 2005. En septembre 2010 a été mise en place l'obligation de dépôt pour l'ensemble du personnel d'Ifremer.

- **Historique**

Avant le développement d'Archimer, il existait:

- Une base de données accessible sur internet et dédiée à la littérature grise
- Une base interne qui référençait les publications dites de « rang A » (indexées par le Web of Science). Cette base recensait les publications et y associait des indicateurs mais ne donnait pas accès au texte intégral.

En 2005, Archimer ouvre et recense littérature grise et publications de « rang A » donne accès au texte intégral quand c'est possible.
Quand ça ne l'est pas, ils sont placés en visibilité intranet.
A l'heure actuelle, un système d'information de la recherche est constitué avec différentes ressources bancarisées et interopérables : images, vidéo sous-marines, données en sciences marines, catalogue des campagnes à la mer.

L'annuaire LDAP est lui aussi lié à cet écosystème via Archimer qui en alimente en temps réel la section bibliographie. Archimer référence à la demande toutes les publications des auteurs qui ont eu une carrière aussi hors Ifremer.

- **Où hébergez-vous votre SI ?**

Tout est stocké sur serveur hébergé et géré par le service informatique.

- **Quelles sont les facilités de paramétrage et d'évolution**

Tout est pris en charge par l'informaticien maître d'œuvre qui a développé la solution.
Depuis quelques années, le développement est sous-traité et l'informaticien Ifremer assure alors la maîtrise d'ouvrage.

- **Alimentation :**

Archimer est alimenté par:

- un import hebdomadaire à partir du WOS
- Des alertes quotidiennes depuis les sites des revues scientifiques dans lesquelles sont publiées les articles des chercheurs.

Cet import est semi-automatique car il faut repérer les doublons et corriger certains éléments comme les notations de formules, les affiliations, la pagination, les numéros de revues…
En outre les documentalistes doivent aussi rattacher le texte intégral.

- Des publications non issues du WOS, qui sont à saisir manuellement.

Pour les publications non issues du WOS, la saisie des métadonnées est manuelle, les personnels Ifremer sont autonomes dans leurs dépôts.
L'administrateur assure le contrôle et complément des notices.
En complément de chargements de références depuis le WOS, il est possible de pré-renseigner une publication à l'aide de son DOI

- **Les formats:**

Dublin core et export aux formats : TXT, RIS, RTF, CSV, Excel + un format de reporting
L'Import se fait par les webservices du Web of Science
Les formats en interne sont développés au fur et à mesure des besoins. Il est possible de faire correspondre ces formats.
Il est possible d'importer des références tirées du WOS, ainsi que des pré-enregistrements à partir du DOI, puisque quand la publication est accompagnée du DOI, Archimer peut aller rechercher les métadonnées sur CrossRef.
Les utilisateurs peuvent exporter des notices bibliographiques en Word, Excel, Ris, Txt...
Le système est donc assez flexible en termes d'imports et d'exports de formats.

- **La gestion des droits d'auteur**

La gestion des droits d'auteur se fait au cas par cas (l'administrateur ne dépose le texte intégral qu'avec l'accord des auteurs). Archimer n'a pas intégré un choix de licences. Les CGU et mentions légales sont en cours de révisions.
Les documents en intranet sont communiqués à la demande sur accord des auteurs (rôle de l'administrateur)
Il est possible de poser un embargo qui peut ensuite être levé par l'administrateur (rapport de test de validité quotidien produit par une routine, ce rapport signale entre autre les documents dont la date de fin d'embargo est franchie).
Archimer n'a pas intégré d'API sherpa/Romeo car elle ne présente pas vraiment d'intérêt pour ce modèle. C'est une documentaliste qui se charge de la veille dans le WOS et récupère toutes les publications dont un auteur au moins est chercheur Ifremer. Cette personne vérifie manuellement les droits d'auteur.

C'est une API intéressante pour les publications internationales, mais qui ne présente pas d'intérêt majeur pour le modèle Archimer.

- **Possibilité d'édition d'un dépôt par tous les auteurs mentionnés**

Une fois validées, les demandes de modifications doivent être adressées par email à l'administrateur.

- **Gestion du dédoublonnage**

La gestion des doublons est assurée par l'administrateur.

- **Interopérabilité**

Les données sont déjà liées via les identifiants. Ce n'est pas tout à fait du web sémantique mais cela permet déjà de repousser les métadonnées vers la plateforme Datacite qui à son tour va pousser ces métadonnées vers d'autres plateformes, dont certaines font du web sémantique.
Les données sont exposées en OAI-PMH.

