Développement et Usage des Archives Ouvertes en France. 1e partie : Développement

Joachim Schöpfel, Hélène Prost

Université Charles-de-Gaulle Lille 3
Laboratoire GERIICO
Groupe d'Etudes et de Recherche Interdisciplinaire en Information et Communication

Développement et Usage des Archives Ouvertes en France

Rapport

1e partie : Développement

Joachim Schöpfel, Université Lille 3
Hélène Prost, INIST-CNRS

Lille, Juillet 2010

Résumé : Le rapport présente les résultats d'un projet de recherche mené en 2009 à l'université Charles-de-Gaulle Lille 3. L'objectif du projet : évaluer les résultats de la politique en faveur des archives ouvertes en France. La 1e partie du rapport intitulé « Développement » fournit des éléments chiffrés sur la typologie, la taille et le contenu des archives ouvertes, en comparant si possible l'information de 2009 avec 2008. L'enquête s'appuie sur un échantillon quasi-exhaustif des sites français, constitué à partir de répertoires et annuaires. Les données ont été collectées en ligne, sur chaque site.

Mots clés : archives ouvertes, information scientifique, publication scientifique, accès libre.

Abstract : The report contains the results of a research project conducted in 2009 at the university Charles-de-Gaulle Lille 3. The objective of the project: evaluate the results of the policy of open repositories in France. The 1st part of the report produces figures on the typology, size and content of open archives, compared whenever possible to 2008 data. The survey has been conducted on a quasi-exhaustive sample of French websites, based on directories and lists of the open access movement. Data have been collected on line, on each web site.

Keywords : open repositories, scientific information, academic publishing, open access.

Contact : Joachim Schöpfel joachim.schopfel@univ-lille3.fr

Financement : Projet BQR 2009 Université de Lille 3

Remerciements : Le projet a été subventionné en 2009 par le service recherche de l'université de Lille 3 que nous remercions pour son conseil et aide logistique. Nous remercions également la direction et l'équipe du SCD de Lille 1 pour leur soutien ainsi que tous les participants – enseignants-chercheurs, professionnels et étudiants - ayant contribué au succès de cette étude. Nos remerciements également à Jacques Creusot pour une relecture attentive du manuscrit.

Information sur le projet

Acronyme	DUAO-F
Nom	Développement et usage des archives ouvertes en France – une étude empirique
Date début	1er mars 2009 **Date fin** 31 décembre 2009
Etablissement porteur	Université Charles de Gaulle Lille 3
Coordinateur projet	Joachim Schöpfel
Coordonnées	Laboratoire GERiiCO, Domaine Universitaire du Point de Bois, BP 60149, 59653 Villeneuve d'Ascq Cedex Email: joachim.schopfel@univ-lille3.fr Tél.: 0320 416 153 / 0688 350 147
Partenaires	Université Charles de Gaulle Lille 3 Laboratoire GERiiCO équipe Savoirs, Information, Document (SID) Chérifa Boukacem-Zeghmouri (MCF) Email boukacemc@yahoo.fr Tél.: 0620 621 812 Aude Sauer-Avargues (étudiante) Email saueravargues.aude@free.fr Mickaël Malandran (étudiant) Email mickael.malandran@laposte.net Université des Sciences et Technologies de Lille Service Commun de Documentation Julien Roche (directeur) Email julien.roche@univ-lille1.fr Tél. : 0320 434 410 / 0320 337 199 Isabelle Le Bescond Email Isabelle.Le-Bescond@univ-lille1.fr François Lefebvre Email francois.lefebvre@univ-lille1.fr Université d'Amsterdam DAREnet Saskia Woutersen-Windhouwer (Electronic Publishing & Repository Manager) Email S.Windhouwer@uva.nl Tél.: 0031 20 525 Institut de l'Information Scientifique et Technique du CNRS Hélène Prost (chargée de ressources documentaires) Email prost@inist.fr Tél.: 0383 504 600 ou 0672 427 630
Site Web	http://usageao2009.pbworks.com/ Liste de diffusion duao@univ-lille1.fr
Programme de recherche	Bonus Qualité Recherche 2009 Université Lille 3
Coordination programme	Michel Crubellier

Table des matières

1. Introduction 9
2. Méthodologie 13
3. Résultats 15
    3.1. Développement des archives en termes d'offre 15
    3.2. Institutions 16
    3.3. Typologie 17
    3.4. Domaines scientifiques et thématiques 19
    3.5. Évolution des contenus 20
    3.6. Articles 21
    3.7. Littérature grise 22
    3.8. Données scientifiques 24
    3.9. Autres contenus 24
    3.10. Absence de documents en texte intégral 25
    3.11. Langue du site et application 25
4. Discussion 27
    4.1. Développement des sites et dépôts 27
    4.2. La part de la littérature grise 28
    4.3. Comment identifier une archive ouverte ? 28
    4.4. A propos de la définition d'une archive ouverte 29
    4.5. La collecte d'information sur le Web 32
    4.6. Le système HAL 32
    4.7. Représentativité et obligation 33
5. Conclusion 35
6. Bibliographie 37
Annexe A – Publications 39
Annexe B – Répertoires de référence 41
Annexe C – Sites analysés 43
Annexe D – Données d'analyse 47

Liste des tableaux

Tableau 1 : Développement des archives ouvertes en France (2008-2009) 15
Tableau 2 : Années de création des archives ouvertes en France 15
Tableau 3 : Nombre d'archives par type d'institutions (2008-2009) 16
Tableau 4 : Nombre de dépôts (items) par type d'institutions (2008-2009) 16
Tableau 5 : Année de création et type d'institutions 16
Tableau 6 : Typologie des archives ouvertes en France (2008-2009) 17
Tableau 7 : Nombre de dépôts (items) par typologie d'archives (2008-2009) 17
Tableau 8 : Nombre de sites et d'items, par type d'archives et d'institutions (2009) 18
Tableau 9 : Les archives dédiées à un type de document (2008-2009) 18
Tableau 10 : Nombre d'archives par politique de dépôt (2008-2009) 18
Tableau 11 : Les disciplines scientifiques des archives ouvertes en France (2008-2009) 19
Tableau 12 : Type d'institutions et disciplines scientifiques des archives (2009) 19
Tableau 13 : Type d'archives et disciplines scientifiques (2009) 20
Tableau 14 : Les cinq premiers sites en nombre d'items (2008-2009) 20
Tableau 15 : Nombre d'articles dans les archives ouvertes (2009) 21
Tableau 16 : Nombre d'articles dans HAL par rapport au WoS (revues avec comité de lecture) 22
Tableau 17 : Nombre d'articles dans HAL par rapport au WoS (revues avec comité de lecture et audience internationale) 22
Tableau 18 : Nombre d'articles dans HAL par rapport à SCOPUS (revues avec comité de lecture) 22
Tableau 19 : Présence de la littérature grise par type d'archives (2009) 23
Tableau 20 : Types de littérature grise dans les archives (2009) 23
Tableau 21 : Thèses et mémoires dans les archives ouvertes, par type d'archives (2009) 24
Tableau 22 : Thèses et mémoires dans les archives ouvertes, par taille d'archives (2009) 24
Tableau 23 : Thèses et mémoires dans les archives ouvertes, par type d'institutions (2009) 24
Tableau 24 : La langue des sites en 2009 25

Synthèse

La France figure parmi les pays fortement engagés dans le mouvement vers l'accès libre à l'information scientifique, par le biais de la communication scientifique directe, c'est-à-dire la mise en place d'archives ouvertes sur Internet et la création de revues gratuites en ligne. Néanmoins, à ce jour et contrairement à d'autres pays, il n'existe que peu d'études empiriques sur les résultats de cet investissement public. Notre étude tente de contribuer à l'évaluation du développement des archives ouvertes en France. La première partie du rapport final fournit des éléments sur la typologie, la taille, le contenu et le développement des archives ouvertes. Leur utilisation par les communautés scientifiques fera l'objet de la 2e partie du rapport.

L'étude s'appuie sur une enquête menée en 2009 sur un échantillon quasi-exhaustif des sites français, à partir d'information et de données publiées et/ou disponibles en ligne, sur chaque site. L'échantillon a été construit à partir de 19 sites de références français et internationaux (répertoires, annuaires, listes…). La collecte d'information a été réalisée à partir d'une grille regroupant 58 critères d'analyse.

Le nombre d'archives ouvertes et de leurs dépôts connaît en 2009 une croissance significative. Le nombre de sites répertoriés passe à 150, avec presque de 1,9 millions items (documents, métadonnées sans texte intégral, matériel audiovisuel, données scientifiques). La plupart des sites (54%) sont des archives institutionnelles ; la majeure partie des dépôts (56%) se trouvent dans des sites relevant des organismes de recherche (CNRS, INRA etc.).

81 sites affichent une politique incitative ; seulement 16 sites se positionnent avec une contrainte forte (obligation de dépôt). 18 sites poursuivent une politique patrimoniale, avec une mise en ligne de collections voire de matériels plus anciens.

