CollEx-Persée
Numérisation – Projet collaboratif

Exhaustivité électronique du fonds documentaire de l'IRD « 2eFDI »

Rapport scientifique de fin de projet

Pier Luigi ROSSI
IRD
Mission culture scientifique et technologique
Service de l'information scientifique et technique

Juillet 2021

1. Fiche d'identité du projet

Thématique	Numérisation – Projet collaboratif
Titre du projet	Exhaustivité électronique du fonds documentaire de l'IRD « 2eFDI »
Établissement porteur du projet	IRD (Institut de Recherche pour le Développement)
Equipe de recherche concernée	Service Information Scientifique et Technique (Mission Culture Scientifique et Technique de l'IRD)
Corpus concerné	FDI - Fonds Documentaire de l'IRD
Indiquez si le corpus concerné est labellisé CollEx	Oui (identifiant labellisation: Fonds documentaire de l'IRD)
Actions à financer	☑ Numérisation
Durée globale du projet (avec reports « COVID »)	12/03/2019 – 30/06/2021
Liste des partenaires	IFRIS (Institut Francilien Recherche Innovation Société)
Résumé du projet	Le Fonds Documentaire de l'IRD (FDI) est l'archive institutionnelle de l'institut, il recense l'ensemble des publications scientifiques produites par les personnels de l'organisme depuis sa création. Au mois de juillet 2021 plus de 75 500 documents du FDI sont en libre accès sur internet (environ 66 500 documents au mois de septembre 2018 lors de la soumission du projet). Le projet 2eFDI visait à rendre accessible environ 10 000 nouveaux documents du FDI afin d'atteindre une mise en libre accès de la quasi-totalité des productions scientifiques de l'IRD depuis sa création. Cet ensemble de documents constitue probablement la plus importante collection française (et internationale) en matière de recherche pour le développement des régions intertropicales disponible en texte intégral et en libre accès sur Internet. Le public visé est composé en priorité par les chercheurs, doctorants et étudiants, du Nord comme du Sud, travaillant sur les problématiques des pays des Suds. La diffusion en libre accès de cette collection répond également à une logique de restitution des documents et données concernant de nombreux pays du Sud vers les équipes de recherche, les citoyens, les internautes des pays concernés, tout particulièrement en Afrique.

2. Description scientifique du projet

2.1 Environnement institutionnel

L'IRD (Institut de recherche pour le développement) est un établissement public français à caractère scientifique et technologique (EPST) placé sous la double tutelle du Ministère de l'Enseignement supérieur et de la Recherche et du Ministère des Affaires étrangères et du Développement international.
Depuis 1955, l'Institut a créé (décret du 1er décembre 1955) une archive institutionnelle avec la volonté de préserver et de diffuser ses productions scientifiques. Ce **fonds documentaire patrimonial** se compose actuellement de 105 000 documents et constitue **le fonds documentaire de l'IRD (FDI)**. Le FDI est géré par le Service de l'information scientifique et technique (Mission Culture Scientifique et Technique de l'IRD). Cette collection, **labellisée Collex en décembre 2017**, a été informatisée depuis 1986 avec la création de la base de données bibliographiques Horizon[1]. Dès 1996, sa numérisation a été lancée[2] et 75 500 documents sont aujourd'hui disponibles gratuitement sur Internet.

2.2 Orientations scientifiques du projet

Objectifs

Le Fonds Documentaire de l'IRD (FDI) est l'archive institutionnelle de l'institut, il recense l'ensemble des publications scientifiques produites par les chercheurs de l'organisme depuis sa création en 1943. Ce fonds est pluridisciplinaire, à l'image des recherches menées par l'institut (Sciences de l'Ingénieur, Sciences de la Santé, Géosciences, Sciences animales et végétales, Sciences Humaines et Sociales, Climatologie et Environnement, Océanographie, Hydrobiologie et Halieutique). L'autre spécificité du FDI concerne le contexte géographique des documents, lié aux missions de l'IRD : la plupart des travaux ont été réalisés dans les pays du Sud et portent sur la zone intertropicale et la région méditerranéenne.

Le FDI se compose actuellement de 105 000 (98 400 documents au moment de la soumission du projet en octobre 2018). **La numérisation du fonds et une politique de libre accès ont été mises en place dès 1996 avec le projet Pleins_Textes**[3], et au moment du bilan du projet 2eFDI 75 500 documents sont disponibles en texte intégral, au format pdf et gratuitement sur Internet (66 500 documents au moment de la soumission du projet en octobre 2018).

Le projet 2eFDI visait à numériser de façon systématique environ 10 000 documents du fonds documentaire de l'IRD pour rendre disponible de façon exhaustive en mode électronique la production scientifique de notre Institut depuis sa création. Le projet Pleins_Textes a été initié en

[1] ROUX-FOUILLET, Jean-Paul (1988). Horizon : base bibliographique ORSTOM : présentation. In : *Séchet Patrick (ed.). Séminfor 1, premier séminaire informatique de l'ORSTOM : bases de données et systèmes d'information : quelles méthodes ?* Paris: ORSTOM, pp. 285-296. Disponible à : https://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_4/colloques/26249.pdf.