- **Workflow**

Un développement interne en Java permet de gérer des workflows à deux niveaux

- **HAL et autres connecteurs IST**

L'icône ORCID qui renvoie vers le compte ORCID du chercheur.
Ifremer s'est abonné à ORCID, puis tous les chercheurs Ifremer qui disposaient déjà d'un compte ORCID ont été recensés, un numéro leur a été attribué. Les chercheurs non abonnés ont été incités à se créer un compte.

- **Référencement**

L'informaticien en charge du système est très attentif à la visibilité Google.
Les publications ne sont pas indexées par Ifremer. Les données sont liées grâce à un système de marquage manuel et de tags.
Pourquoi ? car on part du principe que le texte intégral est interrogeable.
Ce système connait des limites, car on ne peut par conséquent pas lier certains concepts entre eux.

- **Les référentiels**

Un maillage a été créé pour permettre des échanges entre les différentes bases. Il existe notamment une passerelle entre l'annuaire Ifremer et Archimer qui permet, grâce au système d'auto-complétion alimenté par le LDAP, d'attribuer un matricule au chercheur. Le N° ORCID du chercheur est également enregistré dans le LDAP.
C'est un script JAVA permet de mettre à jour les pages web des chercheurs chaque nuit.

- **Les affiliations**

Les affiliations sont importées de l'annuaire pour les personnels Ifremer, les affiliations des autres institutions sont normalisées manuellement.
Le système ne permet pas d'adosser le référentiel mis en place par Ifremer aux référentiels WOS. C'est pourquoi les affiliations sont compliquées à gérer, car tout ce travail doit être fait à la main.
Chaque fois qu'un dépôt est réalisé il y a un mécanisme d'affiliation des auteurs qui se met en marche, ce qui fait qu'Archimer dispose d'une base propre et que les équipes Ifremer sont identifiées de manière non ambigües.

- **Passerelles entre les référentiels maison et référentiels externes**

Il existe une passerelle entre l'annuaire Ifremer et Archimer. Lorsque l'on saisit un auteur dans Archimer, un système d'auto-complétion alimenté par le LDAP permet d'attribuer un matricule au chercheur. La correspondance se fait par ce matricule.
Un script JAVA permet de mettre à jour les pages web des chercheurs chaque nuit.
Le N° ORCID du chercheur est enregistré dans le LDAP. Ainsi, quand on saisit un auteur, grâce à l'auto-complétion basée sur le LDAP, on récupère nom etc, le N°ORCID, les références labos..

- **L'indexation des listes dynamiques**

Une URL pérenne est attribuée à chaque notice. A la demande, il est possible d'obtenir un DOI pour son dépôt.
L'annuaire créé par l'Ifremer permet en effet une mise à jour automatique de la bibliographie de ses chercheurs (le profil est à mettre à jour par les chercheurs eux-mêmes).

- **Création des URL/URI des PDF**

Oui mais ce ne sont pas les URLs que nous incitons à partager : nous incitons à partager le DOI si le dépôt en bénéficie, à défaut l'URL de la Landing Page.

- **Le moteur de recherche**

Une base de données est toujours accompagnée d'un moteur de recherche, toutefois le moteur proposé par Oracle présentant des faiblesses face à une recherche trop poussée, le moteur SolR a été implémenté pour améliorer les fonctions de recherche documentaire. Ce moteur a relativement bien marché mais est aujourd'hui remplacé par Elasticsearch, pour répondre à un choix d'entreprise.
Le recours à des requêtes SQL et possible et réservé à quelques personnes autorisées.

- **Options de fidélisation**

L'abonné reçoit une alerte par dépôt. l'utilisateur peut s'inscrire à un flux RSS.
Un Twitter est également alimenté automatiquement. Des bibliographies sont mises à jour dans les fiches annuaire et sur les pages des unités de recherche, des campagnes à la mer, des TGIR etc.

- **Bibliométrie**

A l'Ifremer, les chercheurs ne sont pas évalués sur la base d'indicateurs bibliométriques. La bibliométrie proposée a une visée d'appui à la recherche via des états de l'art, des cartographies thématiques, des comparaisons thématiques ou institutionnelles. La bibliométrie vient en appui au pilotage de la recherche au niveau des unités de recherche et de l'institution (HCERES, contrat quinquennal, Processus Qualité)
Archimer sert de base aux demandes bibliométriques.
Le système permet de développer facilement des scripts afin d'extraire des informations comme les publications issues de projets menés en collaboration avec l'Europe par exemple.

