Stéphanie LÉVÊQUE

Master DEFI

RAPPORT de STAGE

[stage effectué du 30 mars au 30 juin 2015]

Centre International de Recherche sur l'Environnement et le Développement

Jardin Tropical - Nogent sur Marne

Archivage électronique :

numérisation et valorisation d'un fonds scientifique.

Sous la direction de : M. Franck Lecocq

M. Naceur Chaabane (tuteur professionnel)

Soutenu le 10/09/2015 à l'UFR Phillia

Université Paris Ouest Nanterre La Défense

200 Avenue de la République 92001 Nanterre cedex

Année Universitaire 2014 - 2015

Remerciements

Je tiens à remercier dans un premier temps toute l'équipe pédagogique du Master DEFI de l'Université de Paris 10 ainsi que tous les intervenants professionnels.

Je tiens à remercier tous les membres du laboratoire, le centre international de recherche sur l'environnement et le développement (CIRED), plus particulièrement Monsieur Franck Lecocq, directeur de l'unité et Messieurs Naceur Chaabane et Franck Nadaud pour leur accueil et leur disponibilité. Je remercie également toute l'équipe documentaire de l'École Nationale des Ponts et Chaussées (ENPC).

Je remercie plus personnellement toute ma famille de m'avoir encouragé et soutenu durant cette année.

A Stéphane, à mes enfants.

Résumé :

L'archivage électronique de documents scientifiques répond à la demande croissante de chercheurs d'avoir accès de façon pérenne à toute une documentation primordiale pour leurs travaux. La dématérialisation des documents permet de nouvelles pratiques de recherches. Ces archives ainsi constituées sont valorisées et diffusées sur des bases ressources. Elles accompagnent les jeunes chercheurs dans leurs travaux. La conservation de la mémoire scientifique du laboratoire est assurée.

Mots-clés :Conservation, archivage électronique, numérisation, médiation scientifique

Abstract :

The electronic archiving of scientific documents responds to a growing demand of researchers for having permanent access to a whole literature, essential for their work. The changing to a paperless system allows new research practices. The archives constituted this way are highly valued and they circulate in the databases. They escort the young researchers in their work. So, the conservation of the scientific memory of the laboratory is ensured.

Keywords :.Conservation, electronic storage, digitizing, scientific mediation

Droits d’auteurs

Table des matières

Remerciements.....1

Résumé.....2

Liste des abréviations.....6

Introduction.....7

1 – **Présentation des structures d'accueil**.....9

1.1 – Le CIRED – laboratoire de recherche.....9

1.2 – L'ENPC et le projet R2DS.....11

2 – **Présentation de la mission**.....15

2.1 – Le projet R2DS et l'archivage électronique.....15

2.1.1 – L'existant.....16

2.1.2 - Les besoins.....18

2.1.3 - L'apport du stage.....19

2.2 – Les données exploitées.....20

2.2.1 – L'existant.....20

2.2.2 - L'exploitation des données et les limites rencontrées.....23

3 – **Le travail réalisé**.....27

3.1 – L'organisation du projet.....27

3.1.1 – La gestion du projet.....28

3.1.2 – Les tâches.....29

3.1.3 – Le planning.....31

3.2 – L'interaction avec l'équipe.....32

3.3 – La nécessité des choix....33
3.3.1 – Des choix techniques....33
3.3.2 – Des choix méthodologiques....35
3.4 – La mise en œuvre du projet....35
3.4.1 – L'exploitation de plusieurs sources....36
3.4.2 – Alimentation de la base ressource....40
3.4.3 – Exploitation et valorisation de la base ressource....42
3.5 – Les résultats du processus....43
3.5.1 – Les pistes de réflexion....43

Conclusion....44

Bibliographie....45

Annexes....46

Liste des abréviations

AERES : Agence d'évaluation de la recherche et de l'enseignement supérieur.

AgroParisTech : Institut des sciences et industries du vivant et de l'environnement.

BIOEMCO : laboratoire de Biogéochimie et écologie des milieux continentaux.

CINES : Centre informatique national de l'Enseignement supérieur.

CIRAD : Centre de coopération internationale en recherche agronomique pour le développement.

CIRED : Centre international de recherche sur l'environnement et le développement.

CNRS : Centre national de la recherche scientifique.

DIM : Domaine d'intérêt majeur.

EHESS : École des hautes études en sciences sociales.

ENPC : Ecole nationale des Ponts et Chaussées.

GIEC : Groupe d'experts intergouvernemental sur l'évolution du climat.

GIS : Groupement d'intérêt scientifique.

IFSTTAR : Institut français des sciences et technologies des transports, de l'aménagement et des réseaux.

INREST : Institut nordique de recherche en environnement et en santé du travail.

IPSL : Institut Pierre Simon Laplace.

MEDDE : Ministère de l'écologie, du développement durable et de l'énergie.

R2DS : Réseau francilien de recherche sur le développement soutenable.

Introduction

Mon stage de fin d'étude Master2 DEFI s'est déroulé sur la période du 30 mars au 30 juin 2015 dans un organisme d'enseignement supérieur. J'ai eu l'opportunité de mettre en place l'archivage électronique d'un fonds « chercheurs » pour constituer une base ressource dans le domaine de l'économie de l'environnement. J'ai été amenée à travailler dans deux lieux distants : le laboratoire de recherche du Centre International de Recherche en Économie du Développement (CIRED) et l'École Nationale des Ponts et Chaussées (ENPC).

L'objectif principal du stage consiste en la numérisation et la valorisation d'un fonds d'archives scientifiques issues du CIRED.

Le CIRED est un laboratoire qui aborde les questions de développement durable.

Cet archivage électronique permet à terme aux chercheurs de disposer d'une « mémoire documentaire » sur les sujets majeurs du laboratoire.

Un premier fonds avait été identifié avant le début du stage. J'ai eu à analyser ce fonds. Pour une meilleure valorisation, j'ai proposé la mise en place de mots-clés (tags) pour proposer une catégorisation du fonds. Cela m'a semblé pertinent au vu de la quantité de documents disponibles au final sur la plate-forme Zotero.

Zotero, outil collaboratif sélectionné pour la réalisation de la mission est une base de données de gestion bibliographique. Cette plate-forme est un choix particulièrement adapté pour la constitution de la base ressource. Elle offre la possibilité de constituer des références structurées, interrogeables sur certains champs (titre, auteur, année...) et sur le texte intégral. Elle permet aussi un travail individuel et en groupe. Enfin, elle offre la possibilité de faire participer les chercheurs dans la constitution de la base, par l'ajout des tags. Les références ainsi disponibles sur Zotero sont interopérables.

Tout d'abord, je vais présenter l'organisme de recherche, le CIRED qui souhaite archiver et valoriser son fonds chercheurs ainsi que le projet R2DS soutenu techniquement par l'ENPC.

Dans une seconde partie, j'aborderai la mission qui m'a été confiée, le contexte, l'objectif et les moyens utilisés.

Enfin dans une dernière partie, je vous présenterai le travail réalisé, la mise en œuvre du projet et les résultats obtenus.

1 - Présentation des structures d'accueil

1.1 - Le CIRED, laboratoire de recherche

Le CIRED, laboratoire de recherche en économie de l'environnement a deux tutelles (le CNRS et l'ENPC) et trois institutions partenaires (l'EHESS, AgroParisTech et le CIRAD).

Le CNRS organisme de recherche apporte son soutien au laboratoire en attribuant des postes d'ingénieurs administratifs et techniques. L'École des Ponts et Chaussés, Haute École en génie civil intervient quant à elle plus particulièrement sur les questions documentaires. Elle apporte son soutien technique au développement de l'archivage électronique des fonds scientifiques du laboratoire.

La documentation scientifique est localisée dans les locaux du CIRED et la plate-forme de numérisation est elle dans les locaux de l'ENPC.

Le CIRED a été fondé en 1973 par Ignacy Sachs, chercheur travaillant sur les questions de développement durable, suite à la Conférence Internationale de Stockholm sur l'Environnement et le Développement. Ignacy Sachs souhaite faire émerger les questions de « développement durable ». Dès la création du laboratoire, les chercheurs travaillent sur l'écodéveloppement, sur des domaines liés à l'énergie, aux déchets, à l'économie informelle… Les travaux du CIRED s'intéressent depuis le début à la notion de développement durable en adoptant une démarche prospective. Ces travaux sont réalisés en faisant travailler ensemble plusieurs disciplines (sciences sociales, sciences de l'univers, sciences de la vie…). Le CIRED porte ainsi, depuis le début, **l'interdisciplinarité**. Il continue de le montrer aujourd'hui par l'organisation de ses équipes de recherche[1].

Le CIRED en quelques dates :

Ce laboratoire de recherche, fondé par l'EHESS en 1973, a été reconnu par le CNRS en 1978. Trois autres tutelles interviennent dans l'organisation du CIRED depuis les années 1990 : l'École Nationale du Génie Rural, des Eaux et des Forêts (aujourd'hui

[1] A1 – Organigramme du CIRED.

AgroParisTech), l'École Nationale des Ponts et Chaussées (ENPC) et le Centre International de la Recherche Agronomique pour le Développement (CIRAD).

Aujourd'hui, seuls le CNRS et l'ENPC sont tutelles du CIRED, les trois autres organismes (EHESS, AgroParisTech et le CIRAD) sont partenaires. Le CIRED accueille dans ses locaux du Jardin Tropical de la Ville de Paris environ 90 personnes.

Dans sa politique scientifique, le CIRED travaille sur quatre axes de recherche :

-**Prospectives sectorielles : énergie, ville, usage des terres, eau urbaine.** Ici sont abordés toutes les questions en lien avec les stratégies de développement sous contraintes sociales et environnementales.

-**Stratégies de développement sous contrainte climatique, environnementale et sociale**. Ici sont abordées les questions de prospectives liées aux relations entre environnement et développement.

-**Incertitudes, controverses, négociations, décisions**. Ici est examinée la façon dont l'économie, les sciences sociales peuvent aider à la mise en place de processus de décision dans un contexte d'incertitudes et de controverses.

-**Modèles, outils et bases de données**. Il s'agit ici de concevoir et maintenir des modèles, des outils d'analyse des données pour mener à bien les travaux scientifiques des autres axes.

Parmi les faits marquants du laboratoire, on peut souligner la participation de cinq chercheurs comme *lead authors* aux rapports du GIEC (Groupe d'experts intergouvernemental sur l'évolution du climat), la reconnaissance du modèle Imaclim-R comme modèle de référence internationale et l'analyse de l'impact de l'EU *Emissions Trading Scheme* sur la compétitivité industrielle.

Ces différents événements ne peuvent que conforter la place importante qu'occupe le CIRED sur le plan international.