[2] ROSSI, Pier Luigi (1997). Economie et portabilité : une chaîne d'édition électronique destinée à la dissémination de l'information primaire. In : *Forum initiatives 97*. Hanoi : AUF, 6 p. multigr. Disponible à : https://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_6/divers1/010022348.pdf.

[3] ROSSI, Pier Luigi ; NGOMA-MOUAYA Marcel (2000). "Pleins_Textes": IRD (Institut de Recherche pour le Développement) electronic library. In: *Online information 2000 proceedings*, pp. 201-206. Disponible à : https://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_5/TAP/010024168.pdf.

1996 et actuellement cette collection de fichiers pdf en libre accès constitue une des plus importantes collections scientifiques concernant les problématiques des pays en développement et les régions intertropicales du monde. Cette collection a été labellisée « Collex » pour 5 ans (2018-2022) en décembre 2017.

Etat de l'art

Le projet 2eFDI constitue la dernière phase de numérisation en masse du FDI (projet Pleins_Textes) : après son lancement en 1996, le projet Pleins_Textes a bénéficié du soutien du Ministère de la Recherche via des projets comme « Griseli[4] » et « BSN5[5] », mais l'IRD a consenti au cours de plus de 20 ans à des investissements conséquents pour rendre disponibles aux communautés scientifiques, notamment celles des pays des Suds, ses productions scientifiques originales et fondamentales pour le développement.
En parallèle au projet Pleins_Textes, nous avons entrepris le transfert de compétences vers nos partenaires africains en matière de numérisation des documents et de création de bibliothèques électroniques. Depuis 2001, nous avons pu installer et mettre en service 45 ateliers de numérisation dans 13 pays d'Afrique francophone (Algérie, Bénin, Burkina Faso, Cameroun, Côte d'Ivoire, Madagascar, Mali, Maroc, Niger, Sénégal, Tchad, Togo, Tunisie). Ces activités nous ont permis d'effectuer un transfert de compétences au bénéfice de plus de 200 professionnels de l'information scientifique et technique (bibliothécaires, documentalistes, techniciens) dans 82 institutions[6].
En 2010, nous avons mis en exploitation le serveur Internet BEEP (Bibliothèques électroniques en partenariat), hébergé sur les serveurs de l'IRD. BEEP donne accès à plusieurs collections produites des partenaires du projet[7].
Les expériences acquises dès 1996 en matière de numérisation et de création de bibliothèques électroniques nous ont permis de mettre au point des processus de production de documents électroniques (numérisation ou productions de fichiers pdf à partir de sources électroniques) que nous utilisons en interne et que nous avons transmis à nos sous-traitants (définition des caractéristiques des fichiers pdf, mode de compression des images, segmentation des éléments des pages numérisées, injections des métadonnées, optimisation des fichiers pour l'internet).
Enfin, nous accordons une très grande importance à l'analyse des consultations des documents ainsi disponibles en libre accès, tant pour la base Horizon que pour les collections hébergées par BEEP[8].

Publics cibles

En mettant en libre accès sur internet les publications scientifiques de l'IRD, on peut considérer que **les publics cibles sont les internautes** (des chercheurs, des étudiants, des citoyens, des décideurs, …) qui recherchent des documents scientifiques et techniques concernant les problématiques de la

[4] Griseli : *Mise en place d'un système d'accès la littérature grise publique, scientifique et technique* (projet du Ministère de l'Enseignement supérieur et de la Recherche), 1996-1997.
[5] FARGIER, Nathalie (2015). Numériser la littérature grise scientifique. In *I2D Information, données et documents*, vol. 52, n° 1, pp. 61-62. Disponible à : https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-1-page-61.htm
[6] ROSSI Pier Luigi (2018). Numérisation, bibliothèques électroniques et libre accès : entre renforcement de capacités et perspectives en Afrique francophone. Bondy : IRD, 12 p. multigr. Conference West and Central African Research and Education Network (WACREN), 2018/03/12-16, Lomé. Disponible à : https://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers18-03/010072585.pdf
[7] ROSSI Pier Luigi (2011). Electronic libraries in partnership: BEEP for Africa. *African Research and Documentation*, 2011, (115), p. 69-75. ISSN 0305-826X. Disponible à : https://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-11/010052810.pdf
[8] ROSSI Pier Luigi, TRAORE M., DIALLO F.M. (2018). Publications en libre accès des universités du Burkina Faso : analyse d'impact et visibilité internationale. *027.7 : Zeitschrift für Bibliothekskultur*, 2018, 5 (1), art. no 027.7 - p. 52-64. ISSN 2296-0597. Disponible à : https://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers18-02/010072183.pdf

recherche pour le développement dans les pays des Suds. Cette analyse est confirmée par les études statistiques que nous effectuons sur les consultations de nos documents qui sont déjà en libre accès sur internet. Sur les trois millions de « accès-utilisateurs »[9] enregistrés et validés en 2020, 50 % des consultations relèvent de pays d'Afrique, 70 % des pays en développement[10].
Tous les documents sont au format pdf et en texte intégral : au-delà des citations bibliographiques dont bénéficient les documents en libre accès et des usages des contenus (découvertes, thèses soutenues, données, cartographies, …), l'utilisation et l'extraction du texte intégral offrent une perspective d'exploitation déjà largement appliquée tant en interne que par des utilisateurs externes à l'IRD.