- **Ressources Humaines**

Tout est pris en charge par l'informaticien maître d'œuvre qui a développé la solution. Depuis quelques années, le développement est sous-traité et l'informaticien Ifremer assure alors la maîtrise d'ouvrage.
Le service informatique héberge les serveurs et stocke les données
C'est aussi le service informatique qui développe les formats en interne au fur et à mesure des besoins.
Cet informaticien était membre du personnel de la bibliothèque puis a été rattaché il y a quelques années au service informatique, ce qui lui a permis de s'investir également

dans d'autres projets de systèmes d'information et de lier les publications à tout un ensemble de ressources:
http://data.ifremer.fr/SISMER

- http://www.seanoe.org/
- http://campagnes.flotteoceanographique.fr/

Le soutien d'un prestataire SSII a été nécessaire pour les développements et la maintenance de cet écosystème.
Les documentalistes, de leur côté, assurent :

- 1,2 ETP administration et alimentation
- 0,3 ETP maîtrises d'œuvre et d'ouvrage
- non évalué les ETP dédiés à l'auto-dépôt

➢ **Archivage sécurisé et pérenne**

Seul le format PDF est autorisé. Vu le volume de documents dans ce format publié au niveau mondial, nous supposons qu'il existera des solutions pour les convertir quand ce format deviendra obsolète.
L'accès aux fichiers est garanti par le service informatique (duplication des données dans deux bâtiments …)

➢ **Comment voyez-vous l'avenir ?**

Il y a besoin de faire évoluer certaines fonctionnalités de back office afin d'alléger la charge de l'administrateur, en front office l'interrogation d'Archimer est perfectible, une mise à plat des questions juridiques est nécessaire ainsi que la clarification de la politique éditoriale.
Archimer est un système en place depuis 14 ans et qui prouve sa solidité. Des évolutions sont toujours en cours à court terme, mais pas d'évolution sur le moyen ou long terme pour un système qui fait preuve de solidité.

➢ **Comment envisagez-vous votre passage au web sémantique?**

Nous n'envisageons pas de mise en place de techniques du web sémantique à court terme (ex : Triplestore).
Cependant, les liens avec d'autres ressources (ex : auteurs, jeux de données, campagnes à la mer, …) sont mis en place à l'aide d'identifiants pérennes (Orcid, DOI). Ces couples d'identifiants s sont postés à Datacite quand le document dispose d'un
DOI. Datacite réexpose ces données qui sont ingérées par des moteurs sémantiques.

Entretien n° 8: PMB - https://www.sigb.net/

Entretien réalisé le 3 septembre 2018 avec la responsable commerciale.

Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger.

- **Présentation**

PMB est un SIGB open source. L'éditeur propose une prestation de service autour, mais l'application peut être développée par des informaticiens.

- **Les formats**

Tous les formats de métadonnées sont supportés.

- **Interopérabilité**

La particularité de PMB est de s'être tourné vers le web sémantique. C'est l'un des premiers éditeurs à avoir FRBérisé son logiciel.
PMB s'est penché sur l'opprortunité pour les bibliothèques numériques de se lancer dans la modélisation de données et rendre leur catalogue interopérable, et avons interrogé la responaible du pôle informatique de l'entrepôt DataPersée.

- **Connecteurs API**

PMB offre aussi une solution de GED complète, et peut s'interfacer avec d'autres systèmes de GED.
En revanche, il n'y a pas de connecteur WOS pour les imports en masse.

- **Les référentiels**

La solution gère les affiliations en liant ce champs à une notice d'autorité. Elle permet aussi de gérer autant de thesaurus que nécessaire.

- **Bibliométrie**

PMB est branché directement sur la partie gestion. On peut aussi y intégrer des statistiques de fréquentation, de consultation et connecter le système avec MATOMO et Google analytics.

- **Archivage sécurisé et pérenne**

Enfin, PMB offre une solution d'archivage en mettant à disposition un entrepôt d'archives au format du SIGB qui peuvent être convertis dans d'autres formats, notamment les formats MARC, PDF/A et XML.

Comme la plupart des SIGB, les options sont multiples et personnalisables. Mais comme la plupart, c'est un système qui va s'adapter à chaque type de bibliothèques, mais n'est pas spécifique aux exigences d'open access.