29% des archives, dont HAL, ont un caractère plus ou moins multidisciplinaire. 31% font partie des SHS. Le reste (40%) contient un ou plusieurs domaine(s) STM.

Au moins 35% de l'ensemble des dépôts (plus de 650 000) sont des articles publiés dans des revues. Les articles déposés dans HAL représentent autour de 10% de la production scientifique française.

Trois sites sur quatre dont surtout les archives institutionnelles contiennent des documents à caractère non-commercial. La part de cette « littérature grise » augmente de 11% à 18% pour atteindre plus de 300 000 thèses, communications, rapports etc. en 2009. Paradoxalement, 78% des thèses se trouvent dans les archives des organismes de recherche et pas dans celles des universités.

Les données scientifiques (datasets) représentent déjà 7% du contenu des archives ouvertes. D'autres catégories (cours, vidéo, enregistrements sonores…) sont pour l'instant au moins négligeables. Par contre, 16% des dépôts ne contiennent que les métadonnées, sans texte intégral.

La discussion porte entre autre sur l'identification et la définition des archives ouvertes, sur la méthodologie et sur le système HAL.

Le projet vient renforcer et compléter les projets en cours de l'équipe Savoirs, Information, Document (SID) du laboratoire GERIICO (université Lille 3), en termes d'objets d'études, de méthodologie et de résultats. Il soutiendra le développement du positionnement et des partenariats du laboratoire GERIICO au niveau national et européen. A ce jour, il a donné lieu à plusieurs communications et publications (cf. annexe A).

1. Introduction

Dans l'environnement du mouvement vers l'accès libre à l'information scientifique, les revues gratuites en ligne et les archives ouvertes sont devenues en quelques années une partie significative du paysage de la recherche, estimée à 15-20% de la production scientifique (cf. Aubry & Janik, 2005; Willinsky, 2006; Lutz, 2009). Institutions, organisations et gouvernements investissent dans la mise en place, l'infrastructure, la gestion et la maintenance de ces nouveaux outils, car l'accès libre à l'information scientifique, la communication rapide, directe et non restrictive entre chercheurs ne sont pas seulement utiles à la recherche mais sont devenus un enjeu politique d'envergure. La Commission Européenne soutient le développement d'une infrastructure qui facilite la libre circulation de l'information, la 4^{e} liberté de l'espace de l'Union et condition indispensable de la société de l'information.

Néanmoins les études empiriques sur l'impact réel de ce mouvement, en termes de développement et surtout d'utilisation, sont plutôt rares. Certains chercheurs considèrent les archives ouvertes comme "*napster* de la recherche", d'autres ne les connaissent pas du tout. Il n'y a que peu d'études chiffrées sur le développement des archives ouvertes en France, et leurs résultats sont partiels, divergents et non-exhaustifs.

Pierre Baruch a publié en 2007 une "revue, à l'intention d'un public érudit, mais non forcément au courant, des principales voies retenues pour permettre l'accès libre aux documents scientifiques présents sur Internet" dont l'annexe contenait une liste de 40 archives ouvertes, classée par logiciel, institution et type de documents.

De son côté, le consortium COUPERIN a mené une enquête en avril 2007 dont les résultats ont été publiés en ligne (Bruley et al., 2007). Parmi les 74 établissements de l'enseignement supérieur et organismes de recherche ayant participé à cette enquête, 53% menaient une réflexion globale sur les archives ouvertes ou avaient un projet en cours. 18% étaient en train d'installer un tel site tandis que 29% avaient déjà une archive ouverte en service. Le rapport de COUPERIN conclut que "le paysage des archives ouvertes (...) réunit de nombreuses caractéristiques typiques d'un nouveau domaine en forte croissance et disposant d'un fort potentiel de développement" (idem).

Cette enquête ne s'adressait qu'à une partie des organismes du "paysage". Ses résultats fournissaient une photographie de la situation à un moment donné (avril 2007) sans pouvoir chiffrer son développement. Une partie des membres de COUPERIN n'a pas répondu, et certaines réponses restent ouvertes à interprétation.

Une autre étude publiée en 2009 a estimé le nombre d'archives ouvertes à 80 (dont 54 à caractère institutionnel), avec presque 500 000 publications en texte intégral (Bertignac & Gac, 2009). Plus récemment encore, le Ministère de la Recherche et de l'Enseignement Supérieur a publié un état de l'art sur les archives ouvertes en France qui évalue leur nombre à 68 (MESR, 2010). Les chiffres restent incohérents, produits à partir d'échantillons non-exhaustifs.

Quant à l'utilisation des archives, l'enquête de COUPERIN donne très peu de résultats. Seulement six projets sur 74 (= 8%) ont produit des statistiques de consultation : "La fréquentation indiquée se situe entre 200 et 130 000 consultations moyennes. Même si le nombre de document présent dans les archives n'est pas forcément très important, le nombre de consultation pour les projets qui ont fourni l'information lui est élevé" (idem). Pourtant, les

enjeux ne manquent pas et les statistiques d'utilisation intéressent l'ensemble des acteurs concernés (cf. Carr et al., 2008). -

Afin de créer un socle solide d'information empirique, nous avons mené une première analyse en 2008 (Schöpfel & Stock, 2009a, b). D'après nos chiffres, le nombre des archives ouvertes en France s'élevait en 2008 (mars-mai) à 56 dont 48% relevaient de l'enseignement supérieur. Le nombre total des dépôts dépassait 700 000 documents, données (observations, résultats d'analyses), notices etc. En 2008, nous n'avons trouvé que peu d'information sur l'usage des archives ouvertes.

Nous avons donc répliqué cette étude en 2009, en affinant la méthodologie et en ajoutant quelques critères d'analyse. Cette 2e phase de notre étude poursuit un double objectif : continuer l'analyse du développement des archives ouvertes en France et faire le lien avec les autres recherches de notre équipe à Lille sur les usages des ressources numériques en ligne.

Quatre hypothèses de travail ont orienté les choix de méthodologie de ce projet.

1. Le nombre d'archives ouvertes et de leurs dépôts connaît en 2009 une croissance significative, par rapport à 2008.

Depuis 2007, on assiste sur le plan international à une forte augmentation de l'information scientifique en accès libre[1]. En France, les archives ouvertes avaient encore en 2007 globalement une taille plutôt faible, avec un pourcentage important de métadonnées sans texte intégral (Bruley et al., 2007). Néanmoins, les auteurs de l'enquête de 2007 arrivaient déjà à la conclusion que "le nombre de projets en phase de démarrage (la moitié) laisse prévoir un accroissement rapide du secteur. Le nombre de dépôts est encore faible, peu de projets sont d'ores et déjà opérationnels et l'on ne peut pas considérer que la masse critique ait été atteinte. Et pourtant la demande est présente, témoignant du fort potentiel de rayonnement des établissements à travers les diverses initiatives d'archives ouvertes (…)." Notre étude de 2008 a confirmé ce pronostic (Schöpfel & Stock, 2009b). La première hypothèse s'appuie donc sur l'idée que l'évolution en France suivra la courbe des autres pays. Une augmentation renforcée par la nouvelle politique active des universités et la prise de conscience du potentiel de ces sites.

2. La part de la littérature grise est plus importante.

L'étude de 2008 a montré la part non négligeable des documents à caractère non-commercial (= littérature grise) contenus dans les archives ouvertes (Schöpfel & Stock, 2009a). Qu'en est-il en 2009 ? Comme l'objectif principal d'une archive ouverte est de renforcer la visibilité de la production académique, il est probable que la proportion de la littérature grise soit plus importante en 2009.

3. Il y a davantage d'information sur l'utilisation de ces archives, en termes d'accès (visites) et de téléchargements.

En 2008, il n'y avait pratiquement pas de données sur l'utilisation réelle des archives ouvertes. Ce n'est pas une particularité française[2]. Notre deuxième hypothèse part donc du principe que généralement, la situation n'a pas bougé depuis 2008 mais compte tenu du

[1] Cf. le suivi des revues en accès libre et des archives ouvertes assuré par H. Morrison avec les chiffres pour le premier trimestre 2009 http://poeticeconomics.blogspot.com/2009/03/dramatic-growth-of-open-access-march-31.html

[2] Bath (2009) a mené une enquête sur des archives ouvertes dans un domaine particulier (informatique, mathématiques) ; une seule archive sur neuf fournit des statistiques d'accès aux utilisateurs (Caltech Computer Science Technical Reports http://caltechcstr.library.caltech.edu/) ; quatre le font sur demande, quatre autres ne proposent rien. Il faut s'attendre à trouver la même situation en France.

nombre de projets et d'initiatives nouvelles, il y a une certaine probabilité de trouver davantage de statistiques d'utilisation qu'avant.