Cette émulsion scientifique se retrouve aussi dans les travaux publiés par le laboratoire. Sur la période 2008-2013, le CIRED a produit 801 items, dont 304 de rang A (au sens AERES). Parmi les disciplines choisies pour la diffusion de la production scientifique du laboratoire, l'économie arrive en tête avec 45 %, puis les sciences sociales autres que l'économie avec 15 %. Par ailleurs 40 % des articles sont publiés dans des revues interdisciplinaires.

Enfin, le CIRED a eu l'occasion de participer à des réseaux de recherche internationaux et nationaux. On peut notamment citer ici sa participation au GIS R2DS. Ce domaine d'intérêt majeur (DIM) a permis au laboratoire de tisser des liens forts avec notamment le BIOEMCO, laboratoire en sciences de la terre et l'IPSL, laboratoire en sciences du climat.

1.2 - L'ENPC et le projet R2DS

L'École Nationale des Ponts et Chaussées a été créée en 1747 par Daniel-Charles Trudaine. Cette Grande École d'ingénieurs est placée sous la tutelle du Ministère de l'écologie, du développement durable et de l'environnement. Cette École, se définissant comme généraliste, forme des ingénieurs, des étudiants en master, en MBA (*master of business administration*) et de futurs docteurs. L'École des Ponts est fortement tournée vers l'international avec plus de 40 % des effectifs qui obtiennent un double diplôme à l'international et 30 % d'une promotion de cycle d'ingénieur vient de l'étranger.

L'École des Ponts en chiffres : 350 enseignants chercheurs, plus de 1200 étudiants dont 560 élèves-ingénieurs et plus de 180 diplômes délivrés.

L'histoire des Ponts est marquée par quelques grandes dates :

A sa création en 1747, Jean-Rodolphe Perronet souhaite développer un enseignement fondé sur l'apprentissage et le tutorat. Il reste à la tête de l'École pendant 47 ans puis à la Révolution le mode de recrutement devient « national », un concours d'entrée est ouvert et les futurs élèves reçoivent un salaire fixe pendant leur formation.

En 1796, Jacques-Élie Lamblardie propose la création des deux premières chaires de l'École. Elles sont respectivement consacrées aux Sciences Appliquées et aux Constructions.

1831 marque la création du premier laboratoire de recherche en génie civil dans le monde. Ce laboratoire devient en 1949, laboratoire central des Ponts et Chaussées. En 2011, il prend le nom d'IFSSTAR lors de sa fusion avec l'INREST.

Plus près de nous, c'est en 1993 que l'École est rattachée au Ministère de l'écologie, du développement durable et de l'énergie (MEDDE). Sous la tutelle de ce ministère, elle a pour vocation de former de futurs ingénieurs possédant des compétences scientifiques, techniques et générales pour exercer des fonctions dans les métiers de l'équipement, de l'aménagement, de l'environnement... En 2007, l'École est membre fondateur de deux PRES : l'Université Paris-Est et d'un réseau thématique de recherche avancée.

Dans la valorisation de la recherche et dans le développement de relations industrielles, l'École des Ponts poursuit ses actions vers les organismes publics et les entreprises publiques et privées. L'École souhaite ainsi mettre ses formations et sa recherche au service de la compétitivité des entreprises. Les laboratoires rattachés à l'École ont signé des partenariats avec des entreprises publiques (EDF, …). L'École participe au pôle de compétitivité « AdvanCity – Ville et mobilité durable » et marque ainsi son engagement pour développer une recherche plus collaborative et au service des entreprises. Enfin plusieurs chaires marquent la relation très forte entre l'École et de grandes entreprises privées : la Chaire « Hydrologie pour une ville résiliente » avec Veolia, la Chaire « Sciences pour le Transport Ferroviaire » avec Euro-tunnel...

En 2013, l'École a fait évoluer son offre de formations pour développer une pédagogie par projet. De nombreux intervenants extérieurs interviennent dans l'enseignement. L'École offre ainsi de réelles passerelles entre la formation théorique et le monde des entreprises (SNCF, Vinci, Suez Environnement, EDF...).

Dans le développement de sa stratégie scientifique, l'École des Ponts a donné également une grande place au Développement durable. Parmi les seize Chaires de l'École, huit traitent de cette question. L'École est également un membre actif de l'Institut de la Mobilité Durable (création conjointe de ParisTech et Renault).

Sur ces questions, l'École rejoint les préoccupations scientifiques du CIRED. Elle soutient d'ailleurs ce laboratoire ainsi que le laboratoire de Météorologie Dynamique (LMD) dans leurs participations au groupe Intergouvernemental d'experts sur le climat (GIEC).

Dans son organisation administrative, l'École des Ponts est organisée en neuf directions, un secrétariat général et en différents départements d'enseignement[2]. Onze laboratoires sont rattachés à l'École et parmi eux, le CIRED.

Dans la mise en place de mon stage, j'ai eu à travailler avec la **Direction de la documentation, des archives et du patrimoine** qui apporte son soutien technique au projet R2DS. Cette direction pilote deux développements (la gestion de la documentation, le patrimoine historique) mais aussi la gestion des archives de l'École. Elle s'organise en trois pôles et elle soutient une mission archives. La direction de la documentation en quelques chiffres : 200 000 documents imprimés et multimédias, 11 500 dossiers d'archives, 15 000 dépôts de publications scientifiques sur Hyper Articles en ligne (HAL), 700 000 pages numérisées issues des collections patrimoniales.

Les grandes missions de la Direction de la documentation consistent :

- Amélioration des services au public
- Contribution à l'animation sur le campus
- Offre de ressources pédagogiques numériques
- Formation des élèves à la maîtrise de l'information
- Enrichissement des catalogues et la valorisation des collections.

La Direction de la documentation a lancé deux projets d'envergure : la plate-forme bibliométrique en partenariat avec 5 autres établissements d'enseignement supérieur ainsi que l'archivage électronique des archives scientifiques de laboratoires de recherche dans le domaine **du développement durable**. Le CIRED a été naturellement sollicité pour participer à ce dernier projet. C'est dans ce cadre que j'ai pu réaliser mon stage de fin d'études.

Ce projet de dématérialisation de la documentation scientifique de laboratoires de recherche a pu se développer avec le soutien de la Région Île de France au travers du **programme R2DS** (Réseau de recherche sur le développement soutenable).

Ce programme de recherche soutient depuis 2006 des projets scientifiques qui abordent les questions de développement soutenable, les relations entre les activités

[2] A2 - Organigramme de l'ENPC.

humaines et leur environnement naturel. Au travers de ces actions, il ambitionne de devenir un acteur incontournable de la production de connaissances.

Ce réseau de recherche est constitué de dix-neufs partenaires (Instituts, Universités, Grandes Écoles). Quatre grandes thématiques de recherche structurent le projet ainsi que de nombreuses manifestations scientifiques (ex : colloque « RIO 2012 : Sustainable Development Out of the Age of Innocence » ; colloque « Métropolisation, cohésion et performances : Quels futurs pour nos territoires ? »...).

Ce projet de recherche souhaite avant tout pouvoir porter différentes missions et augmenter ainsi la capacité d'intervention de la recherche francilienne à l'échelle européenne. Il met également tout en œuvre pour implanter les thématiques de recherche dans l'environnement local, pour favoriser les échanges entre la communauté scientifique et le public et jouer ainsi le rôle de médiateur scientifique en soutenant l'émergence de cercles d'échange (scientifiques, experts et citoyens).

Enfin, ce réseau ne se définit pas comme un club fermé. Bien au contraire, il travaille en relation avec un autre Groupement d'Intérêt Scientifique (GIS) : Climat – Environnement et avec trois labex qui traitent des mêmes thématiques : le labex « Institut Pierre et Simon Laplace », le labex « Futurs urbains », le labex « Ouvrir la science économique ».

2 - Présentation de la mission

Deux missions principales :

- Archivage de fonds chercheurs (conservation de documents qui constituent la mémoire du laboratoire) ; numérisation ; dépôt sur le serveur ENPC avec le numéro d'inventaire.

- Diffusion et médiation de ce fonds documentaire auprès de la Communauté des chercheurs et jeunes chercheurs du laboratoire ; constitution d'une base ressource qui serve de référence ; océrisation des documents électroniques, dépôt sur la plate-forme Zotero, lien vers le serveur CIRED ; cotation des documents, enrichissement avec des métadonnées, import et export des données.

Le but du stage est d'initier un processus de valorisation des fonds. Il s'agit de poser les bases et de réaliser une première tranche. J'ai été amenée à initier un processus qui servira de modèle pour d'autres actions de numérisation.

2.1 - Le projet R2DS et l'archivage électronique

R2DS Île-de-France est un réseau de recherche sur le développement soutenable. Créé en 2006 à l'initiative du Conseil régional d'Île-de-France, il aide la recherche sur les questions de développement soutenable.

La notion de développement soutenable reste encore aujourd'hui méconnue du grand public même si elle est diffusée depuis plus de trente ans, sous d'autres noms (écodéveloppement, développement durable...). Elle renvoie au souci de protéger les conditions de vie et d'épanouissement des générations futures. Elle mobilise des communautés scientifiques à l'échelle internationale issues de disciplines très diverses (sciences de l'univers, sciences pour l'ingénieur, sciences humaines et sociales...). Elle marque l'agenda politique international à travers de grandes conventions sur le climat, la biodiversité.

L'émergence du DIM R2DS doit permettre de soutenir la recherche francilienne sur ces questions. Le DIM ambitionne d'apporter une réelle originalité et de concentrer sur son territoire la production de connaissance sur le sujet.

Le DIM « développement soutenable » opère une distinction entre les dépenses de fonctionnement, les petits et moyens équipements (inférieur à 200 K€) et les équipements semi-lourds (de 200K€ à 5M€ d'investissement).

Dans le cadre de ce projet, il s'agit pour R2DS d'investissements de petits et moyens équipements. Le DIM R2DS apporte son soutien financier à des laboratoires de recherche qui souhaitent valoriser les publications scientifiques et les travaux de recherche.

L'outil mis en place et soutenu par R2DS doit permettre la capitalisation de la mémoire accumulée par les chercheurs sur les questions de développement soutenable, thématique phare du réseau. R2DS connaît toutes les difficultés dans la mise en place de cette mémoire, notamment l'hétérogénéité des travaux. En effet, ces questions scientifiques mobilisent des disciplines très diverses et qui n'ont pas les mêmes pratiques de publications. Le souhait de R2DS en soutenant ce projet d'archivage électronique est d'offrir un corpus facile d'accès et harmonisé. Un corpus dont la qualité est assurée par le processus même de sélection des documents (validation scientifique à chaque étape).

R2DS apporte ainsi son accord à l'acquisition d'une station de numérisation (ou scanner professionnel) permettant la numérisation d'un grand nombre de pages de documents de formats divers avec des fonctionnalités avancées de gestion de documents. Cette station de numérisation doit faciliter la conservation et la mise à disposition auprès des chercheurs du laboratoire CIRED et des laboratoires partenaires d'une mémoire des activités de recherche.