Entretien n°9: PolarisOS - https://www.mysciencework.com/polaris-os

Entretien réalisé le 7 août 2018 avec Directeur de l'innovation et du développement.

Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger.

- **Présentation**

Le projet Polaris est né en 2014 et est complètement intégré à MyScienceWork. L'idée de départ était qu'une page pouvait tout gérer (publications etc).
La 2ème version est sortie après avoir répondu à un AO par l'INED.

- **Hébergement**

Polaris propose un hébergement sur Cloud ou sur des serveurs dédiés chez le client.

- **L'installation et la maintenance sont-elles gérées par l'éditeur ? un prestataire ? Quelles sont les facilités de paramétrage et d'évolution**

Polaris a souhaité mettre en place une solution «User friendly» basée sur la technologie ouverte, facilement exploitable par le client, de manière à ne pas le placer dans une situation de dépendance face aux prestataires.
La brique CMS permet de gérer des contenus et des workflows qui permet de facilement intégrer des couleurs, par exemple. Une part des paramétrages et options doivent pouvoir être réalisables par l'administrateur. L'idée étant de pouvoir administrer sans forcément avoir une équipe technique énorme, maisque quelqu'un qui ait de l'expérience et de l'appétence pour ces technologies puisse le faire.
Possibilité de maitriser complètement les flux de données et les workflows de publication, tout en assurant une 'user experience' pour que les chercheurs comprennent l'intérêt de déposer leurs publications.
Un des axes de développement de Polaris est de créer une solution qui permette d'ajouter facilement de nouvelles fonctionnalités et de capitaliser sur ce qui a été fait.
L'objectif était aussi d'offrir une alternative sur le marché des solutions ouvertes.

- **Alimentation**

POS permet d' alimenter un catalogue gérant un fonds papier ET une archive ouverte institutionnelle.
Le système permet une pré complétion des champs grâce à une extraction des métadonnées du PDF pour remplir les champs.
Le système peut aussi faire des exports de données que les chercheurs peuvent réutiliser dans leur propre bibliographie.

- **Gestion des doublons**

En cours de développement. L'identification des doublons se fera à différents niveaux: dans les publications, mais aussi dans les référentiels (mots-clés libres, thesaurus...)

- **Les formats**

POS est une solution basée sur la donnée.
La brique ExtraTransformLoad permet d'attribuer un format de son choix aux données récupérées à partir de sources diverses. Cette fonctionnalité permet de coller très finement au besoin.
Le modèle de données est donc entièrement flexible et rend la plupart des formats importables et exportables.

- **La gestion des droits d'auteur**

POS gère les deux niveaux.
Une fonctionnalité permet d'aller chercher des données dans Sherpa Romeo.
Les chercheurs peuvent aussi déclarer la licence qui va avec l'article et ont même la possibilité de rentrer la date de fin d'embargo afin que l'article soit automatiquement diffusé en OA.
Polaris propose aussi un système de request copy. L'utilisateur peut envoyer une demande à l'auteur pour partager sa publication.
Il existe à ce niveau un workflow pour gérer cette fonctionnalité.

- **Interopérabilité**

L'objectif de l'éditeur était de créer une solution rendant les archives interopérables.
Il est également possible de mettre au point les développements nécessaires pour que toute base de données puisse moissonner les données du client, et faire ressortir ces données dans les résultats des recherches. C'est possible par exemple avec Repec.
POS est une solution entièrement interopérable, aux niveaux entrée et sortie.
Pour Polaris, l'expression « nouvelle génération » signifie que les archives doivent être interopérables.
Le logiciel peut ainsi aller chercher dans les bases fermées car la plupart des gros éditeurs proposent des API.
Il est aussi possible de mettre en place un triplestore.

- **Indexation des listes dynamiques**

On peut générer un code HTML pour mettre à jour automatiquement les listes de publications sur le blog du chercheur.
Grâce à la spécification Handle, Polaris peut attribuer automatiquement un identifiant unique au moment du versement de l'archive, via un lien pérenne.
La BNF a développé ARK qui peut être aussi intégré dans le logiciel.