4. L'information sur l'utilisation n'est ni exhaustive, ni normalisée.

Même constat. Nous n'avons pas connaissance d'initiatives dans le paysage français en faveur d'une généralisation ou d'une normalisation de ces statistiques. Nous partons donc dans l'expectative de trouver des données hétéroclites, non comparables, non représentatives. D'après l'expérience dans d'autres pays, certains chiffres sont diffusés sur le Web, d'autres par liste de diffusion, dans un rapport, plus rarement dans un article ou une communication. -

Le projet tente de fournir des réponses à ces quatre hypothèses. D'autres questions se posent, sans que nous ayons de véritables hypothèses (cf. Smith, 2008) : Quel est le bénéfice réel d'une archive institutionnelle (IR), et comment le mesurer ? Les archives institutionnelles vont-elles devenir un jour le modèle normal des publications universitaires ? Comment s'articuleront-elles avec les archives thématiques qui actuellement se développent mieux ? Quelle est la probabilité d'une obligation de dépôt (mandat obligatoire), et quel impact aurait une telle politique ? De quelles métadonnées les archives institutionnelles auraient-elles besoin, et quels sont leurs liens avec d'autres systèmes d'information ? Puis, quels seront leurs services aux utilisateurs ?

2. Méthodologie

Notre analyse s'appuie sur une adaptation de la méthodologie développée pour l'étude en 2008. Comme en 2008, nous avons sélectionné les archives ouvertes à partir d'un choix de sites Web de référencement (répertoires), au lieu de procéder à une recherche directe sur le Web ou de mener une enquête auprès des institutions. L'analyse porte donc uniquement sur des sites référencés et répertoriés, validés par d'autres comités professionnels ou scientifiques, avec un rayonnement et une visibilité nationale et/ou internationale.

Cette approche correspond à d'autres études, comme par exemple celle de Bueno-de-la-Fuente et al. (2009) sur les métadonnées dans les archives ouvertes.

Aux répertoires de 2008[3], nous avons ajouté onze autres sites dont le wiki « Archives Ouvertes » des URFIST, la liste des archives ouvertes de la Mission IST du MESR, le site du Groupe de Travail Archives Ouvertes du consortium COUPERIN et la liste sur le site du CCSD. La liste des 19 répertoires de référence figure en annexe B.

En recoupant ces répertoires, nous avons établi une liste d'archives ouvertes. Pour chaque entrée nous avons vérifié l'URL, la localisation en France et la présence de dépôts récents. L'échantillon ainsi obtenu – 151 sites - se trouve en annexe C.

Cette sélection de critères correspond à la spécificité de notre projet. D'autres choix restent possibles ; Bath (2009) par exemple, dans sa recherche sur des archives ouvertes dans le domaine informatique et technologie de l'information, applique quatre critères de sélection : le domaine, la langue (anglais), la taille (minimum 100 documents), et l'alimentation (vivant).

Nous n'avons pas non plus, d'emblée, voulu limiter l'analyse à des sites avec dépôt de publications par les auteurs.

Chacun de ces sites a été caractérisé selon 58 critères d'une grille d'analyse (cf. annexe D). Ces critères se répartissent comme suit :

Information générale : 8 critères (nom, acronyme, URL, institution etc.)

Information spécifique : 13 critères (type d'archive, contenu, logiciel, taille etc.)

Contenu : 26 critères (littérature grise, rapports, articles etc.)

Données qualitatives : 7 critères (politique, métadonnées, validation etc.)

Commentaires : 4 critères (date de l'enquête, responsable de la collecte, commentaires)

Ceci correspond à peu près au procédé d'une autre enquête (Bath, 2009) qui a regroupé les critères d'analyse en sept catégories :

1. Background (4) - nom, objectif, gestion
2. Resources (4) - RH, logiciel, financement
3. Content management policies (13) - matériel, métadonnées, formats, dépôts, accès...
4. Preservation policies (3) - content migration, sauvegardes
5. Rights management (2) - accord avec éditeurs, responsabilité pour l'aspect juridique (institution ? auteur/dépositaire ?)

[3] Parmi ces répertoires figuraient notamment Eprints, OpenDOAR et ROAR.

6. Promotion and services (2) - accès via site institutionnel, services complémentaires (numérisation, assistance...)

7. Feedback (2) - commentaires des utilisateurs, statistiques d'usage

Les résultats ont été intégrés dans une base de données, vérifiés, validés et le cas échéant, modifiés et/ou complétés par une ou deux personnes avant l'analyse statistique ou qualitative. Cette base de données contient avec 8700 entrées quatre fois plus d'information que celle de 2008 (1960 entrées). Trois types d'analyse sont effectués : une analyse statistique des données chiffrées, une analyse comparative entre les données 2008 et 2009, et une analyse qualitative.

Pour toutes les archives, nous avons cherché des données d'usage (accès en ligne). Suivant l'approche du réseau DINI (cf. Henneberger, 2009), nous avons choisi de nous référer - là où c'était possible - sur une évaluation à partir des téléchargements de documents (download) comme meilleure mesure d'impact.

La collecte des données en ligne a été faite en équipe, par quatre personnes différentes et avec un contrôle mutuel.

L'étude s'est déroulée en quatre étapes :

* Choix des sites de référencement : mai 2009

* Sélection des archives ouvertes : mai-juin 2009

* Collecte des données des archives ouvertes : juillet-octobre 2009

* Validation et exploitation des données : octobre-décembre 2009

Une partie de l'exploitation statistique des données empiriques a été effectuée entre janvier et mars 2010.

Le planning de la 2^{e} thématique du projet, l'utilisation des archives, sera décrit dans la 2^{e} partie du rapport.

Cette enquête s'insère dans une étude longitudinale sur trois ans (2008-2011).

3. Résultats

Nous présentons ici l'évolution du paysage des archives ouvertes en France entre 2008 et 2009. Les résultats concernant leur utilisation font l'objet de la 2e partie du rapport.

3.1. Développement des archives en termes d'offre

La comparaison des résultats entre 2008 et 2009 montre un spectaculaire accroissement de l'offre, tant qu'en nombre d'archives qu'en nombre de documents. En 2008, 56 archives sont recensées; en 2009, s'y ajoutent 94 autres sites. On dénombre 703 178 items en 2008 ; puis 1 878 520 en 2009.

	2008	**2009**
Sites	56	150
Contenu	703 178	1 878 520

Tableau 1 : Développement des archives ouvertes en France (2008-2009)

Ce développement est significatif : en un an, le nombre de sites a progressé de +174%, celui de leur contenu de +167%.

Nous avons relevé l'année de création et mise en place des sites. Pour 58 sites, il était impossible de déterminer cette date avec précision. Voici le tableau pour les autres :

1995	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
1	1	3	3	4	6	6	5	15	13	11	16	8
8				21				39			24	
9%				23%				42%			26%	

Tableau 2 : Années de création des archives ouvertes en France

On constate une lente progression depuis le milieu des années 90, puis une accélération après la déclaration de Berlin et l'accord entre l'enseignement supérieur et la recherche sur le développement des archives ouvertes en France. Avec un peu de recul, on pourrait diviser cette courte histoire des archives ouvertes en quatre périodes :

1. **1995-2000 : Les précurseurs**. Création des premiers sites à caractère institutionnel ou scientifique comme CITHER, CNUM-CNAM et LACITO Archive.

2. **2001-2004 : Le développement**. Une première prise de conscience politique avec une accélération du mouvement, autour d'archives à caractère national comme HAL, TEL, NUMDAM, Cyberthèses, EduTICE et ArchiveSIC.

3. **2005-2007 : L'essor**. Mise en place de plus en plus d'archives ouvertes, surtout institutionnelles. Parmi ces sites figurent Archimer (IFREMER), PERSEE (SHS), LARA (rapports) et plusieurs des archives institutionnelles de HAL (INSERM, IRD, Pasteur).

4. **2008-... : La consolidation.** Après une forte croissance, on constate un certain ralentissement de la mise en place de nouveaux sites qui n'est pas nécessairement définitif. On

repère plusieurs projets de taille, objectif, architecture et contenu assez différents : SPIRE (Sciences Po Paris), OATAO (Toulouse) ou encore DUMAS (mémoires).

3.2. Institutions

La plupart des différentes institutions qui ont en charge la gestion des archives appartiennent soit à la recherche publique, soit à l'enseignement supérieur.

	Sites en 2008	Sites en 2009
Enseignement supérieur	27	61
Recherche	26	66
Autres	2	23

Tableau 3 : Nombre d'archives par type d'institutions (2008-2009)

La même analyse en termes de contenu (dépôts) donne une toute autre image. En taille relative (nombre de dépôts par site), les archives de l'enseignement supérieur ont même reculé de 10% tandis que celles de la recherche publique ont progressé de 3%. Autrement dit, le nombre des archives ouvertes universitaires s'est développé au détriment de l'alimentation; enfin, l'alimentation n'a pas suivi la mise en place des nouveaux sites.