2.1.1 - L'existant

Le laboratoire de recherche CIRED travaille depuis plus de quarante années sur les questions de développement durable. Riche d'expériences scientifiques de renommées internationales, il souhaite à la fois pouvoir conserver sa documentation scientifique et la diffuser aux membres de son équipe. Le CIRED a identifié plusieurs fonds scientifiques qu'il évalue comme indispensable pour la mémoire du laboratoire. Parmi ces fonds, on peut citer

le fonds Sachs (fondateur du laboratoire), le fonds Godard, le fonds Hourcade et enfin le fonds général (sont repris ici les différents travaux scientifiques des chercheurs du laboratoire autres que les trois principaux cités plus haut).

Le fonds Sachs représente une documentation assez homogène. On y retrouve des articles fondateurs du laboratoire (année 1973), des travaux de recherche antérieurs au CIRED et une publication plus récente du chercheur mais toujours sur les mêmes problématiques liées au développement durable. Pour la documentation d'avant 1973, les documents sont de moins bonne qualité. Le travail sur le copybook a demandé une attention plus particulière avec notamment le choix des options de traitement (amélioration d'image, correction géométrique, accentuation des détails...).

Ignacy Sachs, fondateur et directeur du laboratoire pendant quatorze années est un professeur de socio-économie. C'est lui qui lors de la Conférence de Stockholm en 1972 met en garde contre une économie effrénée risquant à terme d'avoir un impact sur l'environnement. Il est professeur honoraire à l'EHESS.

Le fonds Godard représente une documentation riche de plusieurs volumes (supports papier et électronique). Les documents sélectionnés couvrent la période allant de la moitié des années 1970 aux années 2010. Olivier Godard est directeur de recherche au CNRS et professeur d'économie à l'École Polytechnique et à Sciences-Po Paris. Il est diplômé de l'ESSEC et docteur en sciences économiques.

C'est au cours de ses études à l'ESSEC qu'il a eu l'opportunité de suivre un séminaire d'Ignacy Sachs sur les stratégies de développement économique face à la protection de l'environnement. Après l'obtention de son diplôme, il intègre le CIRED. Il travaille dès lors sur les questions d''environnement et de développement durable, sujets qu'il aborde sous le prisme des sciences économiques et sociales. Olivier Godard rejoint en 1998 un laboratoire d'économétrie à l'École polytechnique. Il s'intéresse alors plus particulièrement au principe de précaution et à la décision en univers controversé.

Le fonds Hourcade se caractérise par une documentation fournie et dense avec des documents papiers et électroniques. Les documents ont connu une bonne conservation. Il n'a pas été apporté d'attention particulière lors de la manipulation du Copybook. Les documents référencés couvrent la période des années 1980 à nos jours.

Jean-Charles Hourcade a été directeur du CIRED de 1987 à 2012. Il est docteur en Sciences sociales (1977) et en sciences économiques (1984). Ses travaux de recherche portent sur les enjeux entre énergie et environnement. Chercheur en sciences sociales, il s'intéresse aux questions de changement climatique. Il a été membre du GIEC (Groupe d'Experts Intergouvernemental sur l'Évolution du Climat). Il a reçu à ce titre, le prix Nobel de la Paix en 2007, conjointement avec Monsieur Al Gore.

Enfin, **le fonds Général** se différencie avec une documentation plus récente, du début des années 1990 à nos jours. C'est sur ce fonds que l'on a le plus de documents électroniques natifs. Les documents sont de bonne qualité. Compte tenu de la nature des documents, il a été réalisé très peu de numérisation. Les versions électroniques disponibles ont servi à l'océrisation. On retrouve les quatre thèmes emblématiques du laboratoire dans ce fonds. Parmi les chercheurs référencés, on peut citer Dominique Finon, Daniel Théry, Louis Puiseux, Alexandre Nicolon...

Le laboratoire a estimé la totalité de ces fonds à environ 100 000 pages. Plusieurs types de documents caractérisent ce fonds scientifique. Il a pu être identifié des rapports techniques, des publications scientifiques, de la documentation interne au laboratoire, des créations iconographiques, cartographiques, photographiques, des formules mathématiques qui viennent appuyer des raisonnements, des mémoires de chercheurs...

L'ENPC a été sollicitée pour accompagner la démarche du CIRED dans l'archivage de son fonds scientifique. L'École a été membre actif dans le dossier de demande petit et moyen équipement pour l'année 2013 et elle accompagne toujours aujourd'hui le laboratoire.

2.1.2 - Les besoins

Le CIRED en identifiant sa documentation scientifique a également exprimé le besoin de constituer un corpus élargi. Il s'agit de constituer un recueil de documents disponibles sur une thématique donnée.

Le laboratoire a fait remarquer le manque aujourd'hui de pouvoir consulter un corpus reprenant les questions de développement durable. Il a donc sollicité un autre laboratoire de recherche, le LEESU (Laboratoire Eau Environnement et Systèmes Urbains) et un Labex « Futurs urbains » pour étoffer le corpus existant et proposer un ensemble intéressant pour la communauté scientifique. A terme, on pourra récupérer la mémoire accumulée par les chercheurs de ces différents laboratoires. La constitution de cette mémoire est aujourd'hui un enjeu important face aux défis de la prospective et de la durabilité.

Le CIRED souhaite donc sauvegarder et valoriser ses travaux de recherche mais aussi ouvrir des partenariats avec d'autres laboratoires préoccupés par les mêmes enjeux scientifiques.

Dans les missions qui m'ont été confiées pendant le stage, je n'ai pas eu à travailler sur la documentation des autres laboratoires (LEESU et Labex « Futurs urbains »). Mais il est important de souligner qu'à terme, la base bibliographique sous Zotero sera ouverte aux chercheurs et enseignants chercheurs de tous ces laboratoires.

Le CIRED a également exprimé le besoin de sauvegarder la mémoire scientifique constituée sur quarante années. Dans les cinq prochaines années, plusieurs chercheurs vont partir à la retraite et cette mémoire risque de disparaître. Elle doit être capitalisée pour garantir la continuité, la transmission entre les générations de chercheurs et de doctorants. Ce besoin exprimé est repris dans les deux missions principales du stage : l'archivage électronique pérenne du corpus constitué et la médiation scientifique auprès des membres de l'équipe du CIRED.

2.1.3 - L'apport du stage

Le stage doit concourir à la mise en place de l'archivage électronique et de la valorisation de la base ressource (base bibliographique Zotero). Le travail réalisé pendant le stage doit aussi permettre l'identification des différents corpus du CIRED, la sélection des documents, la validation scientifique et la proposition d'une taxonomie pour le corpus final.

Les processus élaborés pendant ce stage seront repris par la suite. Les critères de sélection, les paramètres de numérisation et d'océrisation sous LIMB, la constitution des numéros d'inventaire pour l'archivage sur le serveur de l'ENPC et l'établissement de

l'adresse URL pour la consultation sur le serveur du CIRED du texte intégral du document, tous ces éléments serviront de base pour la suite de l'archivage électronique des documents du CIRED.

Au delà du stage, le prototype mis en place avec la direction du CIRED pourra être ensuite étendu à un réseau plus large que les équipes de recherche déjà constituées actuellement (avec le LEESU et le Labex « Futurs urbains »).

2.2 - Les données exploitées

2.2.1 - L'existant

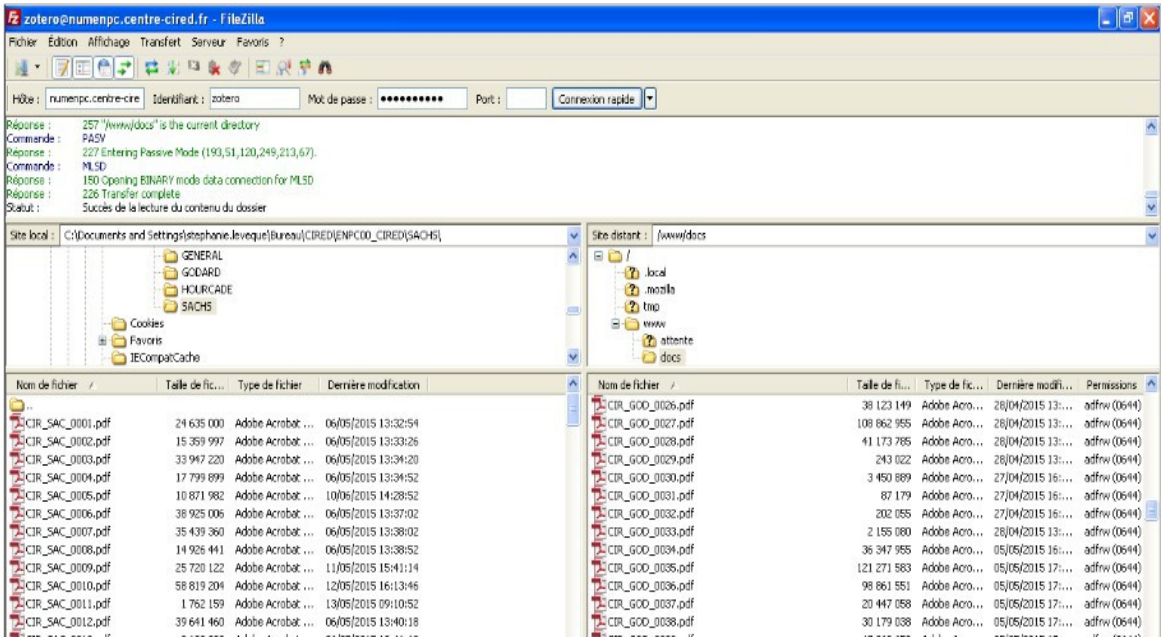
Avant le début de mon stage, une quarantaine de documents avaient été archivés et déposés sur la base bibliographique Zotero. Il y avait également des documents sélectionnés par le laboratoire et accessibles en version électronique sur le serveur du CIRED mais non exploités. L'ensemble de ces ressources électroniques représente plus d'une centaine de documents.

Enfin, il y avait aussi les documents papiers archivés par le laboratoire. Sur l'ensemble de ces documents papiers, un premier tri avait été fait par auteur mais il restait à appliquer les différentes étapes du processus d'archivage électronique.

Pour s'assurer du bon déroulé, il a été défini un espace de stockage supérieur à 20 GB et en accès restreint à la communauté scientifique concernée par le projet sur le serveur du CIRED. Un accès via Filezilla permet de se connecter au serveur à distance et de travailler sans trop de contrainte sur les données.

Filezilla est un outil simple d'installation (libre et gratuit) qui permet de se connecter à distance sur le serveur. Ce client FTP est disponible sur la plupart des systèmes d'exploitation. Pour pouvoir interroger le serveur web à distance, il est nécessaire de connaître plusieurs éléments : l'identifiant utilisateur FTP, le mot de passe FTP, l'adresse du serveur FTP (ordinateur distant), le répertoire de publication (www ou htdocs) et l'adresse web du site (https://www.zotero.org). Il est également indispensable d'avoir sur ordinateur local les fichiers à déposer sur le serveur FTP. Le chemin vers les fichiers doit être donné sur Filezilla.