- **HAL et autres connecteurs API**

Le socle de Polaris est composé entre autres d'une brique ERP grâce à laquelle il est possible de récupérer n'importe quelle ressource sur le web sous forme d'API.
- Polaris permet aussi au chercheur de se créer un numéro ORCID et de s'y connecter pour synchroniser les informations.
- Possibilité de reversement dans Pubmed et Hal.
- Polaris est aussi compatible avec un moissonnage Google Scholar.
- Possibilité d'imports en lots de données extraites de Zotero (RIS) et EndNote.
- Possibilité également d'importer et exporter du CSV.
- Connexion avec la base crossRef via le DOI du document.

- **Les référentiels**

Le thésaurus a été intégré sous SKOS et répond donc aux recommandations de la COAR.
Les référentiels sont gérés depuis la plateforme administrateur. Il est possible d'intégrer n'importe quel type de référentiel.

- **Les affiliations**

Possibilité de gérer les affiliations et même d'ajouter un auteur si on ne trouve pas l'affiliation (ce process peut être soumis à workflow).

- **Bibliométrie**

Le cœur de l'expertise de Polaris est d'offrir la possibilité de développer un dashboard sur mesure (avec graphes relationnels), des statistiques de productivité et aussi de mettre en relation les partenariats les plus fréquents.

Grâce àElasticsearchet la manière dont on récupère les données, on peut mettre en place des bases de données relationnelles. Les données sont ainsi transformées dans un autre modèle et on peut créer un graphe pour lesanalytics.
De plus, en adaptant la technologie de la blockchain, le logiciel permet de connaitre les tendances de la recherche et offre ainsi la possibilité de faire des croisements et recoupements.
L'évaluation se fait par une extraction souple à partir du profil du chercheur.

- **Moteur de recherche**

Polaris a choisi Elasticsearch qui permet de la collecte de la donnée et d'analyse. Beaucoup de données sont stockées, et ce moteur de recherche permet une bonne rapidité et une bonne finesse de recherche.

- **Archivage sécurisé et pérenne**

L'archivage est garanti grâce à l'identifiant pérenne ARK (de la BNF).
Le client peut mettre en place une solution d'archivage pérenne à partir de son propre système d'archivage en attribuant un identifiant ARK pour chaque publication.

- **Fluidité**

La solution doit être responsive: sur tablette et mobile.
C'est aussi une solution réactive car seuls les nouveaux éléments se chargent et permettent donc une navigation fluide.

- **Communauté d'utilisateurs**

POS est une solution récente, ne disposant pas encore d'une très grande communauté.
La constitution d'une communauté de contributeurs est un axe de développpement.
Les créateurs de POS souhaitent pouvoir pousser les universités et les instituts de recherche à recourrir à la solution Polaris sans forcément faire appel aux éditeurs ou à des développeurs pour l'intégration.

- **Evolution du système**

La gestion des données de la recherche peut être un projet futur, qui peut être implémenté.
D'autres fonctionnalités comme le système de peer reviewing pour pouvoir publier depuis l'archive ou faire le lien avec une base de peer reviewing sont en cours.

- **Web sémantique**

Polaris permet de gérer un corpus énorme de data en entrée comme en sortie.
C'est un système qui s'inscrit dans le web sémantique, grâce à son interopérabilité.
D'autres fonctionnalités comme le système de peer reviewing pour pouvoir publier depuis l'archive ou faire le lien avec une base de peer reviewing sont en cours.

Entretien n° 10: DataPersée : http://data.persee.fr/

Entretien réalisé le 31 août 2018avec responsable pôle informatique.

Les propos retransmis ici sont de l'entière responsabilité de l'auteur et ont été transcris avec le plus de fidélité possible. Si des malentendus ou des contresens sont à signaler dans le texte ci-dessous, l'auteur en endosse l'entière responsabilité et mettra tout en œuvre pour les corriger.

Le portail Persée est complet et fonctionne bien. Toutefois, en allant sonder les utilisateurs, les gérants du portail ont réalisé que le corpus exposé devenait très étoffé et qu'il n'était plus seulement utilisé comme une bibliothèque.
L'approche bibliothèque n'était plus suffisante face aux besoins des chercheurs. Le portail était en train de devenir un entrepôt de données, et beaucoup de demandes allaient dans ce sens. Les chercheurs demandaient de l'aide sur des requêtes dont le but était de récupérer tous les documents mis à disposition par Perséesur tel sujet.

Le but aujourd'hui consiste souvent à récupérer un corpus et le téléchargement de PDF devient secondaire. L'objectif derrière ces requêtes est de croiser des métadonnées ou des données de recherche, et permettre aux chercheurs de constituer leurs propres bibliographies thématiques, ou de réunir tous les ouvrages d'un même auteur, ou encore toutes les informations concernant un lieu, etc.