	Contenu en 2008	Contenu en 2009
Enseignement supérieur	260 358	534 122
Recherche	405 319	1 057 435
Autres	37 501	286 370

Tableau 4 : Nombre de dépôts (items) par type d'institutions (2008-2009)

Plus de 56% des ressources sont archivées dans des sites relevant des organismes de recherche publique, et seulement 28% appartiennent aux sites de l'enseignement supérieur. Le rapport entre les deux secteurs est de 66:33 en faveur de la recherche publique ; en 2008, il a été de 60:40. Ces chiffres reflètent le poids du CCSD avec le serveur HAL, mais aussi la production scientifique plus importante des EPST et un certain retard des universités à mettre en place et surtout, alimenter des archives institutionnelles.

L'étude des années de création et mise en place confirme ce développement différent.

	n.d.	1995-2000	2001-2004	2005-2007	2008-2009
EnsSup	16	2	6	17	17
		5%	14%	40%	40%
Recherche	28	4	12	20	4
		10%	30%	50%	10%

Tableau 5 : Année de création et type d'institutions

Les sites issus des organismes de recherche ont été mis en place plus tôt et plus vite, tandis que le secteur de l'enseignement supérieur commence à rattraper son retard depuis 2-3 ans seulement.

3.3. Typologie

Une 2e analyse s'appuie sur une typologie publiée par Armbruster & Romary (2009) qui proposent quatre catégories d'archives ouvertes.

Une archive est institutionnelle si elle regroupe les différentes publications d'une même institution ; au nombre de 31 en 2008, puis 81 en 2009, ce type est le plus représenté dans l'échantillon de 2009.

Une archive est thématique (subject-based) si sa caractéristique principale est de regrouper des documents se rapportant à un sujet commun ; nous en avons identifiées 13 et 19, respectivement en 2008 et 2009.

D'autres sites ont une vocation nationale, supra-institutionnelle sans être liée à une discipline particulière. C'est par exemple le cas d'un site comme HAL pour l'ensemble des publications ou TEL, pour les thèses.

	Sites en 2008	**Sites en 2009**
Archives institutionnelles	31	81
Archives thématiques	13	19
Archives à vocation nationale	7	11
Archives à caractère scientifique	4	23
Autres, non spécifié	1	16

Tableau 6 : Typologie des archives ouvertes en France (2008-2009)

Parmi les sites à caractère national, on trouve des archives comme LARA, I-Revues ou Revues.org.

Parmi ceux à caractère scientifique, on trouve des sites comme Symposcience qui contient des colloques en ligne de plusieurs organismes (CIRAD, IFREMER, CEMAGREF, INRA) ou les Bibliothèques Virtuelles Humanistes de l'université de Tours.

	Contenu en 2008	**Contenu en 2009**
Archives institutionnelles	202 338	474 834
Archives thématiques	301 339	737 641
Archives à vocation nationale	194 241	428 528
Archives à caractère scientifique	5 269	236 924

Tableau 7 : Nombre de dépôts (items) par typologie d'archives (2008-2009)

Quand on compare ces chiffres, on constate, d'une part l'importance du nombre des archives institutionnelles mais en même temps, l'importance prépondérante (et grandissante) des sites à vocation nationale et à caractère thématique. Non seulement la taille moyenne des archives institutionnelles a régressé (cf. plus haut) mais en plus, celle des autres types est entre deux fois (archives de recherche) et six fois (sites nationaux et thématiques) plus importante.

En croisant les deux critères - type d'institutions et type d'archives - nous obtenons un tableau intéressant :

		Institutionnel	Thématique	National	Scientifique
EnsSup	**Sites**	50	7*	0	3
	Items	113 555	390 794	0	29 773
Recherche	**Sites**	36	18**	6***	8
	Items	332 943	328 772	206 309	161 117

* dont PERSEE et Cristallography

** dont NUMDAM, HAL-SHS et INHA

*** dont HAL

Tableau 8 : Nombre de sites et d'items, par type d'archives et d'institutions (2009)

Les sites institutionnels sont plus répandus dans l'enseignement supérieur.

Les sites à caractère national sont (pour l'instant au moins) du ressort des organismes de recherche, avec en particulier HAL.

La taille moyenne des archives ouvertes des organismes de recherche est de deux à quatre fois plus importante que celle des sites de l'enseignement supérieur - à l'exception des sites thématiques. Ceci s'explique par les deux grands sites PERSEE et Cristallography hébergés tous les deux par des universités.

Certaines archives contiennent un seul type de document, des thèses, rapports etc. Au nombre de quatre en 2008, on en compte 20 en 2009.

Une archive peut à la fois être institutionnelle et ne contenir que les thèses soutenues dans cette institution.

Archives dédiées à un type de document	2008	2009
Nombre de sites	4	20
Nombre d'items	34 101	50 975

Tableau 9 : Les archives dédiées à un type de document (2008-2009)

Comparée aux autres types, cette catégorie est de moindre importance.

Nous avons également défini les archives selon la politique de dépôt : les trois principales sont, l'incitation pour 81 archives en 2009, l'obligation pour 16 archives ou la politique patrimoniale pour 18 sites. D'autres archives sont régies par une politique mixte : patrimoniale et incitative pour IRIS, incitative ou obligatoire selon les établissements regroupés sous l'archive OATAO[4], etc. Certains sites n'affichent pas de politique de dépôt ou d'alimentation.

	Sites en 2008	Sites en 2009
Incitation	26	81
Obligation	12	16
Patrimonial	7	18
Mixte ou indéterminé	11	35

Tableau 10 : Nombre d'archives par politique de dépôt (2008-2009)

[4] Archive institutionnelle de l'Institut National Polytechnique de Toulouse et de l'Ecole Vétérinaire de Toulouse

Parmi les institutions qui affichent une politique d'incitation forte voire d'obligation se trouvent trois organismes de recherche et 13 universités. A l'exception du site Cemadoc (CEMAGREF), il ne s'agit pas d'archives très importantes en termes de dépôts; par contre, elles affichent une croissance deux fois plus forte que la moyenne (+329%).

En ce qui concerne le signalement des documents, l'interface la plus répandue est le dépôt (103 sites) où le chercheur peut signaler ses publications de sa propre initiative ou aidé par les gestionnaires du site. Ce modèle reflète le principe de base d'une archive ouverte.

3.4. Domaines scientifiques et thématiques

L'étude de 2008 avait classé les sites selon la ou les discipline(s) scientifique(s) de leurs contenus en trois catégories : multidisciplinaires, sciences humaines et sociales (SHS) et autres. Cette dernière catégorie recouvre tous les sites avec des contenus "sciences, techniques et médecine" (STM), sciences appliquées (sciences de l'ingénieur) ou mixte (sciences de la vie et SHS etc.). Voici l'évolution 2008-2009 :

	2008	2008 en %	2009	2009 en %
SHS	18	32%	46	31%
multidisciplinaire	18	32%	44	29%
autres	20	36%	60	40%

Tableau 11 : Les disciplines scientifiques des archives ouvertes en France (2008-2009)

La répartition des domaines est restée relativement stable. La part des sites SHS a légèrement reculé de 32% à 31%. Parmi ces sites, on trouve HAL-SHS, PERSEE et l'INHA. Mais des ressources SHS ont été déposées dans 70 archives (= 47%).

Phénomène identique pour les sites clairement identifiés comme multidisciplinaires, comme par exemple TEL, HAL et une série d'archives institutionnelles. Leur part a reculé de 32% à 29%.

Par contre, la part des sites couvrant un domaine STM ou plusieurs disciplines sans être vraiment multidisciplinaires a progressé de 36% à 40%. On trouve ici les sciences de l'ingénieur avec 41 sites dont 17 dédiés. Quant aux sciences de la vie, elles se retrouvent dans 27 sites mais seulement cinq leur sont explicitement consacrés - l'absence d'une grande archive française comparable à PubmedCentral reste un des traits caractéristiques du paysage français.

Si on compare le domaine scientifique avec l'institution et le type des archives, on obtient les résultats suivants:

	EnsSup	Recherche
SHS	18	23
Multidisciplinaire	24	11
Autres	18	29

Tableau 12 : Type d'institutions et disciplines scientifiques des archives (2009)

La part des sites avec un ou plusieurs domaine(s) clairement défini(s) est plus élevé pour les organismes de recherche (83%) que pour l'enseignement supérieur (60%). Tous les sites multidisciplinaires de l'enseignement supérieur sont des archives institutionnelles.

	Institutionnel	Thématique	National	Recherche
SHS	20	17	2	7
Multidisciplinaire	29	0	9	6
Autres	37	13	0	10

Tableau 13 : Type d'archives et disciplines scientifiques (2009)

La majeure partie des sites thématiques relève des sciences humaines et sociales. Les sites à vocation nationale par contre se positionnent majoritairement comme multidisciplinaire même si ce positionnement est relativisé par la réalité des dépôts mais aussi par l'historique notamment du plus grand site, HAL.[5]

3.5. Évolution des contenus

Entre 2008 et 2009, le nombre d'items contenus dans les archives ouvertes augmente de 167%. Voici les 5 sites les plus importants en terme de dépôts en 2008 et en 2009.