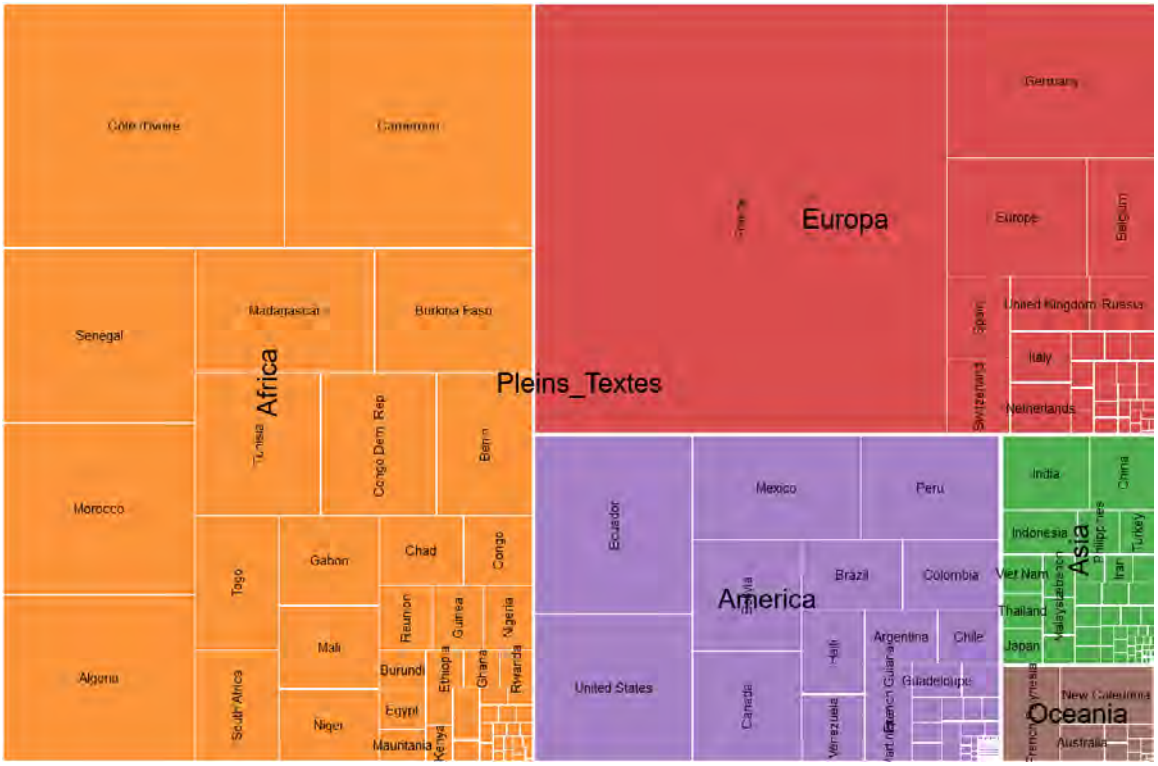

Figure 1. Géolocalisations des consultations des documents en libre accès de l'IRD (Horizon-Pleins_Textes) pour l'année 2020

2.3 Description du corpus qui a été numérisé

Le projet 2eFDI visait à atteindre l'exhaustivité en version électronique afin de rendre accessible en libre accès la plupart des documents qui composent ce fonds patrimonial original et unique. La dynamique du projet faisait suite à une période de trois années (2016-2018) au cours de laquelle nous avions mis en place une stratégie de numérisation systématique des documents concernant la Pédologie (projet financé avec l'appui de BSN5) et l'Hydrologie. Avec le projet 2eFDI, il s'agissait de poursuivre ces dynamiques et d'étendre la démarche de numérisation systématique à toutes les disciplines couvertes par les recherches menées à l'IRD.

[9] Les « accès-utilisateurs » représentent des consultations des documents que nous estimons avoir été réalisées par des humains. Pour identifier ces accès nous avons développé des procédures qui permettent de filtrer les fichiers de log du serveur afin d'isoler les accès faits par les robots d'indexation et ceux pouvant être identifiés comme des spams.
[10] Le tableau de bord complet des statistiques de consultations pour l'année 2020 est accessible à : https://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/2020.htm.

La sélection des documents à numériser comportait des rapports scientifiques, des thèses, des mémoires, des ouvrages, des travaux de littérature grise (études de terrain) et des articles portant sur des travaux de recherche réalisés dans différents pays d'Afrique, d'Amérique et des Régions et Communautés d'Outre-mer. Les études de terrain concernent des territoires très spécifiques pour lesquels peu de connaissances scientifiques sont disponibles. Elles contiennent des informations hautement originales en matière de données scientifiques, de cartes thématiques, d'iconographie, de méthodologies d'approche dans un contexte tropical. Dans bien des cas ces études sont uniques tant pour la zone géographique concernée que par les connaissances qu'elles véhiculent.