Figure 1 – Vue du logiciel Filezilla, connexion au serveur CIRED.

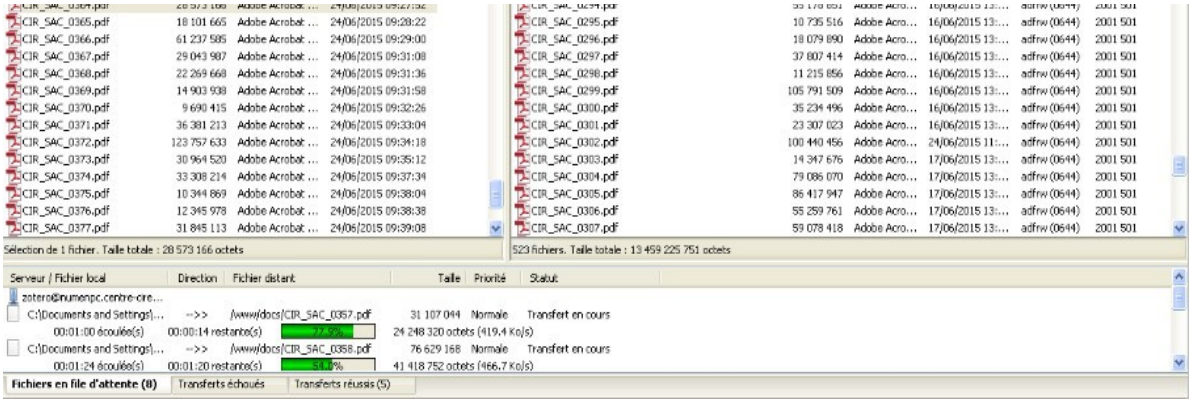


Après avoir renseigné les différents champs, une connexion rapide permet d'ouvrir une session distante sur le serveur du CIRED. Lorsque la connexion est établie, Filezilla affiche le contenu des ordinateurs connectés : l'ordinateur local et l'ordinateur distant. La navigation se fait au niveau des panneaux « site local » et « site distant ». Les chemins d'accès aux répertoires courants sont affichés sur ces deux panneaux. Les fichiers de chaque répertoire sont eux détaillés sur les deux autres panneaux du dessous.

Les transferts de fichiers se font habituellement de l'ordinateur local vers l'ordinateur distant. Il est nécessaire de choisir dans l'arborescence de l'ordinateur local le répertoire. Ensuite apparaît la liste des fichiers dans le panneau du dessous. Il devient alors très aisé de faire glisser le fichier sélectionné vers le répertoire souhaité dans l'ordinateur distant. Lorsque cette étape est lancée, un panneau de message de suivi apparaît. Il permet de suivre la progression du transfert.

Dans le travail a effectué pendant le stage, il s'agira principalement de transférer des fichiers pdf pour alimenter ensuite la base bibliographique Zotero. Mais il est également possible avec Filezilla de transférer des répertoires avec le même procédé.

Figure 2 – Vue des transferts de fichiers pdf de l'ordinateur local vers le serveur.



Le choix de la **base bibliographique Zotero** pour recevoir les documents sélectionnés a été fait en préambule du stage.

Une étude avait été faite par l'ENPC pour choisir la plate-forme de consultation des documents. La plate-forme devait répondre à plusieurs critères :

- avoir une possibilité de stockage supérieure à 20GB.

- proposer un référencement des documents sous un format bibliographique standard.

- proposer une recherche simple (sur auteur, titre, date...) et sur le texte intégral du document.

- donner un accès web simple et sans contrainte d'installation particulière.

Une étude comparative a été faite entre Mendeley et Zotero, deux outils gratuits pour la version de base.

Les résultats de l'étude montrent que Zotero se positionne devant Mendeley dans l'utilisation notamment de la recherche plein texte. Zotero offre la possibilité de consulter et d'interroger la base bibliographique directement sur le site web, sans aucune installation sur un ordinateur individuel, à la différence de Mendeley. Les fichiers pdf sont déposés sur un serveur FTP (le serveur du CIRED). Un pointeur (adresse web du fichier sur le serveur) est placé dans la référence Zotero au niveau du champ URL. C'est à partir de ce pointeur que l'on peut consulter le fichier pdf et interroger le FullText sur Zotero.

Zotero reste le logiciel le plus répandu dans le milieu de l'enseignement supérieur et de la recherche. Cet élément a été un point important dans le choix de la plate-forme, à côté évidemment des possibilités techniques offertes par le logiciel. Une attention particulière a été apportée au confort d'utilisation des chercheurs. Dès le départ, il a été souhaité que la communauté scientifique du laboratoire s'approprie la plate-forme et la recherche plein texte pour que très vite cet outil devienne indispensable.

La comparaison a été faite dès le début entre ces deux gestionnaires bibliographiques mais il en existe d'autres moins connus comme Papers, EndNote, JabRef, Reference Manager...

2.2.2 - L'exploitation des données et les limites rencontrées

Dans l'exploitation des documents et avant de commencer réellement le travail d'archivage électronique, il y a lieu de reprendre l'existant, l'appréhender pour pouvoir parfaitement l'intégrer dans le projet final et assurer ainsi une cohérence entre la centaine de références et mon travail lors du stage.

Avant d'aborder de façon plus précise l'exploitation des données, j'ai voulu différencier les deux grands types de documents (formats papier et électronique natif). Cela dans le but, de mieux présenter les différents points réalisés au cours de ce travail :

- Exploitation des données papiers. Pour ces documents devenus numériques à partir d'un original papier qui a été scanné, la version numérique ne représente qu'une copie. L'original reste la version papier. La consultation du Vade-mecum des Archives de France peut guider le laboratoire et l'ENPC dans les choix faits dans le traitement de la documentation papier numérisée. Il reste à noter que toute destruction de documents doit s'accompagner du visa de la direction des Archivages.

- Exploitation des données électroniques. Ces documents électroniques dits « natifs » sont archivés classiquement sur le serveur.

Pour ces deux types de documents, il doit être réalisé différentes actions :

- Sélectionner les documents constitutifs de la mémoire du laboratoires en associant les chercheurs du laboratoire. Un travail plus particulier a été fait avec les chercheurs du laboratoire pour cerner au mieux leurs attentes et proposer une sélection pertinente. La lecture en amont de différents rapports de recherches du CIRED (bilan annuel, rapports d'évaluation CNRS, AERES...) m'ont permis de mieux appréhender les thématiques de recherche.

- Réaliser les opérations de numérisation sur la station professionnelle en tenant compte des caractéristiques matérielles des documents. La station de numérisation permet de scanner des livres ouverts et de réaliser ainsi des numérisations de qualité professionnelle. La station est dotée d'un système de plateaux ajustables automatiquement. Cette fonctionnalité permet de remédier à des problèmes d'ouverture des documents. La puissance du capteur permet d'atteindre en standard une définition en 300 dpi (dot per inch ou point par pouce). La numérisation est réalisée en mode image (pdf multipage). Le choix de ce format a pu se faire en raison de ses nombreux avantages : protection en écriture, souplesse d'utilisation, téléchargement plus aisé... Puis, le passage en mode texte se fait avec l'application LIMB qui propose une solution d'océrisation (Optical Character Recognition).

- Déposer les documents électroniques sur le serveur de l'ENPC en respectant la Charte de nommage de l'École. Un document réalisé par l'ENPC reprend les principales consignes de nommage des documents sur le serveur. Chaque document numérisé a un premier identifiant correspondant à l'organisme, ici ENPC. Puis le deuxième identifiant attribué correspond à l'opération, au lot de numérisation, ici 00. Ensuite le troisième identifiant donné correspond au type de document, ici CIRED. Enfin, le quatrième identifiant correspond à la cote du document, ici CIR_GOD_0020 (par exemple). La Charte de nommage est disponible et consultable dans la salle de numérisation de l'ENPC. A chaque document déposé sur le serveur ENPC est attribué un nom de fichier sur ce modèle. Par exemple : ENPC00_CIRED_CIR_GOD_0020.pdf.

- Réaliser les opérations d'océrisation sous le logiciel LIMB en maîtrisant tous les enjeux de l'OCR. Cette étape reste primordiale dans le développement du projet de création de la base ressource. Tous les documents sélectionnés (papiers et électroniques) ont été océrisés en faisant reconnaître des chaînes de caractères par la machine pour ensuite valoriser la recherche plein texte.

- Déposer les fichiers de sortie sur le serveur du CIRED en s'assurant de l'accès plein texte sur Zotero. Les fichiers pdf après traitement sont stockés sur le serveur via Filezilla. Une attention particulière est apportée au nommage de ces fichiers pdf pour garantir un accès fiable vers le texte intégral du pdf depuis la base bibliographique Zotero.

Une autre exigence est d'empêcher GoogleBot d'indexer le contenu de la base. Pour cela, il convient de créer un fichier 'robots.txt' où on définit un user-agent et ce qu'il peut ou non indexer avec les instructions allow et disallow. (cfhttp://www.user-agents.org/)

- Diffuser les documents finaux sur la plate-forme Zotero et s'assurer ainsi de l'interopérabilité des données. Les documents sont référencés sur la base ressource. Pour chaque notice, différents champs sont renseignés. Les métadonnées associées aux fichiers pdf sont importées. Le lien vers le serveur de dépôt des fichiers pdf est donné manuellement. Ce point peut fragiliser la base. Toute erreur de saisie rend la consultation du fichier pdf en texte intégral et la recherche plein texte inaccessibles.

Les limites identifiées au démarrage du stage sont de natures diverses. Elles ont été de plusieurs ordres.

Limites organisationnelles avec la gestion du temps. La réalisation du stage sur trois mois et non sur quatre comme prévu lors de l'annonce va nécessité de faire des choix. Dès le départ, il a fallu réajuster le travail au vu de cette contrainte organisationnelle. Je propose de travailler très vite sur la sélection des documents à numériser. Il est ainsi décidé de travailler tout d'abord sur les fonds Godard et Sachs et dans un deuxième temps sur les fonds Hourcade et Général.

Limites structurelles avec l'absence d'un module archivage électronique au cours de la formation. Afin de contourner cette difficulté, il me faut consacrer un peu de temps à consulter de la documentation. J'ai voulu consulter une documentation professionnelle avec une expérience propre à l'enseignement supérieur et au milieu de la recherche.

Il est possible de m'appuyer sur différents sites : le site de la BNF et les rubriques destinées aux professionnels, celui de l'ENSSIB et différents rapports autour de l'archivage électronique ainsi que des documents publiés par les Archives de France.