	Acronyme	Nom	Items 2008
1	PERSEE	Revues Scientifiques en Sciences Humaines et Sociales	172 215
2	HAL	Hyperarticle en Ligne	108 590
3	ProdINRA	Base de données des publications de l'INRA	100 000 (appr.)
4	COD	Cristallography Open Database	70 295
5	IRD	IRD Horizon Pleins Textes	68 519
	Acronyme	**Nom**	**Items 2009**
1	PERSEE	Revues Scientifiques en Sciences Humaines et Sociale	259 816
2	Gallica	Gallica	215 422
3	HAL	Hyperarticle en Ligne	143 341
4	ProdINRA	Base de données des publications de l'INRA	118 543
5	COD	Cristallography Open Database	110 210

Tableau 14 : Les cinq premiers sites en nombre d'items (2008-2009)

Ce classement a deux particularités. D'une part, la présence d'une archive de données (*datasets*) sans documents (COD), précurseur, peut-on dire, de la cyberinfrastructure de

[5] 49% des documents de HAL relèvent des sciences physiques, 19% de l'informatique et 16% des sciences humaines et sociales ; mais seulement 5% sont indexés dans les sciences du vivant, 1% en chimie et 1% dans les sciences de l'environnement. La médecine humaine avec 850 dépôts compte pour 0,6%. Ces pourcentages ne sont pas représentatifs de la production scientifique française mais fortement biaisés en faveur des sciences physiques, l'informatique et les sciences humaines et sociales. A titre comparatif, les pourcentages issus de la base SCOPUS (2008) : sciences physiques 12%, informatique 7%, SHS 3%, sciences du vivant 23%, chimie 8%, sciences de l'environnement 4% et médecine 24%.

recherche. L'autre particularité est la présence de deux sites référencés comme archive ouverte mais qui n'ont pas la vocation d'une communication scientifique directe : Gallica de la BNF d'une part (avec surtout des documents et images numérisés tombés dans le domaine public), et PERSEE, l'archive numérique des revues en SHS.

Nous pouvons distinguer trois grandes catégories de dépôts : des articles, de la littérature grise et des données scientifiques (*datasets*). Mais il est difficile de préciser quel genre de document ou support les archives ne contiennent pas. Tout y est. Voici une liste non exhaustive issue de notre enquête :

Revue, article, livre, chapitre, encyclopédie, dictionnaire, norme, brevet, thèse, HDR, mémoire, rapport, conférence, communication, poster, actes de colloque, cours, séminaire, annales, prépublication, travail non publié, note, carnet de terrain, manuscrit, lettre, épistemon, fac-similé, schéma, plan, carte, médaille, monnaie, herbier, photo, image, affiche, croquis, vidéo, enregistrement sonore, ressource linguistique informatisée, document oral, partition, logiciel, données, objet.

L'identification de ces différents contenus n'est pas toujours facile. Souvent la précision des métadonnées pour identifier les contenus reste de moins bonne qualité qu'un catalogage traditionnel même si la formule "métadonnées + référentiel = qualité" fait partie du concept de HAL. Mais l'enjeu de ces référentiels est souvent ailleurs, notamment pour les auteurs et les institutions afin d'utiliser les archives pour l'évaluation de la production scientifique.

3.6. Articles

Quel est le nombre d'articles déposés dans les archives ouvertes en France ? Voici une estimation à partir de nos résultats :

	2009		**en %**
Articles dans archives ouvertes	654392		35%
Dont publiés en 2007		*29 638*	*5%*
dont publié en 2008		*25 357*	*4%*

Tableau 15 : Nombre d'articles dans les archives ouvertes (2009)

D'après ce tableau, les articles représenteraient seulement 35% des publications déposées. Mais il est assez difficile de déterminer avec précision le nombre d'articles déposés dans les archives ouvertes. Pour 72 sites (= presque la moitié), nous ne disposons tout simplement pas d'information. Certains sites comme les sites pour les thèses, mémoires ou rapports, n'en contiennent pas ; pour d'autres, il était impossible d'obtenir cette information (absence de métadonnées et/ou fonctionnalité de recherche ou de consultation). Ce chiffre est donc probablement sous-estimé. -

Essayons donc une autre estimation, à partir des articles dans HAL. Quelle partie de la production scientifique française représentent-ils ? Nous avons mené cette analyse sur trois années (2006-2008) en comparaison avec les deux bases de références, le Web of Science (chiffres publiés par l'OST) et SCOPUS (statistiques générées par SCImago). Nous avons sélectionné dans HAL tous les articles publiés dans des revues avec un comité de lecture pour être en cohérence avec les critères de sélection du WoS et de SCOPUS. Voici les résultats.

	WoS (OST)*	HAL	% WoS
2006	39 068	5 750	14,7%

* WoS : Rapport 2008 de l'OST

Tableau 16 : Nombre d'articles dans HAL par rapport au WoS (revues avec comité de lecture)

D'après ces chiffres, presque 15% des articles publiés en 2006 dans des revues de qualité se trouvent dans HAL. Nous avons ajouté une 2e analyse, plus restrictive et peut-être, plus réaliste par rapport au contenu de HAL et aux critères d'inclusion très strictes du WoS, en limitant l'extraction de HAL aux seules revues avec comité de lecture **et** avec une audience internationale. Les résultats :

	WoS (OST)*	HAL	% WoS
2006	39 068	3 337	8,5%

* WoS : Rapport 2008 de l'OST

Tableau 17 : Nombre d'articles dans HAL par rapport au WoS (revues avec comité de lecture et audience internationale)

Avec cette contrainte supplémentaire, la représentativité se limite à 8-9%. Une comparaison du même genre avec SCOPUS et sur trois ans révèle des chiffres assez semblables :

	SCOPUS*	HAL	% SCOPUS
2006	73 391	5 750	7,8%
2007	75 908	6 236	8,2%
2008	78 887	5 579	7,1%

* SCOPUS : rercherche avec SCImago

Tableau 18 : Nombre d'articles dans HAL par rapport à SCOPUS (revues avec comité de lecture)

Entre 7 et 8% de la production scientifique française sous forme d'articles publiés de 2006 à 2008 et indexés dans la base SCOPUS se retrouvent dans HAL. Même s'il faut manier ces chiffres avec prudence, l'estimation paraît assez réaliste. Par contre, il est difficile à dire quelle part se trouve dans les autres archives.

3.7. Littérature grise

La part des sites contenant de la littérature grise reste relativement stable, autour de 75% en 2008 et 74% en 2009. Ces 74% en 2009 correspondent à 110 sites. L'analyse par type d'archive montre deux particularités :

Archives institutionnelles	94%
Archives thématiques	37%
Archives documents	80%
Autres	23%

Tableau 19 : Présence de la littérature grise par type d'archives (2009)

D'une part, pratiquement toutes les archives institutionnelles contiennent ce type de document. Il s'agit surtout des communications des enseignants-chercheurs et des thèses des doctorants mais aussi de rapports etc. (cf. tableau 20). D'autre part, quatre sur cinq des archives dédiées à une seule catégorie de contenu ont été créées pour ces documents, surtout pour les thèses. Néanmoins, leur nombre est limité.

Depuis 2008, le rapport du nombre de documents gris à celui des documents textuels (écrits) a augmenté, passant de 11% en 2008 à 18% en 2009.

En terme de chiffre brut : le nombre total de documents gris augmente de 301%, de 79 005 en 2008 à 316 751 en 2009. Alors que les thèses constituaient le type de littérature grise le plus important en 2008 (46%), ce sont les conférences qui arrivent en tête en 2009, en représentant 49% des documents gris.

Type de littérature grise	**Contenu**
Thèses et mémoires	70 488
Rapports	36 186
Conférences	157 257
Working papers (non publié)	6 637
Cours	2 875
Divers	43 308

Tableau 20 : Types de littérature grise dans les archives (2009)

Les cours (*learning objects* dans les archives internationales) sont peu nombreux et correspondent à 0,9% des documents à caractère non-commercial, c'est-à-dire à 0,15% de l'ensemble des dépôts.

Environ 70% des sites proposent des métadonnées spécifiques à ce type de documents, en particulier pour les thèses (date de soutenance, directeur) et les conférences (lieu et nom de la conférence). 41% des documents gris ont été validés avant leur dépôt ou ont "subi" une autre forme de contrôle de qualité (sélection d'une communication par un comité scientifique pour une conférence, soutenance pour une thèse etc.). -

Les thèses et mémoires occupent une place à part dans l'information scientifique. Pièce centrale des travaux universitaires, ces documents constituent en même temps les premières publications scientifiques, diffusées la plupart du temps en dehors de l'édition scientifique traditionnelle, dans le cadre d'un dispositif particulier et spécifique. Voici quelques éléments[6].