Le tableau 1 résume la répartition thématique des documents qui ont été numérisé et le tableau 2 résume leur distribution en fonction de quatre périodes de publication.

Sciences de l'ingénieur	684
Sciences de la santé	1434
Géosciences	871
Sciences animales et végétales	2351
Sciences humaines et sociales	2392
Climatologie, environnement	501
Océanographie, hydrobiologie	2199

Tableau 1 : Répartition des documents par domaines scientifiques

1943-1959	489
1960-1979	1540
1980-1999	4315
2000-2012	4088

Tableau 2 : Répartition des documents par année de publication

Les documents publiés entre 2013 et 2018, répertoriés au Fonds documentaire de l'IRD, n'étaient pas concernés par ce projet de numérisation. L'équipe du Service IST de l'IRD a mis en place, depuis juin 2018, une procédure de recherche des sources numériques des documents les plus récents (éditeurs scientifiques, archives ouvertes, sites des unités, contacts avec les chercheurs) pour les déposer en libre accès ou en accès réservé sur les serveurs du FDI.

2.4 Résultats et impact

Le tableau 3 présente les principales étapes qui étaient identifiées pour la réalisation du projet 2eFDI avec leurs indicateurs de réussite.

Tâche	**Personne / partenaire responsable**	**Livrables**	**Indicateurs de réussite**
Préparation des documents	Equipe projet IRD + sous-traitant	Documents papier identifiés code barre	10432 (100 % des documents prévus)
Numérisation par le sous-traitant	Sous-traitant	Fichiers pdf et retour des originaux papier	10432 (100 % des documents prévus)
Contrôle qualité	Equipe projet IRD	Fichiers pdf contrôlés et validés	8000 (77 % des documents prévus)
Diffusion : mise en accès internet	Chef de projet	Fichiers pdf « optimisés web »	8000 (77 % des documents prévus)

Tableau 3. Principaux indicateurs de réussite du projet 2eFDI

En ce qui concerne la sélection et la préparation des documents ainsi que leur numérisation chez le sous-traitant, **tous les documents prévus ont été traités**.

En ce qui concerne le contrôle qualité et la diffusion des documents sous la forme de fichiers pdf, au moment de la rédaction de ce bilan (juillet 2021), environ 8 000 documents ont été traités et sont disponibles en libre accès sur Internet. Ces traitements de finalisation des documents déjà numérisés sont actuellement en cours : compte tenu des difficultés de production et des décalages des calendriers à la suite de la crise sanitaire « COVID », nous envisageons la mise en accès intégrale sur Internet à la fin de 2021. Toutefois ces actions sont continues et progressives : notre capacité de finalisation et de mise en libre accès sur Internet est d'environ 450-500 documents par mois.
Par rapport aux données et aux fichiers disponibles en libre accès, en relation avec le projet 2eFDI, au moment du dépôt du projet (octobre 2018) nous avions indiqué que 66 500 documents étaient disponibles en libre accès sur Internet. A la date de la rédaction de ce rapport, nous avons donc pratiquement atteint les objectifs du projet : dans la fiche du résumé du projet nous indiquions une progression d'environ 10 000 documents pour atteindre un seuil de 76 500 en libre accès.

La dynamique de mise en libre accès sur Internet des documents du Fonds documentaire de l'IRD est une constante du projet Pleins_Textes. Le tableau 4 en illustre les performances depuis 2010.
L'IRD avait identifié dans son « Contrat d'objectifs et de performance à l'horizon 2020 »[11] l'indicateur 13 qui prévoyait une progression de 25 % du nombre de publications en libre accès sur la période 2015-2020. Grâce aux performances de production de Pleins_Textes combinées au soutien du Collex-Persée via le financement du projet 2eFDI, sur la période 2015-2020, la progression a été de 31,4 % : ainsi les résultats attendus ont été dépassés de façon significative. Ces résultats contribuent de façon efficace à la visibilité internationale de la production scientifique de l'IRD.

Nous réalisons des statistiques avancées sur les consultations des documents en libre accès. Nous

Année	PDF en libre accès	PDF numérisés sur l'année	Progression	Progression 2015-2020
2010	43311			
2011	45811	2500	5.8%	
2012	47711	1900	4.1%	
2013	49041	1330	2.8%	
2014	51288	2247	4.6%	
2015	55619	4331	8.4%	
2016	60157	4538	8.2%	
2017	64765	4608	7.7%	
2018	67236	2471	3.8%	
2019	69974	2738	4.1%	
2020	73077	3103	4.4%	**31.4%**

Tableau 4 : Progression du nombre des documents en libre accès sur les dix dernières années

associons les données produites par les logs de nos serveurs aux métadonnées décrivant d'un point de vue bibliographique et catalographique les documents disponibles. Cette approche nous permet de produire des tableaux de bord selon plusieurs entrées et proposant des indicateurs innovants et originaux[12].

[11] Le document est accessible à : https://www.ird.fr/documents-strategiques.
[12] Voir, par exemple, le tableau de bord complet des consultations des documents en anglais accessible à : https://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/ES-2020.html ou un tableau de bord « auteur » accessible à : https://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/Favreau.htm.