Ces limites sont quantitatives (quantité de dossiers traités) mais elles ne sont pas techniques. En effet, tout le processus technique : du document sélectionné à sa publication sous forme océrisée est traité. Ceci constitue un réel enrichissement personnel du fait de l'acquisition de nombreuses connaissances nouvelles.

3 - Le travail réalisé

Le travail d'archivage électronique des documents du CIRED a permis de référencer plus de 7500 pages, 505 documents consultables en texte intégral et interrogeables en plein texte. Sur l'ensemble de ces documents, il a été réalisé un travail d'identification, de numérisation, d'océrisation et de référencement documentaire. La constitution de cette base ressource doit devenir un outil de recherche pour l'ensemble de la communauté scientifique.

3.1 - L'organisation du projet

Le descriptif du stage ainsi que les attentes du CIRED et de l'ENPC ont demandé dès le départ une organisation rigoureuse. Mais j'ai pu dès le début connaître **les moyens** qui m'ont été accordés (accès serveurs, accès base ressource, manipulation du Copybook, formation au logiciel d'océrisation Limb,...). J'ai également été impliquée dans la mise en place de **l'organisation** du stage. J'ai pu proposer un planning. Je connaissais dès le début l'échéance et **le délai** accordé pour réaliser la mission. Ces trois mois de stage, outre l'objectif chiffré des 500 premières références, doivent également servir de tremplin à d'autres actions d'archivage électronique du fonds documentaire du CIRED. Les interlocuteurs souhaitent faire perdurer le travail réalisé par l'embauche de stagiaires et de contractuels.

Je n'ai pas eu à gérer de **budget** particulier. Les questions financières ont été réglées avant le début du stage avec la participation du DIM R2DS au projet de numérisation.

Concrètement, j'ai été amenée à me déplacer sur deux lieux distincts. J'ai proposé de privilégier au début du stage, l'identification des documents au CIRED. Après discussions avec les chercheurs du laboratoire, je savais qu'ils étaient plus disponibles au cours du premier mois de stage. J'ai donc travaillé plus étroitement avec eux à ce moment-là. Cela m'a aussi permis de présenter plus précisément le projet et de le valoriser auprès de la communauté scientifique. Après la constitution du corpus, il a été défini tout le processus pour pouvoir l'archiver électroniquement et le valoriser sur la base ressource.

3.1.1 - La gestion du projet

La gestion du projet s'est faite en plusieurs étape. Lors de l'entretien, il m'a été fait une présentation du projet, les attentes des différentes parties et les actions à mener. Quelques semaines avant le début du stage, **une réunion de lancement du projet** (ou *kickoff meeting*) m'a permis d'être présentée à l'ensemble des interlocuteurs. Il a été abordé tous les points essentiels relatifs au projet. Il a été rappelé le contexte du projet, les grandes tâches à réaliser, le plan qualité décidé. Parmi les questions de gestion et de suivi, il m'a été demandé de réaliser un tableau de bord reprenant les principaux enjeux du projet et l'avancement des tâches.

C'est également au cours de cette réunion, que j'ai pu proposer un planning des actions à mener sur les trois mois de stage.

Puis, je me suis appuyée sur l'expérience des uns et des autres pour commencer le travail d'archivage électronique. Ces informations récoltées au cours de cette réunion m'ont permis de réaliser une note de cadrage. J'ai pu ainsi clarifier l'idée du projet et faire un état des lieux.

Trois autres réunions ont rythmé mon stage au CIRED. J'ai eu des réunions de coordination du projet avec une présentation de ma part du suivi du projet et de sa gestion. J'ai également eu des réunions de travail consacrées à la réalisation de l'objet.

C'est au cours de ces réunions que j'ai exposé l'avancement du projet et montré sa réalisation. Les étapes ont pu être évaluées et validées. Des tests ont été faits par l'ENPC sur la base ressource pour vérifier la conformité technique et fonctionnelle. Un plan qualité a été établi avec les procédures de validation de l'avancement du projet, les procédures de gestion des risques et les procédures de gestion des délais.

C'est également au cours de ces réunions, qu'il a été décidé des outils de communication pour valoriser le projet. A la demande des chercheurs du CIRED une note de présentation a été fournie à la fin du stage. Elle sera présentée à l'ensemble des chercheurs et au COPIL du laboratoire.

Cette communication permet de valoriser le travail réalisé, de transmettre et partager les connaissances et les savoirs-faire. Cette diffusion du projet auprès de la communauté

scientifique peut également servir d'incubateur pour d'autres actions et attirer d'éventuels investisseurs.

L'ensemble des interlocuteurs savent qu'il ne s'agit que de la première étape du processus de création de la base ressource. Il a été discuté de la suite de l'action menée et des possibilités offertes au CIRED. Le but étant de pérenniser les résultats du projet et de capitaliser l'expérience acquise pour la faire vivre au-delà du stage.

Concrètement, j'ai pu apprécier la fréquence des entrevues, cela m'a permis de répondre pleinement à l'objectif des 500 premières références disponibles sur la base ressource.

Pour répondre à la difficulté éventuelle des déplacements sur plusieurs lieux, il a été proposé plusieurs outils de communication en accès restreint sur le web. Outre l'organisation matérielle du stage, ces outils ont permis aux différents interlocuteurs de suivre hebdomadairement l'avancée du stage. Je me suis également assurée d'avoir un accès aux serveurs de l'ENPC, du CIRED. J'ai utilisé Filezilla pour la connexion au serveur du CIRED et j'ai demandé au service informatique de l'ENPC un accès au serveur d'archive. Un planning d'utilisation de la salle de numérisation a été établi à l'ENPC. Cela m'a assuré d'avoir accès en permanence au copybook et au logiciel d'océrisation, Limb. Enfin, j'ai ouvert une session zotero pour pouvoir alimenter la base ressource en m'assurant de l'installation des différents plugin nécessaires.

3.1.2 - Les tâches

Les différentes tâches décrites dans la suite de ce rapport représentent la portée dans la gestion du projet Pour une meilleure compréhension de la mission qu'y m'a été confiée, je vais les décomposer et proposer un ordonnancement des tâches constitutives du projet.

Cet organigramme technique du projet permet de détailler le projet en tâches, en lots de travail (ou *work package*).

Ces activités entretiennent des relations fonctionnelles entre elles. Il y a une notion de dépendance. Certaines tâches doivent être réalisées avec succès à un moment précis

pendant la durée du projet. Il est donc nécessaire de définir les principales tâches et de proposer un calendrier de réalisation[3]. Deux grands types d'activités ont pu être identifiés:

- des tâches techniques (identification, numérisation, dépôt sur la base ressource).

- des tâches de communication (support de présentation).

Pour chacune des ces tâches, il est nécessaire d'identifier :

- le responsable de la tâche

- l'objectif général de la tâche

- la description synthétique du travail réalisé

- la durée de la tâche et sa période de réalisation

- la relation fonctionnelle

- **Identification du fonds CIRED** pour la numérisation. Ce travail scientifique d'identification a été réalisé conjointement avec les chercheurs du CIRED. Ils ont validé le corpus constitué. L'objectif principal est de rassembler sous un même corpus les différents fonds scientifiques du laboratoire après avoir sélectionné les documents qualifiés d'incontournables pour la mémoire du CIRED. Ce travail a nécessité un mois de présence au CIRED avec les chercheurs. Il a été réalisé pendant le premier mois de stage (avril 2015). Cette activité est la première identifiée pour mener à bien le projet. Elle est en relation directe avec les autres tâches identifiées par la suite.

- **Numérisation et Océrisation du fonds CIRED**. Ce travail technique a été réalisé dans les locaux de l'ENPC. J'ai pu gérer en toute autonomie cette activité, j'en étais le principal responsable. L'objectif général est la numérisation et l'océrisation du corpus identifié dans la première étape. Ce travail s'est déroulé pendant deux mois (mai et juin 2015). Cette activité a pu se faire en parallèle de celle identifiée dans le point suivant (valorisation du corpus). Elle est dépendante de la première tâche et sans sa réussite, la dernière activité référencée (médiation du corpus) ne peut être validée.

[3] A3 – Extrait du diagramme de Gantt.

- **Valorisation du fonds CIRED sur la base ressource.** J'ai pu réaliser ce travail technique en toute autonomie, j'en étais le principal responsable. L'objectif principal est le référencement documentaire sur la base Zotero. Cette activité a nécessité deux mois de travail (mai et juin 2015). Elle s'est faite en parallèle avec la numérisation et l'océrisation sous Limb. Elle est dépendante de la première tâche et sans sa réussite, la dernière activité référencée (médiation du corpus) ne peut être validée.

- **Médiation du fonds CIRED auprès de la communauté scientifique**. Ces actions de communication ont été réalisées en direction de la Communauté scientifique du laboratoire. J'en étais la responsable principale. Elles ont pu prendre la forme d'une plaquette de présentation de la base ressource et de son utilisation future par les chercheurs du laboratoire. Elles ont eu lieu à la fin du projet (en juin 2015). Elles pourront être reconduites lors de manifestations scientifiques d'ici la fin de l'année lors de colloques, conférences. Cette étape reste indispensable dans la gestion du projet, elle valorise le travail réalisé et diffuse ce nouvel outil auprès des chercheurs.

3.1.3 - Le planning

Le planning permet de fixer les dates de réalisation du projet, d'identifier les jalons et atteindre les objectifs du projet. Dans l'établissement du planning, il est important de prendre en compte le chemin critique, c'est à dire le chemin continu entre la date de début et de fin du projet. Ici, ce chemin critique court du 30 mars au 30 juin 2015. Tout retard d'une tâche du chemin critique est répercuté sur la durée du projet et sur la date de fin. Dans la mise en place du planning, certains projets portent une attention particulière à la marge : la marge totale et la marge libre. Cela n'a pas été mon cas ici mais pour des projets plus longs, il est intéressant de calculer ces délais supplémentaires.

Les 14 semaines de stages se sont organisées en fonction de l'objectif final et des contraintes des différents interlocuteurs. Le diagramme de Gantt reprend ces différentes étapes. Il a permis de mieux appréhender les différentes tâches et de voir les interactions possibles.

J'ai pu faire une proposition de planning :

- Travail de repérage, sélection des documents sous la validation scientifique de chercheurs du CIRED.

- Travail de numérisation sur le Copybook dans les locaux de l'ENPC.

- Travail d'océrisation avec le logiciel Limb.

Pour ces deux dernières missions, des formations m'ont été proposées. La société Spigraph qui commercialise le logiciel d'océrisation Limb est notamment intervenue pendant toute une journée.

Ces actions de formations ont été intégrées au module numérisation – océrisation.

3.2 - L'interaction avec l'équipe

La communication avec les différents interlocuteurs est indispensable pour le bon déroulé du projet. Son efficacité passe par des outils adaptés au projet. Elle prend la forme de réunions informelles, entre deux parties prenantes du projet. Des réunions plus officielles peuvent être organisées entre les responsables du projet.