91 archives (= trois sites sur cinq) contiennent des thèses ou mémoires. Certains sites comme TEL, Cyberthèses, PASTEL ou MemSIC sont dédiés aux thèses et mémoires et ne contiennent rien d'autre ; d'autres au contraire mélangent ces travaux à d'autres publications.

[6] Pour une analyse détaillée de ce paysage, cf. Paillassard et al. (2007).

C'est le cas notamment des archives institutionnelles qui représentent de loin le type d'archives le plus représentatif pour les travaux des étudiants :

	Nb sites	Nb items
Archives institutionnelles	74	36 365
Archives thématiques	11	2 690
Archives nationales	5	14 737
Archives scientifiques	1	16 660

Tableau 21 : Thèses et mémoires dans les archives ouvertes, par type d'archives (2009)

La taille des sites avec thèses et mémoires est très hétérogène. Mais la plupart de ces sites sont de taille plutôt petite :

Taille en nombre d'items	Nb de sites
< 500	67
500 - 1000	14
1000 - 10000	8
> 10000	2

Tableau 22 : Thèses et mémoires dans les archives ouvertes, par taille d'archives (2009)

Par contre, les thèses et mémoires ne se trouvent PAS majoritairement dans les archives universitaires, du moins au moment de l'étude (2009). Voici la répartition entre les établissements :

	Nb de sites	Nb d'items
Enseignement supérieur	44	15 092
Recherche	41	53 800

Tableau 23 : Thèses et mémoires dans les archives ouvertes, par type d'institutions (2009)

A l'exception de sept sites, toutes ces archives contiennent des métadonnées spécifiques à ce type de publications (établissement et/ou date de soutenance, directeur de thèse etc.).

3.8. Données scientifiques

Nous avons trouvé cinq archives ouvertes avec des données scientifiques (*datasets*). Deux de ces sites contiennent également des ressources documentaires (textuels).

L'échantillon global compte 123 948 *datasets* en 2009, équivalent à 7% du contenu des archives ouvertes en France. Le plus grand réservoir de ces données est la Cristallography Open Database avec 110 210 dépôts. Elle est hébergée par l'université du Maine Le Mans-Laval.

3.9. Autres contenus

Dans notre liste d'archives ouvertes, nous avons trouvé neuf sites avec du matériel audio-visuel, surtout des vidéos mais aussi des enregistrements sonores ; en tout, 6 791 dépôts en accès libre.

3.10. Absence de documents en texte intégral

Une autre particularité des archives ouvertes en France est le pourcentage élevé de dépôts sans texte intégral. 16% des dépôts (items) dans les archives sont des notices bibliographiques (métadonnées), sans accès au texte intégral. Dans 37 archives, la part des notices sans texte intégral est égale ou supérieure à 50%.

3.11. Langue du site et application

A part cinq sites, tous ont une interface en français :

français	français & anglais	multilingue	Anglais
62	77	6	5
41%	51%	4%	3%

Tableau 24 : La langue des sites en 2009

Plus de la moitié des sites ont mis en place des pages en anglais, parfois aussi dans d'autres langues (allemand, espagnol, chinois). Preuve s'il en faut d'une volonté d'ouvrir ces archives à l'international.

Quant aux systèmes d'information, nos résultats soulignent la place importante qu'occupe l'application HAL dans le paysage français (57 sites). Pour le reste, l'information sur les logiciels confirme le constat de Bruley et al. (2007) : "une très grande diversité illustrée par le nombre important d'applications n'apparaissant qu'une seule fois (...)", dont un petit nombre de systèmes sur plusieurs sites (DSpace, EPrints, LODEL).

4. Discussion

Notre étude partait de quatre hypothèses :

1. Le nombre d'archives ouvertes et de leurs dépôts connaît en 2009 une croissance significative, par rapport à 2008.
2. La part de la littérature grise est plus importante.
3. Il y a davantage d'information sur l'utilisation de ces archives, en termes d'accès (visites) et de téléchargements.
4. Cette information n'est ni exhaustive, ni normalisée.

Dans ce premier rapport, nous discuterons les résultats de notre projet par rapport aux deux premières hypothèses, puis nous aborderons quelques problèmes méthodologiques.

4.1. Développement des sites et dépôts

Les résultats de notre étude confirment l'hypothèse. En l'espace d'un an, le nombre des archives ouvertes en France a pratiquement triplé, passant de 56 à 150 sites. Cette augmentation spectaculaire semble surtout portée par les établissements de l'enseignement supérieur ayant pris conscience de l'intérêt et des enjeux de la création d'archives institutionnelles, et par la multiplication des sites hébergés par le CCSD. Elle trouve son reflet dans la croissance des dépôts qui passent de 700 000 à plus de 1,87 millions.

Nos résultats empiriques confirment également le constat par COUPERIN d'un "fort potentiel de développement" et leur pronostic d'une évolution accélérée dans les années à venir (Bruley et al., 2007). Nous y sommes. Les archives ouvertes en France figurent parmi les plus importants sites au monde. En juillet 2009, HAL occupait le 3e rang, HAL-INRIA le 5e rang du Ranking Web of World Repositories (Webometrics). HAL-INRIA était classé première archive institutionnelle au plan international.

A titre comparatif, voici les chiffres du site ROAR (Registry of Open Access Repositories, maintenu par Tim Brody à l'université de Southampton Registry of Open Access Repositories) début juin 2010 :

	France	**Allemagne**	**Royaume-Uni**	**Etats-Unis**
Nb d'archives ouvertes	51	108	165	307
Nb d'items	519 664	1 262 562	2 403 401	8 969 658
Nb de publications 2008*	78 897	103 768	118 831	366 491

* Chiffres extraits du site SCImago basés sur SCOPUS

Tableau 25 : Les archives en France, Allemagne, Royaume-Uni et Etats-Unis (ROAR, juin 2010)

Ces chiffres sont à interpréter avec beaucoup de prudence. Apparemment ROAR sous-estime la réalité en France ; d'après ce site, la France comparée aux autres pays (a) compterait relativement peu de sites par rapport à la production scientifique annuelle (2x moins qu'au Royaume-Uni mais proche des Etats-Unis) ; (b) les sites français auraient une taille moyenne moins importante (mais comparable à l'Allemagne) ; et surtout (c) ils totaliseraient peu de dépôts par rapport à la production annuelle du pays (2-4x moins que les autres pays).

Nos propres chiffres dessinent une autre image ; d'après les résultats de notre enquête, la France se positionne au même niveau que les autres pays, avec un nombre de sites élevé, une taille moyenne de sites comparable aux voisins européens, et autant de dépôts qu'au Royaume-Uni et aux Etats-Unis (et bien plus qu'en Allemagne).

Un rapport récent du Ministère de l'Enseignement Supérieur et de la Recherche (MESR, 2010) semble sous-estimer l'ampleur du phénomène. Nous reviendrons sur cette divergence dans la discussion méthodologique.

4.2. La part de la littérature grise

74% des archives ouvertes contiennent de la littérature grise, c'est-à-dire des publications à caractère non-commercial. Quant aux archives institutionnelles, ce taux s'élève à 93% - pratiquement tous ces sites contiennent des thèses, conférences, travaux non publiés, mémoires etc. Depuis 2008, on compte 200 000 nouveaux documents gris dans les archives ouvertes en France.

Leur part relative à l'ensemble des dépôts textuels a augmenté de 11% à 17%. Ceci confirme donc notre 2^{e} hypothèse selon laquelle ce pourcentage allait augmenter. Avec les projets d'archives institutionnelles et de thèses et mémoires électroniques, cette part risque encore d'accroître dans les mois et années à venir. On est donc loin du précurseur des archives ouvertes, arXiv, dont le but était la communication rapide et directe des pré-publications. D'une certaine façon, la diffusion non-commerciale de l'IST a trouvé un nouveau vecteur de dissémination à travers les archives ouvertes et en particulier, les archives institutionnelles.

4.3. Comment identifier une archive ouverte ?

Le point de départ de notre étude – la sélection d'archives ouvertes à partir de répertoires (annuaires etc.) – induit une certaine hétérogénéité. Il n'existe pas de définition objective et généralement acceptée d'une archive ouverte. Chaque site de référencement, chaque liste ou site de veille applique ses propres critères. Dans certains cas, il s'agit du protocole OAI-PMH. Ailleurs, il suffit d'avoir des documents en accès libre sur une plate-forme.

Nous avons privilégié dès le départ l'exhaustivité et le regard croisé à une sélection stricte. Cette approche exploratoire nous paraissait adaptée à un paysage en transition. Le résultat est qu'on trouve dans l'échantillon aussi bien de « vraies » archives institutionnelles que de sites à caractère patrimonial, 15 bibliothèques numériques ou 12 sites Web mettant en ligne un certain nombre de ressources de tout genre.

N'est-il pas également intéressant de constater que le mouvement du libre accès à l'information, loin d'être monolithique et normalisé, a créé une nouvelle diversité et richesse de sites, de contenus et de services ?