En 2020 nous avons enregistré 2 858 908 de « accès-utilisateurs » avec en moyenne 45 consultations par fichier (tableau 5). Il est à noter qu'il y a eu 668 documents qui ont été consulté au moins 668 fois (facteur D[13] = 668).
La mise en libre accès de 10 000 nouveau documents du fonds documentaire de l'IRD s'inscrit dans la dynamique de consultation déjà largement constatée jusqu'à présent avec les données factuelles des consultations : ainsi cette dynamique pourrait engendrer environ **quatre millions de « accès-utilisateurs » en 2022**.

Indice	Valeur
Nombre de fichiers "Pleins_Textes" consultés	64 103
Nombre de consultations "Pleins_Textes"	2 858 908
Nombre moyen de consultations par fichier "Pleins_Textes"	44.60
Facteur D (nombre de documents consultes au moins "n" fois)	668
Top D (668) sur total fichiers consultés	1.0%
Somme des consultations des Top 668	905 756
Consultations Top 668 sur total des consultations de 2020	31.7%

Tableau 5. Principaux indicateurs de consultation des documents de l'IRD en libre accès pour l'année 2020

2.5 Calendrier du projet

Initialement prévu avec un déroulement sur 14 mois, le calendrier du projet a été revu par deux fois.
En 2019, le sous-traitant devant assurer la numérisation des documents a connu un fort ralentissement de ses capacités productives suite à la fermeture du site de Neuilly sur Marne et une délocalisation des activités à Verrière les buissons. Une partie importante des personnels n'a pas suivi ce déplacement des activités et l'équipe de numérisation n'a été que partiellement remplacée.
A cette situation inattendue s'est greffé, à partir de mars 2020, le confinement lié à la crise sanitaire « COVID 19 ». Face à ces difficultés qui ont fortement limité la numérisation des documents, nous avons demandé un premier report du projet. En juillet 2020 la Direction du GIS Collex-Persée nous a accordé un report de la date de fin de projet au 12 février 2021.
Suite au deuxième confinement lié à la crise sanitaire « COVID 19 » et aux difficultés rencontrées par notre sous-traitant, la Direction du GIS Collex-Persée nous a accordé, en février 2021, un deuxième report de la date de fin de projet au 30 juin 2021.

2.6 Partenariat

L'IFRIS (Institut Francilien Recherche Innovation Société) est, depuis 2007, un consortium d'Unités de Recherche en Ile-de-France qui travaillent sur les questions liées aux interactions entre science, techniques et sociétés ainsi que politiques de recherche et d'innovation.
L'IFRIS a développé une plateforme numérique d'analyse textuelle fondée sur des outils d'analyse novateurs (Cortext).
L'IFRIS était associé au projet 2eFDI en vue d'ouvrir, en partenariat avec l'IRD, des chantiers d'importance non seulement de valorisation classique des résultats de recherche, mais aussi d'analyses sémantiques, bibliométriques et de visualisation qui peuvent avoir une grande portée pour le futur, en permettant de construire des projets dans l'ensemble des domaines disciplinaires sur la constitution des objets de recherche et des régimes de production des savoirs.

[13] Le facteur D (downloading factor) est un indicateur que nous avons mis au point en utilisant une méthode de calcul comparable à celle du facteur H (http://fr.wikipedia.org/wiki/Indice_h).

Le calendrier du projet a été très fortement perturbé par la crise sanitaire et par les difficultés d'organisation de notre sous-traitant en matière de numérisation. Nos efforts se sont donc concentrés sur la réalisation du projet en ce qui concerne la numérisation des documents du fonds documentaire de l'IRD (préparation des documents, suivi de la numérisation, finalisation des documents post-numérisation).
Ainsi la collaboration avec l'IFRIS n'a pas pu se concrétiser sur la période du déroulement du projet. Nous envisageons néanmoins de poursuivre l'exploration de cette collaboration tout en tenant compte des conditions sanitaires dont l'évolution est particulièrement incertaine.