Des bilans d'étapes et un bilan final ont été rédigés. Les documents ont pu être diffusés par e-mail. Un réajustement des tâches a pu être proposé en fonction de l'avancement et du temps écoulé. Une attention plus particulière a été portée sur certains points.

Deux outils collaboratifs ont été lancés : un planning sous Trello et un journal de bord sur Google drive. Des échanges ont eu lieu via le journal de bord. Des points plus particuliers ont pu ainsi être détaillés.

J'ai eu au quotidien une très grande autonomie dans la gestion de ce projet de numérisation mais j'ai toujours eu la possibilité d'échanger avec les différents interlocuteurs via ces outils de communication.

3.3 - La nécessité des choix

Dans toute gestion de projet, l'équipe peut être amenée à réajuster les tâches définies et à proposer des choix pour aboutir au résultat fixé. Parfois certains choix s'imposent d'eux même. C'est ce que l'on a pu expérimenter ici avec l'utilisation de Zotero pour la base ressource. Mais parfois, certains choix deviennent nécessaires même s'ils n'étaient pas envisagés au départ.

3.3.1 - Des choix techniques

L'objectif du CIRED a été dès le départ de permettre la consultation à distance de ses archives scientifiques. Il était tout de suite très intéressé par les possibilités offertes par Zotero et notamment la recherche plein texte sur les documents numérisés.

Les choix techniques opérés, découlent de cet objectif. Ils portent sur deux points en particulier :

- **La numérisation sur le Copybook et l'océrisation sous Limb.**

J'ai utilisé le Copybook Cobalt commercialisé par la société i2S. Il rentre dans la gamme des scanners de moyenne génération. Il permet à la fois une manipulation simple et un résultat de très bonne qualité. Dans le traitement de l'image, différentes options peuvent affiner les résultats. Il s'agit ici d'un scanner avec prise de vue photographique. Le document est déposé sur un plateau balancier permettant ainsi une numérisation à plat.

Une autre gamme de scanner par prise de vue photographique existe mais ce type d'appareil est utilisé pour des documents plus fragiles. Il s'agit alors d'une machine avec un berceau en forme de V. Un robot tourne les pages du documents, limitant ainsi l'intervention de l'opérateur.

Enfin, d'autres appareils proposent une numérisation par balayage. Les capteurs se déplacent sur le document et ils reconstituent l'image pixel par pixel. Ce type de matériel est utilisé pour des document de grande taille, volumineux.

Les principaux concurrents de i2S sont européens : l'entreprise InoTec, l'entreprise Image Access et l'entreprise Kodac...

Le logiciel OCR LIMB permet lui de traiter, d'enrichir, de convertir des fonds documentaires. Dans la manipulation de LIMB, j'ai eu à utiliser la fonction Limb Processing. Trois autres fonctions sont offertes par Limb : Limb Capture, Limb gallery et Limb Maestro.

La fonction Processing permet le traitement, la structuration et la préparation des données. La version 3.1 du logiciel permet l'importation des métadonnées présentes dans le fichier source. Elle permettrait également la détection multi-zones sur des documents de type photos, schémas.

Les principaux concurrents de LIMB sont : ReadIris, FineReader, Omnipage professionnal, PaperPort professionnal.

- **Le référencement sous Zotero et l'accès distant des fichiers pdf en texte intégral.**

Le gestionnaire bibliographique Zotero permet de stocker des références, de les gérer, de les partager et de les exporter pour réaliser notamment des bibliographies. Zotero est disponible en tant qu'extension du navigateur Firefox et en version *standelone*. Lors de son installation, il est important de s'assurer que pdftotext et pdfinfo soient bien configurés. Ces deux éléments permettent la recherche plein texte des documents sur la base.

Il aurait pu être envisagé de mettre en place des bibliothèques numériques sous Oméka ou Wordpress… deux *content management system* (CMS) pour constituer et valoriser le fonds d'archives électronique. Ces outils sont faciles d'utilisation mais ils ne peuvent être manipulés par toute une communauté scientifique. Un administrateur gère les accès des différents utilisateurs. Il y a moins de souplesse à utiliser ces bibliothèques numériques (installation, gestion des comptes, des contenus...). Mais à terme, il pourra être envisagé une présentation de la base ressource dans une bibliothèque numérique.

3.3.2 - Des choix méthodologiques

Sur la période du stage au CIRED et à l'ENPC, j'ai eu à faire des choix méthodologiques.

Il a été prévu dès le départ de consacrer les premières semaines de stage à l'identification des documents du laboratoire. Puis, avec cette « première récolte », le travail de numérisation et d'océrisation sur les documents a été lancé. Mais une deuxième période avait été envisagée pour continuer la sélection des documents scientifiques, à la mi-mai.

Nous avons été très vite confrontés à deux aspects de la mission, peut-être sous estimés lors de l'établissement du planning : le temps nécessaire pour la sélection des documents et le temps nécessaire pour la numérisation et l'océrisation des documents.

Il a été envisagé lors d'une réunion avec les différents interlocuteurs du CIRED et de l'ENPC de procéder à l'ensemble des étapes pour les corpus déjà sélectionnés des fonds Sachs, Godard mais de ne traiter que dans un deuxième temps les documents des fonds Hourcade et Général.

Il reste que dans les fonds Sachs et Godard, j'ai pu identifier des documents des fonds Hourcade et Général. Ces quelques dizaines de références ont été traitées et elles constituent la base ressource comme l'ensemble des autres documents.

3.4 - La mise en œuvre du projet

La mise en œuvre du projet a pu se faire sans gêne particulière. La préparation en amont des différents aspects techniques a permis de lancer le processus très rapidement. J'ai pu rencontrer quelques problèmes avec le logiciel d'océrisation Limb, notamment quelques lenteurs dans l'import des fichiers pdf. Mais cela n'a pas empêché le travail d'ensemble.

3.4.1 - L'exploitation de plusieurs sources

Identification :

Typologie des documents : papiers préparatoires pour publication, articles publiés par les chercheurs du CIRED, actes de colloque, rapports, cours, notes internes au CIRED.

Il s'agit uniquement de documents scientifiques qui ont fait la renommée du laboratoire et qui sont marqués actuellement comme indispensable par les chercheurs actuels.

L'identification s'est faite sous la validation scientifique de chercheurs du laboratoire. Parmi les critères de sélection, j'ai pu lister la date de publication, l'auteur du document, le type de document... Dans ce processus d'identification des documents et pour préparer ensuite les actions de numérisation, j'ai également réalisé un tableau d'inventaire avec différentes rubriques (identifiant ENPC, cote définissant une partie de l'adresse URL sur Zotero, titre, auteur, pages numérisées, valorisation possible sur HAL...).

Numérisation :

Après l'identification des documents à numériser, j'ai dû configurer le Copybook pour optimiser les résultats sur le corpus sélectionné.

Il est nécessaire de créer un dossier pour y rentrer les différents paramètres. Puis, quatre onglets sont à renseigner.

- Onglet général. Je précise le type de document, ouvrage...

- Onglet Format. Je sélectionne ici le cadre en fonction des dimensions du document. On peut définir s'il s'agit d'un cadre manuel ou automatique, le mode couleur ou noir et blanc. C'est également dans cet onglet que l'on donne le format (2xA4) et la résolution (300 dpi).

- Onglet Traitement. J'ai eu la possibilité de définir différentes options qui améliorent la qualité de la numérisation. Parmi les options offertes par le Copybook, on peut citer : la correction de l'inclinaison des pages, la correction géométrique, l'amélioration de l'image et permettre ainsi une meilleure lisibilité des détails en atténuant le bruit numérique de l'image ; le travail sur les contrastes avec la redistribution des niveaux de l'image, pour renforcer justement les contrastes et l'étirement des niveaux qui permettent un ajustement automatique des niveaux de l'image pour optimiser sa dynamique.

- Onglet Archivage. Avec cette dernière option, j'ai défini les caractéristiques de l'image restaurée. C'est ici que l'on peut donner le nom du fichier de sortie (l'identifiant ENPC), le chemin vers le serveur de stockage des documents et le format du document numérisé (TIFF, JPEG, PDF...).

Figure 3 - Vue du Scanner Copybook. Source i2S – Spigraph.

Figure 4 - Configuration du scanner avant la numérisation. Source i2S – Spigraph.

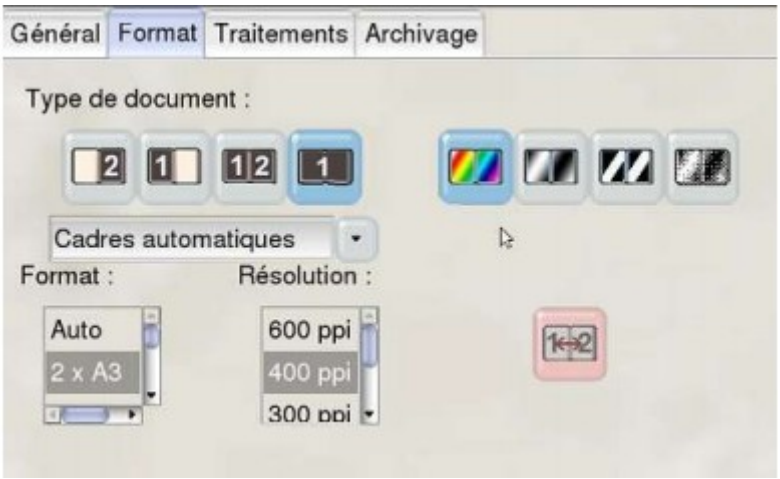
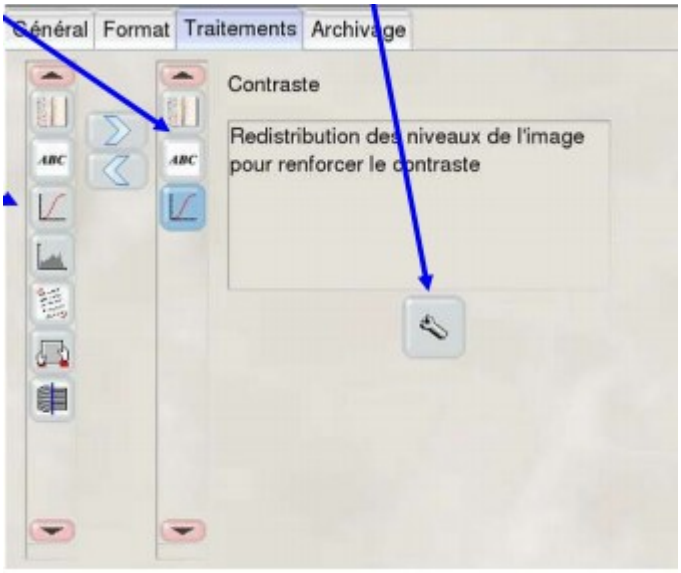


Figure 5 - Configuration des options avancées. Source i2S – Spigraph.