Ceci explique en partie la différence entre nos résultats et les chiffres du rapport MESR dont l'échantillon se limite à deux sources majeures, le répertoire OpenDOAR et la liste des archives sur le site du CCSD. Notre propre étude nous laisse penser que ces deux sites, même pris ensemble, sont loin d'être exhaustifs et qu'il faut aller plus loin si on veut avoir une image plus réaliste sur le développement des archives ouvertes en France.

Un autre inconvénient de notre approche est le décalage entre la mise en ligne d'une archive ouverte et son référencement. Comme en 2008, nous sommes conscients qu'il existait au moment de la collecte des données d'autres archives ouvertes en France qui ne figuraient pas encore dans les répertoires.

4.4. A propos de la définition d'une archive ouverte

L'identification d'une archive ouverte est étroitement liée à la définition de l'objet « archive ouverte ». Mais comme nous l'avons déjà dit, il n'existe pas une définition acceptée mais plusieurs approches. Voici quelques exemples :

« Une Archive Ouverte est un serveur stockant des textes sous version informatique. Il s'agit d'entrepôt d'information, d'archives vivantes, constitué par des articles scientifiques produits par la communauté de chercheurs. Pour ce qui concerne les chercheurs, l'Archive Ouverte leur autorise l'accès libre à leurs publications, via le dépôt de leur production scientifique sur un serveur configuré pour stocker leurs articles, déjà publiés (post publications) ou non (pré-publications), mais également leurs autres travaux de recherche, ainsi que les thèses. »[7]

« (...) le terme archive ouverte désigne un réservoir où sont déposées des données issues de la recherche scientifique et de l'enseignement et dont l'accès se veut ouvert, c'est-à-dire sans barrière. Cette ouverture est rendue possible par l'utilisation de protocoles communs qui facilitent l'accessibilité de contenus provenant de plusieurs entrepôts maintenus par différents fournisseurs de données. »[8]

« Une archive institutionnelle est l'archive d'une institution regroupant l'ensemble de sa production (de recherche, patrimoniale, pédagogique, administrative...) dans des espaces privatifs ou ouverts. »[9]

« Open Access, qui se traduit par Libre Accès, est la libre disponibilité en ligne de contenus numériques. Open Access est principalement utilisé pour les articles de revues ou de recherches universitaires, sélectionnés par des pairs, publiés gratuitement. (…) En ce qui concerne l'Open Access par auto-archivage, aussi appelée la voie « verte » d'Open Access, les auteurs font des copies de leurs propres articles publiés, ouvertement accessible. Ils le font généralement sur une page personnelle ou un dépôt institutionnel ».[10]

« Le terme archive ouverte désigne un réservoir de données issues de la recherche scientifique et de l'enseignement, accessible sur internet et dont l'accès est ouvert. Cette ouverture est rendue possible par l'utilisation de protocoles qui permettent une interopérabilité avec d'autres serveurs. On peut archiver des pré-publications, aussi bien que des publications officielles, tout en respectant certaines conditions des éditeurs (…). » [11]

Les URFIST listent plusieurs éléments constitutifs sans tenter une définition d'une archive ouverte : « traduction de l'expression "open archives", serveur/plateforme permettant le dépôt de documents et leur consultation, notions et outils issus du monde de la recherche, ne recouvre pas la notion française d'"archives" (rétrospectives), diffusion vs conservation, termes en anglais: archive, repository, depository, synonymes: dépôt, réservoir…, "ouvert", mouvement du libre accès : accès libre aux résultats de la recherche scientifique, architecture technique distribuée et interopérable. »[12]

Carr et al. (2008) définissent la finalité d'une archive institutionnelle comme suit: "the aim of institutional repositories has focused on serving the interests of faculty – researchers and

[7] Bibliopédia, http://www.bibliopedia.fr/index.php/Archives_Ouvertes (30 mai 2010)
[8] INIST Libre accès à l'information scientifique et technique, Glossaire http://openaccess.inist.fr/spip.php?mot12 (30 mai 2010)
[9] COUPERIN GTAO http://www.couperin.org/archivesouvertes/spip.php?article104 (30 mai 2010)
[10] http://fr.wikipedia.org/wiki/Open_access (14 juin 2009)
[11] INRIA http://www.inria.fr/publications/archiveouverte/lexique.fr.html (14 juin 2009)
[12] http://urfist-apps.unice.fr/wiki_AO/index.php/Définitions (14 juin 2009)

teachers – by collecting their intellectual outputs for long-term access, preservation and management". Et ils proposent une liste de critères pour « auditer » une archive ouverte.

Armbruster & Romary (2009) précisent trois fonctions fondamentales : "a) the fast and wide dissemination of results; b) the preservation of the record; and c) digital curation for dissemination and preservation."

Smith (2008) expose plusieurs arguments pour la création d'une archive institutionnelle : "Institutions create IRs to keep a wide variety of materials in digital form, such as research journal articles, preprints and postprints, digital versions of theses and dissertations, and administrative documents, course notes, or learning objects."

Plus en amont, la Research Library Group a proposé en 2005 une définition large du terme d'archive numérique, basée sur les travaux d'OAIS (RLG-NARA, 2005) : "Digital repository / Digital archive: These two terms are often used interchangeably. OAIS uses archive when referring to an organization that intends to preserve information for access and use by a Designated Community." -

On pourrait continuer cette liste de définitions. Etablissons plutôt une synthèse des aspects les plus significatifs :

1. Existence d'un serveur comme réservoir de données.
2. Contenus en format numérique (fichiers).
3. Stockage/archivage pérenne.
4. Accès libre sur Internet, diffusion sans restriction.

A ces quatre aspects d'ajoutent d'autres, plus spécifiques et pas partagés par tous les auteurs et projets :

1. Publications, documents, texte.
2. Interopérabilité du site, protocoles communs.
3. Dépôt par auteur ou laboratoire.
4. Lien fort avec l'institution et ses objectifs et besoins.
5. Environnement de service, plate-forme.

Cependant, il reste des questions ouvertes :

Contenu : Il paraît impossible de déterminer une archive ouverte à partir de son contenu. S'agit-il de documents ou de publications ? Mais quid alors des données, résultats de recherche, observations météorologiques, astronomiques etc. ? Si on parle simplement d'information, il devient impossible de distinguer des archives à caractère scientifique d'autres sites à caractère patrimonial ou personnel. S'agit-il uniquement de publications scientifiques ? Mais alors les "datasets" ? Et les enregistrements sonores ou données d'observation, sources d'autres analyses et études ? Et la littérature comme objet d'une recherche scientifique ? Par ailleurs, le site Bibliopédia, après avoir défini les archives ouvertes comme sites où les chercheurs peuvent déposer leurs publications, mentionne un peu plus loin parmi les archives ouvertes "incontournables" des sites comme Gallica, PERSEE, Revues.org et CAIRN. C'est tout à fait illogique - aucun de ces sites ne permet le dépôt individuel ; PERSEE et Gallica sont des sites à caractère patrimonial ; CAIRN est loin d'être un site en accès libre ; et Gallica ne contient pas (que) de publications scientifiques mais plutôt de la littérature. Cependant cette absence de logique "colle" à la réalité, dans la mesure où les différents annuaires,

répertoires et autres sites de référencement recensent, sous l'intitulé d'archive ouverte, toute sorte de sites dont le point commun est surtout l'accès libre aux contenus.

Format : Une archive ouverte doit-elle adopter le protocole OAI-PMH ? Peut-on, doit-on réduire le nombre d'archives ouvertes aux seules archives compatible OAI-PMH ? La réalité est différente, et Armbruster & Romary (2009) ont raison quand ils soulignent l'hétérogénéité des réalisations. Ni format, ni métadonnée, ni logiciel ou système peuvent être d'une grande utilité pour définir une archive ouverte.

Communauté : Une autre approche serait de définir une archive ouverte à partir de sa (ses) communauté(s) - auteurs, dépositaires, hébergeurs, lecteurs-utilisateurs... Appartiennent-ils au secteur scientifique ? Mais dans ce cas, quid de GALLICA ? Où est la frontière entre culture et recherche ?

Diffusion libre : Reste comme dernier critère l'accès libre aux contenus. Un accès non restreint, gratuit, illimité. Mais là encore, la définition se heurte à la réalité. Pour plusieurs raisons[13], les archives contiennent de plus en plus de documents sous embargo et/ou avec une diffusion restreinte, et des notices (métadonnées) sans texte intégral. Michael White de l'université de Stirling (UK) a indiqué que 12% des consultations de leur archive institutionnelle concernaient des documents sous embargo[14]. Dans ce cas, peut-on encore parler d'archive ouverte ?

Nous n'avons pas essayé de sélectionner ou d'éliminer des sites en fonction de tels critères, en acceptant les critères de définition des sites de référencement. Le résultat : parmi les sites répertoriés comme archives ouvertes, on trouve certains sites qui ressemblent plutôt à des bibliothèques numériques, à des sites Web classiques avec mise en ligne de documents, ou encore à des bases de données.