3 Caractéristiques techniques de la numérisation

Les documents du Fonds documentaire de l'IRD sont entreposés sur le site IRD de Bondy et constituent un fonds documentaire physique avec des documents présents en double exemplaire ou en exemplaire unique en format papier. Pour la réalisation du projet 2eFDI, les documents à numériser ont été extraits à partir des rayonnages du FDI, ont été mis en caisses et transportés chez le sous-traitant.
La numérisation des documents a été réalisée avec des scanners de production format A4/A3 après la suppression des reliures (légères ou rigides). Les cartes et les pages hors format (dont la taille dépasse le format A3) ont été numérisées sur des scanners de plan et associées à chaque document dans des répertoires spécifiques. Le processus de numérisation massif a permis de produire directement des fichiers pdf avec un paramétrage de la numérisation à 300 dpi et une colorimétrie adaptée à la nature des pages (numérisation en noir et blanc, en niveaux de gris, en couleur en fonction du contenu visible de la page). Les pages hors format ont été livrées au format tiff avec une résolution de 300 dpi. Ces pages ont été traitées par l'équipe projet pour en améliorer la qualité (traitements sous Photoshop) puis ont été intégrées aux documents originaux au format pdf.
Le sous-traitant a appliqué aux fichiers pdf produits un traitement de reconnaissance optique de caractères de façon systématique et industrielle sans intervention manuelle sur les résultats.
Le sous-traitant a effectué un premier contrôle-qualité pour vérifier que les résultats de la numérisation sont bien une représentation informatique intégrale de l'original papier.
Le contrôle-qualité final, réalisé par le technicien de la numérisation, a permis de vérifier l'intégrité des documents (pages manquantes, pages en double, page mal positionnées, pages mal orientées) ; cette étape a permis également la suppression de tampons, écritures, étiquettes présents sur les couvertures et les pages de garde ; l'assemblage et le désassemblage des documents (chapitres ou parties d'ouvrages). Ces opérations ont été réalisées grâce à l'utilisation d'une version professionnelle d'Adobe Acrobat, et avec Adobe Photoshop.
La finalisation des documents a permis l'injection des métadonnées de la base bibliographique dans les fichiers pdf, l'optimisation web de ces fichiers[14], le contrôle de la bonne identification des fichiers, leurs associations avec les fiches bibliographiques. Ces opérations ont été réalisées par le chef de projet en utilisant une version professionnelle d'Adobe Acrobat qui intègre des « plugins » spécifiques. L'extraction des métadonnées bibliographiques (pour injection dans les fichiers pdf) et l'association des fichiers aux fiches documentaires ont été réalisées avec des scripts en php qui interagissent avec la base de données bibliographiques Horizon.
Les fichiers pdf ont été transférés dans un espace informatique qui leur est dédié. Le système des fichiers sitemaps a été mis à jour à chaque opération de transfert des fichiers pdf pour que les robots d'indexation de l'internet puissent aisément donner accès à ces ressources pour les utilisateurs de l'Internet.

[14] Optimiser la taille des fichiers pdf pour faciliter l'accès internet, est une opération essentielle, surtout vis-à-vis des internautes qui disposent d'une bande passante réduite. "Linear PDF files (also called "optimized" or "web optimized" PDF files) are constructed in a manner that enables them to be read in a Web browser plugin without waiting for the entire file to download, since they are written to disk in a linear (as in page order) fashion". Source : http://en.wikipedia.org/wiki/Portable_Document_Format

Les fichiers de logs du serveur des fichiers pdf en libre accès sont régulièrement exploités pour suivre la montée en charge et les niveaux des consultations de l'indexation des moteurs de recherche et de l'usage qui en est fait par les internautes.
Il est important de constater que sur les trois millions de « accès-utilisateurs » des fichiers pdf de Horizon-Pleins_Textes en 2020, 90 % des accès passent par une requête effectue dans « google » ou « google scholar » et que pour nos collections, 70 % des consultations relèvent d'un pays des Suds.
Les sauvegardes du système d'information documentaire de l'IRD (base de données bibliographiques Horizon, fichiers pdf en libre accès et en accès réservé, données de consultations sont effectuées par la Direction pour le Développement des Usages Numériques Innovants (D-DUNI)[15]. Des copies supplémentaires de ce système d'information documentaire sont effectuées par le service de l'information scientifique et technique de l'IRD.

Le processus de production de la numérisation du projet 2eFDI peut se résumer en quatre étapes :
La préparation des documents a permis l'identification et l'extraction des documents du FDI à numériser. Ces opérations ont permis l'extraction physique des documents du fonds à partir des listes préparées par le chef de projet. Chaque document a été identifié par un code barre qui a servi au nommage du fichier pdf produit lors de la numérisation. Cet identifiant est unique et correspond à l'identifiant du document dans la base bibliographique Horizon.
La numérisation des documents a été réalisée par le sous-traitant selon le protocole de numérisation et les spécifications définies par le chef de projet IRD (ces procédures ont été mises au point et bien rodées au cours des opérations précédemment menées par le sous-traitant dans le cadre du marché public passé par l'IRD).
Le contrôle-qualité a été opéré à l'IRD par le technicien de la numérisation. Cette procédure comportait le découpage des documents constitués de plusieurs communications/chapitres, le nettoyage des couvertures et des pages de garde des documents (suppression des tampons et des identifiants et des « tags » présents), la vérification de l'intégrité et de la cohérence du document (pages manquantes, pages en trop, assemblage des cartes hors textes et des annexes).
La diffusion des fichiers pdf a nécessité une phase de finalisation des documents avec l'intégration des métadonnées, la vérification du bon nommage du fichier, l'optimisation web des fichiers, leur mise en accès sur les dossiers du serveur, leur association avec les fiches bibliographiques.