Océrisation :

Le logiciel LIMB utilisé pour ce module, permet de traiter, d'enrichir et de convertir des fonds documentaires. Cette étape est primordiale dans le processus d'archivage électronique du fonds scientifique du CIRED. En effet, la qualité de l'OCR permet la recherche plein texte sur la base ressource.

Pour optimiser ce module, j'ai suivi une journée de formation auprès du prestataire Spigraph. J'ai ensuite utilisé la fonction LIMB Processing qui permet le traitement, la structuration et la préparation des données. J'ai défini un modèle qui peut être ensuite utilisé pour les autres documents du projet. Le travail d'océrisation s'organise en différentes étapes. Il est important de préparer le document avant le module OCR et l'export final. Parmi les fonctions avancées de la fonction LIMB Processing, on peut noter la possibilité de construire une table des matières à partir des informations issues du document. Cela permet à l'export d'avoir des documents pdf structurés.

Le module OCR s'appuie sur un dictionnaire de 140 langues y compris l'arabe, l'hébreux et les langues asiatiques. Parmi les moteurs utilisés, on peut citer ABBYY, Iris, Sakhr Software.

Le module export permet de choisir plusieurs formats de sortie simultanément et notamment les formats PDF/A1, A2 et A3. Ces formats de fichiers ont été approuvés par l'Organisation Internationale de normalisation (ISO) en 2005. Ils sont particulièrement utilisés pour l'archivage électronique des documents. Avec ces nouveaux formats, les fichiers conservent tous les éléments de couleur, de police, d'image... dans le document lui -même. La principale caractéristique de cette méthode est de rendre le format pdf autonome de tous outils de création, de stockage et de lecture des fichiers.

Figure 6 - Configuration des langues pour OCR.

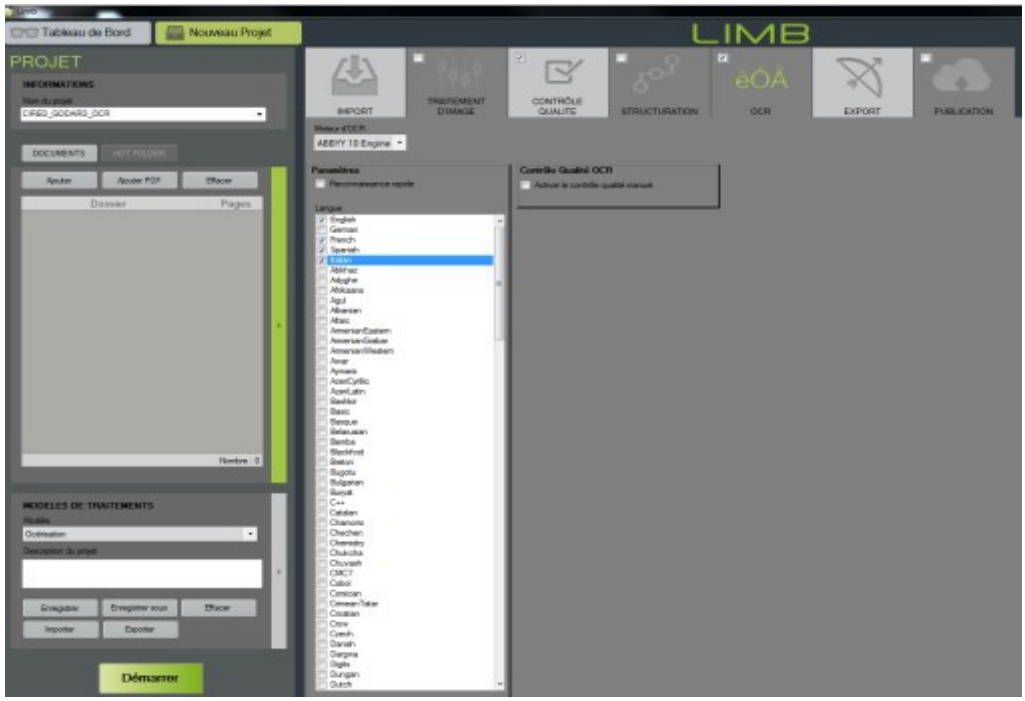


Figure 7 - Configuration des formats pour l'export.

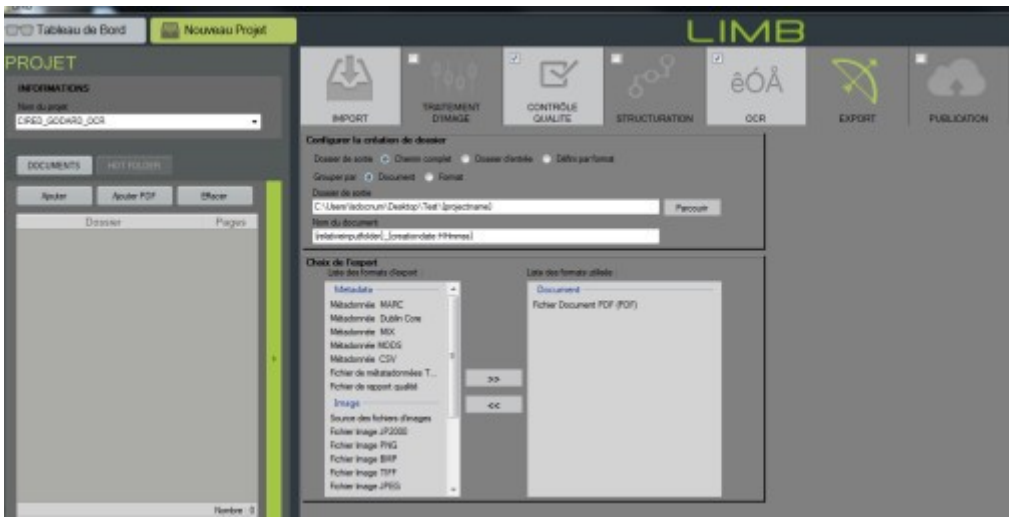


Figure 8 - Vue du tableau de bord avec les différentes phases lors de l'OCR.

3.4.2 – Alimentation de la base ressource

L'alimentation de la base peut se faire de deux façons :

- Manuellement en signalant le document sur la base (création d'une référence bibliographique et alimentation des différents champs bibliographiques).

- Automatiquement en important les métadonnées du PDF.

Cette automatisation reste la façon la plus pertinente pour alimenter la base mais elle n'a pu se faire que sur les documents natifs électroniques.

Pour les documents devenus électroniques, le référencement sur la base s'est donc fait manuellement. Il reste à noter que pour ces deux situations, le champ URL doit être saisi manuellement. C'est à partir de cette zone que le chercheur accède au texte intégral du document.

Actuellement, les documents devenus électroniques par la numérisation sont stockés sur les serveurs de l'ENPC et du CIRED. Il peut être envisagé par la suite une externalisation des données vers les serveurs du Centre Informatique National de l'Enseignement Supérieur (CINES). Ce centre propose un archivage pérenne des données de la recherche sur la plate-forme PAC.

Figure 9 - Vue de la base ressource Zotero, accès professionnel.

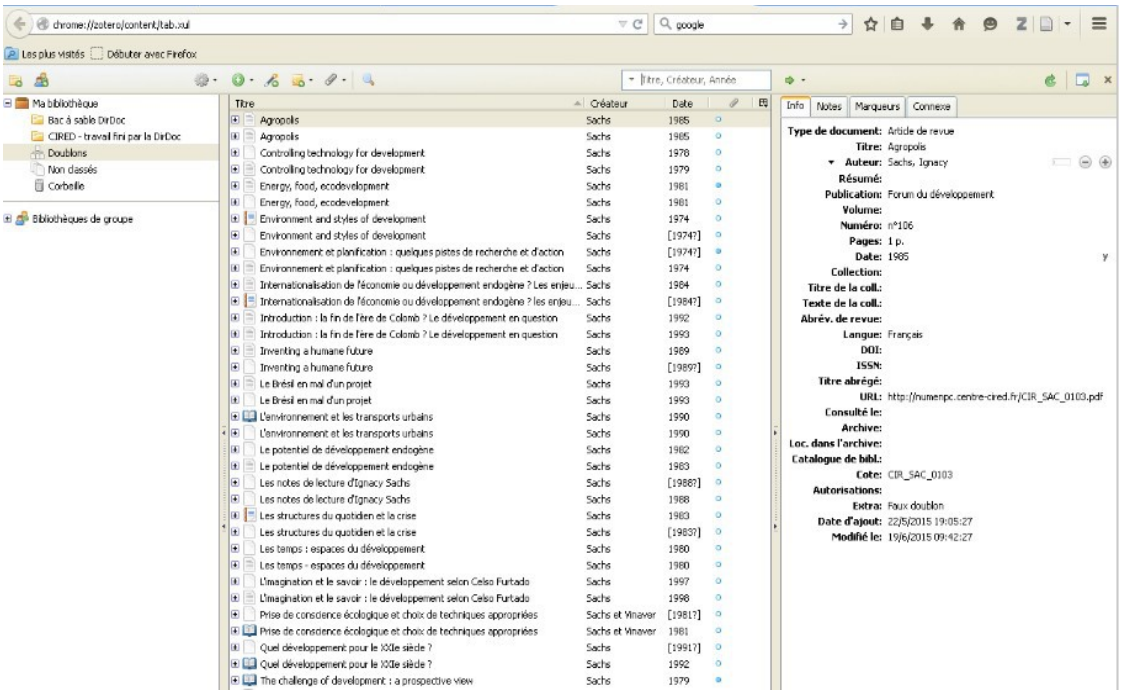
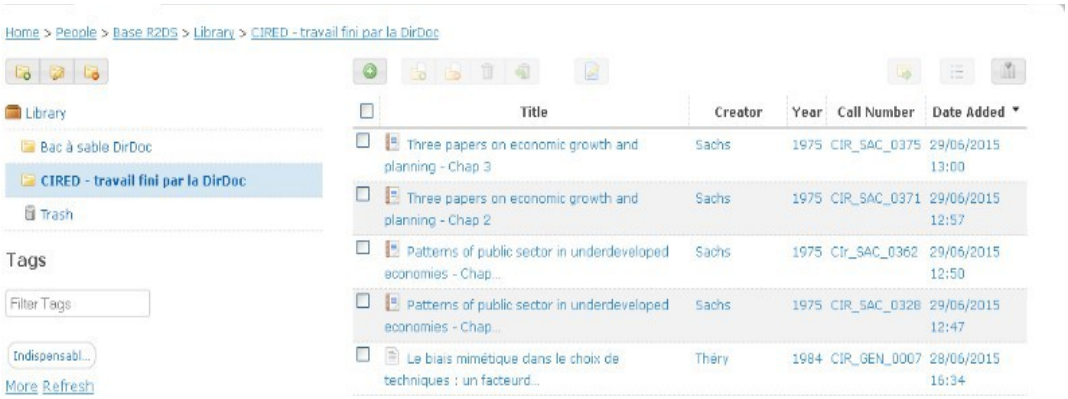


Figure 10 - Vue de la base ressource Zotero, accès Standelone.