Or, le portail Persée ne permet pas d'exploiter son corpus dans ce sens.
Il aurait été possible de brancher le triplestore sur la bibliothèque existante, mais cela aurait entraîné un risque de confusion pourl'utilisateur. Persée a alors choisi d'exposer deux interfaces bien distinctes.

La réflexion sur ce sujet est en cours depuis 2009/10 et Data Persée a ouvert en 2017.
La FRBérisation a été envisagée, mais finalement laissée de côté car non adaptée aux chercheurs qui avaient acceptés d'être testés.
L'équipe projet s'est donc orientée vers un modèle plus allégé, et le format RDA a été choisi.
Des listes d'auteur avaient aussi été mises en place (pour les listes d'auteur d'articles scientifiques) mais l'interrogation de listes en SparQL n'est pas évidente et cette option a été écartée.
Le triplestore a été confié à une architecture del 'information, ainsi que le site web qui l'abrite.

Techniquement, le portail Persée est essentiellement alimenté avec des données en XML, plus d'autres formats comme du Dublin Core.
Un certain nombre de traitements ont été mis en place, leur permettant de produire duR DF qui est ensuite versé sur letriplestore.
Cette opération a été réalisée par le service informatique de Persée, sans faire appel à des développeurs.
Comme les jeux de données du portail Persée, sur lesquels les documentalistes et informaticiens avaient travaillé en amont étaient très propre et bien maitrisé, ce travail n'a pas été très compliqué et a demandé seulement quelques semaines de travail. Ils ne se sont pas appuyés sur des applications pour procéder à la conversion des données et Persée a développé ses outils, notamment des scripts.

Le triplestore repose sur l'outil Virtuoso, installé de façon basique.

L’interconnexion est assurée en amont chez Persée, au niveau de la production des données.
Les identifiants des chercheurs sont associés à Idref, le référentiel de l’ABES. En outre, d'autres alignements vers d'autres référentiels sont recherchés de façon continue.

Les citations des articles sont aussi recherchées pour créer du lien dans les données ancrées dans cet écosystème.
Ensuite, les données sont interconnectées grâce aux référentiels. Ce qui permet d'avoir des données du SUDOC, de DBpedia ou encore de la BNF apparaître sur le portailPersée (ce sont les pages de mashup*).

Au niveau de DataPersée, la requête permet d'interroger à la fois l'ABES et Persée.

Les utilisateurs peuvent lier ensemble les données de leur choix.
Les notices sont importables dans la plupart des formats, cela dépend surtout des outils de l'utilisateur.

Les moyens humains requis comprennent la cellule informatique comprend une équipe de recherche et développement.
L’équipe compte 5/6 informaticiens, parmi eux un professionnel de l'information.
Ils ne font pas qu’héberger la plateforme, mais la gèrent entièrement.

L’objectif de cette entreprise était aussi de valoriser le travail sur la partie liaison des données, mais aussi les partenariats avec lesinstitutionsqui hébergent les référentiels.

Enfin, un réel effort a été porté sur toute la partie pédagogique. Le site de DataPersée comprend toute une partie faite de tutoriels et de pages expliquant la technologie d’un triplestore, ses avantages, ce qu’il offre de plus par rapport à une bibliothèque numérique traditionnelle.
Les chercheurs peuvent effectuer leurs recherches via le sparql endpoint directement, mais l’outil Sparklis leur permet aussi de poser leur requête en langage naturel.
Les statistiques de consultation montrent que les chercheurs sont de plus en plus nombreux à utiliser les services de DataPersée.
Les retours sont surtout quantitatifs, et l’on ne dispose pas encore de retour d’expérience des chercheurs eux-mêmes.
Un programme de formation est en cours d’élaboration, afin d’aller à la rencontre des laboratoires, car l’équipe qui a mis en place DataPerséeaimerait que les chercheurs partagent plus leur expérience.

Le futur sera envisagé quand Persée aura eu un retour des utilisateurs, donc lorsque le programme de formation aura été lancé.
D’autre part, l’équipe informatique recherche actuellement des procédures pour alimenter le triplestore de façon régulière.
L'ontologie vaêtreenrichie pour représenter des ressources variées, comme la biologie.
Toutes les données de taxonomie trouvées dans les documents vont être ajoutées.