Fiche numérisation

La numérisation des documents a été réalisée avec des scanners de production format A4/A3 après la suppression des reliures (légère ou rigides). Les cartes et les pages hors format (dont la taille dépasse le format A3) ont été numérisées sur des scanners de plan et associées à chaque document dans des répertoires spécifiques. Le processus de numérisation massif a permis de produire directement des fichiers pdf avec un paramétrage de la numérisation à 300 dpi et une colorimétrie adaptée à la nature des pages (numérisation en noir et blanc, en niveaux de gris, en couleur en fonction du contenu visible de la page). Le sous-traitant a effectué un premier contrôle-qualité pour vérifier que les résultats de la numérisation constituaient bien une représentation informatique intégrale de l'original papier. Le contrôle qualité réalisé par le technicien de la numérisation doit a permis de vérifier l'intégrité des documents (pages manquantes, pages en double, page mal positionnées, pages mal orientées) ; s'y ajoutent la suppression de tampons, écritures, étiquettes présents sur les couvertures et les pages de garde ainsi que l'assemblage et le désassemblage des documents. La finalisation des documents a été réalisée avec l'injection des métadonnées de la base bibliographique dans les fichiers pdf, l'optimisation web de ces fichiers, le contrôle de la bonne identification des fichiers, leurs associations avec les fiches bibliographiques.
VOLUMETRIE – Environ 862 500 pages pour un ensemble de 10 432 documents (moyenne de 83

[15] Les sauvegardes sont effectuées par la D-DUNI, sous la forme d'une prestation de service réalisée par la société Cloud Temple (https://www.cloud-temple.com/).

pages par document). Le coût unitaire moyen de numérisation des documents (préparation, numérisation, reconnaissance optique des caractères, validations) est de l'ordre de 6,6 €.
Format d'acquisition des images Les fichiers pdf contiennent différents type d'images produites lors de la numérisation des documents (résolution : 300 dpi). D'une façon générale, les pages en noir et blanc sont des images au format tiff ; les pages en niveaux de gris ou en couleurs sont des images au format JPEG 2000.
Résolution : 300 dpi
Prestataire : ☑ prestataire privé : Studia solution selon le marché IRD 2017-2021 n° 20170030CTA004
Métadonnées des fichiers pdf : les fichiers pdf disposent de métadonnées techniques (taille du fichier, nombre de vues, date de création, outil de création, …) et des métadonnées bibliographiques (injection à partir des métadonnées de la base bibliographiques HORIZON). Les métadonnées des fichiers pdf sont disponibles au format DC, XMP, PDFx, …

3.1 Production de métadonnées, structuration

Le projet 2eFDI ne prévoyait pas de production de métadonnées ou de structuration : tous les documents du fonds documentaire de l'IRD sont répertoriés, identifiés, catalogués et indexés. Toutes les métadonnées concernant ces traitements sont saisies dans la base de données bibliographiques HORIZON (crée en 1986 et mise à jour de « façon capillaire » : plusieurs interventions journalières).
Pour l'indexation des documents le service d'information scientifique et technique de l'IRD utilise un vocabulaire multidisciplinaire mis au point depuis 1989 : il comporte 4391 entrées structurées de façon arborescente. Pour la catégorisation thématique un plan de classement a été conçu par les documentalistes de l'IRD dès la création de la base de données bibliographique HORIZON : il comporte 42 entrées thématiques structurées de façon arborescente.
Tout le système d'information documentaire (notices bibliographiques et fichiers pdf) est moissonnable au format OAI-PMH, html, xml-mods, et indexé de façon capillaire par les robots de recherche de l'internet.
Tous les documents qui ont été numérisés dans le cadre du projet 2eFDI disposaient déjà d'un ensemble complet de métadonnées. Dans le cadre du projet 2eFDI aucune nouvelle métadonnée bibliographique n'a été produite. Par contre de nouvelles métadonnées techniques ont été générées (de façon automatique ou logique) par la réalisation d'un fichier pdf : la taille du fichier, le nombre de vues (« pages ») du fichier, la date de fabrication du fichier, l'url du fichier, le texte intégral produit par reconnaissance optique de caractères.
Les métadonnées incluses dans les fichiers pdf sont au format DC, XMP, PDFx et peuvent être extraites avec plusieurs outils gratuits.
L'ensemble des documents présents dans Horizon-Pleins_Textes respectent les règles de la *Charte des bonnes pratiques pour l'édition numérique scientifique*[16] applicables pour ce genre de collection, et notamment :

- Citabilité : chaque document est identifié par une URL stable et courte, constituée comme une URI établie à partir de l'identifiant pérenne du document dans la base Horizon.
- Interopérabilité : chaque document, à partir de son identifiant unique, est accessible de manière standardisée pour sa forme PDF et pour ses métadonnées aux formats génériques Dublin Core, BibTex, Mods et dans des formats adaptés à EndNote et à Zotero, en particulier.
- Accessibilité : les documents numériques sont au format PDF et les pages HTML présentant les documents et leurs métadonnées en XHTML respectant les recommandations d'accessibilité du W3C.
- Ouverture : les documents sont en libre accès, sans DRM et sans quota d'accès.
- Durabilité : des solutions d'archivage pérenne sont à l'étude pour l'ensemble des données de

[16] Voir : https://publications-prairial.fr/arabesques/index.php?id=1110

l'IRD, des contacts ont été pris auprès du Cines ; de plus, l'IRD contribue à l'archive ouverte Hal via son portail Hal-IRD.