3.4.3 - Exploitation et valorisation de la base ressource

J'ai valorisé la base ressource lors de la présentation des résultats obtenus sur la plaquette d'utilisation. D'autres actions seront proposées lors de COPIL ou de manifestations scientifiques d'ici la fin de l'année. La connexion très simple sur le site web de Zotero doit permettre une prise en main et une exploitation rapide de la base par les chercheurs du laboratoire. Des actions de formation ont été par le passé proposées sur cet outil. Elles pourront être reconduites. Cela peut faciliter l'appropriation de la base par les chercheurs et les jeunes chercheurs du CIRED.

Il peut être intéressant de diffuser le contenu de la base ressource sur la plate-forme d'archive ouverte HAL. Le format BibTex semble le mieux correspondre pour réaliser cette manipulation. A l'origine conçu pour laTex, il est utilisé pour gérer et traiter des bases bibliographiques. Bib2HAL permet de déposer des fichiers au format BibTex. Une documentation disponible sur le site web du CCSD peut aider à l'importation des références. Zotero permet également d'exporter les collections au format BibTex.

Dans le cadre de dépôt massif, comme cela peut être envisagé ici, il est évidemment important de s'assurer des droits d'auteur pour verser les documents dans une archive ouverte.

3.5 - Les résultats du processus

Les résultats obtenus correspondent aux attentes fixées par les interlocuteurs au début du stage. Les 500 premières références sont disponibles sur la base ressource. La recherche plein texte sur le texte intégral des documents numérisés donne de bons résultats grâce à l'océrisation. Dans l'avenir, le travail doit être reconduit pour finir l'exploitation des documents du CIRED et proposer ainsi une base reprenant l'ensemble des textes constituant la mémoire scientifique du laboratoire.

3.5.1 – Les pistes de réflexion

Ce travail a permis un enrichissement du fonds scientifique du CIRED. Cela a été également l'occasion de proposer ce processus d'archivage électronique à d'autres laboratoires et de créer ainsi une dynamique.

Parmi les pistes de réflexion qui pourraient être intéressantes pour ce type de projet, on peut citer :

- la correction participative de l'OCR avec le projet CORRECT

- l'utilisation du format ALTO (Analyzed Layout and Text Object) avec la conservation de toutes les coordonnées géométriques du document (textes, illustrations, graphiques...).

- l'utilisation de la norme ISAD (G), norme internationale qui permet à la fois une recherche d'information par les chercheurs et par des services d'archives. La dernière permet de décrire 7 zones (identification, contexte, contenu...).

Conclusion

L'objectif des cinq cents premiers documents interrogeables depuis la base bibliographique Zotero a été atteint. J'ai pu mettre en place un processus de traitement qui sera repris lors de futures actions d'archivage électronique.

La base va perdurer, les chercheurs vont pouvoir se l'approprier et proposer leur participation avec notamment, l'ajout de tags. Les actions de formation ont permis à la communauté scientifique de maîtriser l'outil Zotero, elles pourront être reconduites dans les mois qui viennent.

Le travail réalisé lors de ce stage m'a permis d'acquérir de nouvelles compétences et de gérer un projet de numérisation de la sélection des documents à leur diffusion sur une base ressource. J'ai également eu l'opportunité de suivre une journée de formation sur un logiciel d'OCR. Cela m'a permis d'enrichir mes connaissances techniques et méthodologiques avec la gestion de ce projet.

Références bibliographiques

Ouvrages

- Archives de France. *Autoriser la destruction de documents sur support papier après leur numérisation : quels critères de décision ?* Paris : Archives de France, 2014. 18p.

- Lorène Béchard, Lourdes Fuentes Hashimoto, Edouard Vasseur. *Les archives électroniques*. Paris : Association des archivistes français, 2014. 82p.

- Charles Kecskeméti, Laros Körmendy. *Les écrits s'envolent : la problématique de la conservation des archives papier et numériques*. Lausanne : Favre, 2014. 207p.

- Bruno Racine. *Schéma numérique des Bibliothèque*. Paris : La Documentation française, 2010. 88p.

Sites web

- Archivage numérique

- http://www.piaf-archives.org/espace-formation/

- http://www.bnf.fr/fr/professionnels.html

- http://bbf.enssib.fr/

- Zotero

- https://hal.archives-ouvertes.fr/page/presentation

- http://zotero.hypotheses.org/51

Table des annexes

A1 – Organigramme du Centre International de recherche sur l'environnement et le développement

A2 – Organigramme de l'École des Ponts et Chaussées

A3 – Extrait du diagramme de Gantt

A1. - Organigramme du CIRED

Centre International de Recherche sur l'Environnement et le Développement

Directeur : Franck Lecocq
Directrice adjointe : Catherine Boemare
Secrétaire général : Naceur Chaabane

C.I.R.E.D.
UMR 8568
CNRS · ENPC · CIRAD
EHESS · AgroParisTech

AXES DE RECHERCHE

Prospectives sectorielles : énergie, ville, usage des terres

B. Barraqué, M. Benito Garzon, T. Brunelle, C. Boemare, B. Dorin, P. Dumas, D. Finon, L.-G. Giraudet, M. Ha Duong, J.C. Hourcade, T. Le Cotty, H. Levrel, A. Nadaï, P. Quirion, V. Viguié

Doctorants : P. Avner, N. Berghmans, F. Branger, W. Dang, L. Faucheux, A. Fontaine, J. Hamann, E. Lanckriet, R.A. Lopez Gonzalez, N. Neverre, T. H. A. Nguyen,V. Przyluski, A. Saussay, M. Solignac, A. Vallet

Stratégies de développement sous contrainte climatique, environnementale et sociale

C. Barbier, C. Cassen, N. Chaabane, E. Combet, T. Gasser, F. Ghersi, C. Guivarch, J.C. Hourcade, A. Méjean, P. Quirion, F. Lecocq, J. Lefèvre, E. O Broin

Doctorants : A. Berry, R. Bibas, S. De Lauretis, M. Hamdi-Cherif, G. Le Treut, F. Leblanc, E. Mosseri, Q. Perrier, J. Schers, A. Vogt-Schilb, Y. Quilcaille

Négociations, controverses, processus de décisions sous incertitude

C. Cassen, B. Cointe, P. Dumas, D. Finon, C. Guivarch, M. Ha-Duong, J.C. Hourcade, G. Massard-Guilbaud, A. Nadaï, T. Tazdaït

Doctorants : M.-J. Cardoso de Mendonca, M. Cherbib, M. Domergue, C. Feger, S. Rabaud, I. Shishlov

Modèles, outils, données

R. Bibas, P. Dumas, N. Chaabane, C. Guivarch, J. Lefèvre, P. Mabire, F. Nadaud

Bases de données, Systèmes d'informations

C. Barbier, N. Chaabane, P. Mabire, F. Nadaud

Informatique

R. Hoby

Communication, Valorisation, Diffusion

N. Belalimat, C. Cassen

Gestion, Comptabilité

E. Tyma

Secrétariat de direction

Y. Serfaty

Secrétariat de l'équipe CIRAD

V. Hourmant

Gestion DIM R2DS

C. Boemare, A.Sánchez

A2 - Organigramme de l'ENPC

Les départements d'enseignement de l'École

1re année
Président :
François CHEVOIR
Dir. académique :
N.

Formation linguistique (FL)
Président :
Jörg ESCHENAUER
Adj. : Thomas HARCHARICK
et Amokrane KADDOUR

Sciences humaines et sociales (SHS)
Président :
Gilles CRAGUE

Génie civil et construction (GCC)
Président :
Bernard VAUDEVILLE
Dir. académique :
N.

Génie mécanique et matériaux (GMM)
Président :
Alain EHRLACHER
Resp. académique :
Frédéric TAYEB

Ingénierie mathématique et informatique (IMI)
Président :
Éric DUCEAU
Resp. académique :
Mohammed EL RHABI

Génie industriel (GI)
Président :
Fabrice BONNEAU
Resp. académique :
Aurélie DELEMARLE

Ville, environnement, transport (VET)
Président :
Pierre SALLENAVE
Resp. académique :
Joachim BROOMBERG

Sciences économiques, gestion, finance (SEGF)
Président :
Dominique JACQUET
Resp. académique :
Abdelkader SLIFI

École des Ponts Business School
Alon ROZEN
Président du directoire de MIB Développement

PARIS-EST d.school
Doyenne
Véronique HILLEN

Les laboratoires de l'École

Centre d'enseignement et de recherche sur l'environnement atmosphérique (CEREA)
Directeur :
Christian SEIGNEUR

Centre d'enseignement et de recherche en mathématiques et calcul scientifique (CERMICS)
Directeur :
Jean-François DELMAS

Centre international de recherches sur l'environnement et le développement (CIRED)
Directeur :
Franck LECOCQ

Laboratoire techniques, territoires et sociétés (LATTS)
Directeur :
Olivier COUTARD

Laboratoire eau, environnement, systèmes urbains (LEESU)
Directeur :
Régis MOILLERON

Laboratoire d'Informatique Gaspard Monge (LIGM)
Directeur :
Marie-Pierre BÉAL

Laboratoire de météorologie dynamique (LMD)
Directeur :
Vincent CASSÉ

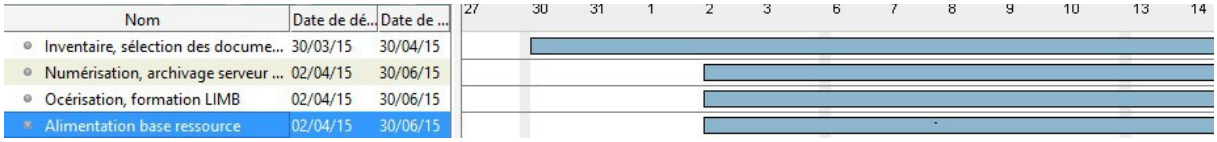
Laboratoire ville, mobilité, transport (LVMT)
Directeur :
Pierre ZEMBRI

Paris-Jourdan sciences économiques (PJSE)
Directeur :
Luc BEHAGHEL

Laboratoire d'hydraulique Saint-Venant
Directeur :
Michel BENOIT

Laboratoire Navier
Directeur :
Karam SAB

A3 - Extrait du Diagramme de Gantt



Nom	Date de dé...	Date de ...
Inventaire, sélection des docume...	30/03/15	30/04/15
Numérisation, archivage serveur ...	02/04/15	30/06/15
Océrisation, formation LIMB	02/04/15	30/06/15
Alimentation base ressource	02/04/15	30/06/15