3.2 Diffusion sur Internet

Tous les fichiers PDF produits à partir de la numérisation des documents concernés par le projet sont mis en libre accès sur le portail documentaire de l'IRD https://horizon.documentation.ird.fr/, chaque document étant associé à sa notice bibliographique de la base Horizon.
Sur ce site la recherche se fait avec un formulaire qui comporte plusieurs champs et des listes de classements thématiques.
La taille des fichiers ainsi que le contenu des fichiers PDF sont optimisés (voir note n° 14) pour faciliter les transferts via le protocole HTTP.
Tous les fichiers pdf sont enrichis avec des métadonnées internes (auteurs, titres, classement thématique, mots clef) extraites de la base de données bibliographique Horizon.
Afin de vérifier le niveau de valorisation des fichiers pdf en libre accès, nous effectuons depuis plus de 10 ans des analyses statistiques des indexations par les moteurs de recherche, des consultations pouvant être considérées des « accès-utilisateurs », des spams.
Grace au système des fichiers sitemap, tous les fichiers pdf sont régulièrement indexés par le robot google (googlebot) et par de nombreux autres robots de recherche ([…]bot).
Les analyses statistiques de consultations, produites avec des modules de filtrages des moteurs de recherche et des spams, nous indiquent qu'en 2020 64 103 fichiers ont été consultés au moins une fois, avec un total d'environ trois millions de consultations (moyenne de 45 consultations par fichier).
Les statistiques produites nous montrent que 90 % des consultations relèvent d'une consultation venant d'une page google (referer google) et que environ 50 % des accès ont comme origine les pays d'Afrique. Environ 70 % des accès relèvent des pays en développement.
Les métadonnées de l'ensemble des documents numérisés sont exposées sur Internet dans différents formats et moissonnables notamment en OAI-PMH, comme c'est le cas pour l'ensemble de la base bibliographique Horizon de l'IRD.
L'entrepôt de toutes les métadonnées, quel que soit le statut du document est :
http://www.documentation.ird.fr/fdi/oai.php?verb=Identify
Cet entrepôt est moissonné par :
Open Aire : https://www.openaire.eu/
Bases : https://www.base-search.net/
Isidore : https://www.rechercheisidore.fr/
Clacso (Consejo Latino Americano de Ciencias Sociales) : http://biblioteca.clacso.edu.ar/
Bneuf (Bibliothèque Virtuelle de l'Espace Numérique Francophone): https://bneuf.auf.org/

4 Pérennité du projet

4.1 Pérennité de la diffusion des corpus numérisés

L'IRD assure l'hébergement du fonds numérisé et des métadonnées dans son système documentaire : https://horizon.documentation.ird.fr.
Le Service d'Information Scientifique et Technique (Mission Culture Scientifique et Technique de l'IRD) assure la gestion du Fonds documentaire de l'IRD, la gestion, la production et les développements du système d'information documentaire : la base de données bibliographiques Horizon et le système de fichiers pdf en libre accès Pleins_Textes.
Le service d'accès au système d'information documentaire a été et est assuré de façon continue (24x24 7x7) depuis la mise en place du projet Pleins_Textes en 1996 (premiers documents de l'IRD en accès en texte intégral).

Auparavant, quand pour le Fonds documentaire de l'IRD il n'existait qu'une base de données bibliographiques (entre 1986 et 1996) cette base était accessible en ligne via le protocole X25 (accès minitel 3617 mode écran 80 colonnes ou connexion en mode terminal sur les serveurs) via les serveurs du site IRD de Montpellier[17].

4.2 Sauvegarde des corpus numérisés

Les sauvegardes du système d'information documentaire de l'IRD (base de données bibliographiques Horizon, fichiers pdf en libre accès et en accès réservé, données de consultations sont effectuées par la Direction pour le Développement des Usages Numériques Innovants (D-DUNI) de l'IRD. Des copies supplémentaires de ce système d'information documentaire sont effectuées par le Service de l'information scientifique et technique de l'IRD.
La D-DUNI assure le fonctionnement du système d'information de l'IRD. Actuellement la D-DUNI a passé des marchés en ce qui concerne l'hébergement et les sauvegardes du système d'information de l'IRD à la société Cloud Temple (https://www.cloud-temple.com/). Les sauvegardes complètes des disques du serveur se réalisent à chaud via un snapshot vmWare.
Dans le dispositif mis en place par la D-DUNI de l'IRD, le système d'information documentaire fait partie des applications institutionnelles prioritaires et inscrites dans le schéma directeur informatique de l'IRD. Les applications institutionnelles prioritaires sont gérées selon les mêmes garanties contractuelles que les applications de gestion administrative de l'IRD.
Des solutions d'archivage pérenne sont à l'étude pour l'ensemble des données de l'IRD, des contacts ont été pris auprès du Cines par la D-DUNI ; de plus, l'IRD contribue à l'archive ouverte Hal via son portail Hal-IRD.

[17] ROSSI Pier Luigi (1992). Servers and online bibliographic databases in developing countries : the African reality. In : Raitt D.I. (ed.). Online information 92. Oxford : Learned Information, 1992, p. 431-435. http://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_6/b_fdi_35-36/41308.pdf