Pour mieux comprendre, voici un schéma descriptif qui permet de distinguer certains sites :

1 IST alimentation courante	1.1 Publications, documents	
	1.2 Données, "information"	1.2.1 Résultats de recherche
		1.2.2 Objets de recherche
2 Patrimoine	2.1 IST	2.1.1 Publications, documents
	2.2 Culture	2.2.1 Objets de recherche

Tableau 26 : Types d'archives en fonction de leur contenu

Parmi les sites répertoriés comme archives ouvertes, nous avons trouvé deux groupes : ceux à caractère scientifique et alimentés régulièrement à partir des dépôts des chercheurs (1), comme par exemple HAL, et ceux dont le contenu a un caractère patrimonial comme Gallica (2).

Le premier groupe se divise en deux catégories, les sites où on dépose des documents, comme la plupart des sites du système HAL ou des archives institutionnelles autres (1.1), et d'autres sites réservés aux données scientifiques (*datasets*), soit comme résultats de la recherche, soit comme leurs objets (enregistrements sonores - ethnologie, linguistique…).

[13] dont notamment l'incitation forte au dépôt, l'objectif d'atteindre vite une masse critique et l'exploitation des archives à des fins d'évaluation scientométrique

[14] message sur la liste JISC-REPOSITORIES du 1 avril 2009

Parmi les sites à caractère patrimonial, on en trouve certains avec du contenu scientifique (PERSEE, IRIS) tandis que d'autres contiennent uniquement des textes littéraires etc. qui sont plutôt objets de recherches scientifiques.

Nous n'avons pas essayé de quantifier ce schéma. Le travail conceptuel et le développement d'une définition à partir de l'évidence empirique devraient se rejoindre afin de clarifier cette situation. Il est certain qu'actuellement le terme d'archive ouverte ou encore plus généralement, d'accès libre à l'information, recouvre une réalité très hétérogène et disparate.

4.5. La collecte d'information sur le Web

La collecte de données sur le Web pose plusieurs problèmes. Quasiment aucun site ne renseigne sur la date de création de l'archive ou sur son évolution chiffrée. Nous avons été confrontés aux différences de qualité de signalement des métadonnées. Décompter les différents types de documents fut parfois un véritable parcours de combattant. Par exemple, l'interface de recherche d'une archive propose comme principal critère le caractère « publié » ou « non publié » d'un document, une autre classe uniquement par auteur l'ensemble des publications ; pour connaître le nombre exact des différents types de documents, la seule solution possible est de les compter « à la main ».

La récolte des statistiques d'usage s'est faite de façon empirique, en recherchant sur le Web. Nous avons découvert une information très hétérogène, des données brutes d'usage en libre accès sur Internet, des diaporamas présentant des données d'usage, des rapports. Il est probable que d'autres éléments appartiennent au domaine de la littérature grise ou se trouvent dans les profondeurs du Web et restent donc pour l'instant inexploités.

4.6. Le système HAL

Nous avons constaté des problèmes liés au système HAL dont le plus étonnant est peut-être la création systématique de doublons :

Le nombre précis des dépôts à un moment donné est difficile à cerner. Les chiffres varient d'un répertoire à l'autre, en fonction de la méthode de comptage (tous les dépôts y compris notices et données primaires, seulement dépôts avec texte intégral, seulement dépôts sans restriction d'accès etc.).

Les chiffres indiqués sur la page d'accueil d'un site HAL ne reflètent pas forcément la volumétrie trouvée dans le cadre de la recherche avancée. Un exemple : au 1er juin 2010, la page d'accueil de HAL indique 278 993 items composés uniquement de la notice bibliographique, alors que la recherche avancée en indique 295 370 (= différence du nombre de documents avec texte intégral et du nombre total d'items). Et on obtient un 3^{e} chiffre à partir du feuilletage par type de documents. Quel est le chiffre fiable ?

D'autre part, selon les différents portails de HAL, l'interface de recherche avancée n'est pas uniforme. Par exemple, dans le cadre du portail de l'INSERM, la recherche avancée ne permet pas de faire la distinction entre les notices bibliographiques seules et les références liées au texte intégral. Le portail de l'INRIA propose en premier lieu les notices avec texte intégral. L'utilisateur doit cocher une case s'il souhaite intégrer les notices bibliographiques dans sa recherche. La recherche avancée dans le portail du CIRAD est construite à l'inverse. L'utilisateur doit préciser s'il souhaite limiter sa recherche aux documents avec texte intégral. Tout cela ne facilite pas l'analyse de l'ensemble du système HAL.

Un autre problème est une certaine imprécision au niveau des métadonnées descriptives, en particulier en ce qui concerne la littérature grise. Pour l'archive nationale HAL, nous avons une image imprécise de la nature des différents documents car un nombre important est classé

sous la catégorie « autres ». A l'inverse, l'archive CemOA du CEMAGREF offre une interface de recherche très fouillée : les critères de recherche permettent de filtrer jusqu'à 32 types de documents différents, classés parmi 14 équipes de recherche. La recherche peut afficher les documents publiés au cours d'une période définie et classer les résultats selon l'ordre chronologique croissant ou décroissant.

4e problème, les doublons. Personne n'empêchera un auteur de déposer son texte dans plusieurs archives afin d'en assurer une plus grande visibilité. Le même document sera donc compté plusieurs fois. A cela s'ajoute une particularité de la plate-forme HAL. Le CCSD a décidé d'y verser systématiquement les dépôts d'autres archives. Ainsi une thèse déposée dans PASTEL ou TEL se retrouvera dans HAL. Un article déposé dans ArchiveSIC sera versé dans HAL-SHS et HAL. Ces doublons gonflent artificiellement le volume de HAL.[15]

Est-il possible de faire une estimation même approximative des dépôts sans doublons ? La manière la plus simple serait d'enlever le contenu de HAL de l'ensemble des dépôts ce qui diminuerait le volume global de 1,87 millions d'items d'environ 7% à 1,73 millions d'items. La réalité se situe probablement quelque part entre ces deux chiffres, dans la mesure où HAL contient des dépôts qui ne sont pas répliqués ailleurs. Concernant le développement global du paysage français, cela ne change pas grand chose.

Nous avons donc préféré compter tous les items dans tous les sites du système HAL, en faisant des doublons.

On rencontre une autre situation à Toulouse où plusieurs universités et grandes écoles ont mis en ligne des archives de thèses (INPT, INSA, UPS). Un site spécifique « Toulouse thèses » moissonne ces dépôts et intègre en plus les thèses de l'école vétérinaire locale. Récemment, le PRES Toulouse a mis en place une nouvelle archive institutionnelle OATAO (Open Archive Toulouse Archive Ouverte) qui recevra surtout les publications scientifiques des chercheurs (articles) mais également des thèses. Une archive unique pour toutes les publications et tous les établissements de l'enseignement supérieur de Toulouse est annoncée pour bientôt. En attendant, il faut vivre (et compter) avec des doublons, comme en 2008 (Schöpfel & Stock, 2009).

4.7. Représentativité et obligation

D'une manière prudente et avec quelques réserves, nous avons estimé la part de la production scientifique dans HAL à 7-10%, sans pouvoir donner une estimation plus précise pour l'ensemble des archives ouvertes en France. Nous rejoignons sur ce point Hélène Bosc (2008) qui avait estimé la représentative de HAL à 10%.

Une politique volontariste qui obligerait les chercheurs à déposer leurs publications dans une archive ouverte (comme c'est déjà le cas dans certaines universités pour les thèses), aurait-elle un impact fort sur le contenu des sites ?

Dans le cadre du *Australian Digital Theses Program*, Sale (2006) a constaté que le taux de dépôt passe de 12% à 100% quelques années après la décision de rendre le dépôt obligatoire pour les thèses de doctorat. Peut-on généraliser sur l'ensemble des publications ?

Nos propres résultats semblent indiquer qu'une archive avec un mandat obligatoire se remplit effectivement plus vite (cf. 4.1.1). Mais il faudrait attendre plusieurs années pour

[15] Une infrastructure comme celle décrite par Bauer (2009) pourrait peut-être améliorer certains aspects, par l'agrégation des données et par la création de services à valeur ajoutée par l'opérateur central. Mais cela reste hypothétique.

évaluer l'impact réel d'une politique contraignante[16]. Pour l'instant, le débat reste ouvert entre ceux qui à l'instar de Romary & Armbruster (2009) défendent un système central tel que HAL, sans contrainte forte mais alimenté (aussi) par les éditeurs eux-mêmes, et ceux qui comme Stevan Harnad plaident pour des archives institutionnelles avec obligation de dépôt.

[16] Même un modèle comme l'université de Liège avec le site ORBI n'atteint pas un dépôt de 100%, malgré une incitation très forte qui lie le dépôt aux promotions et demandes de subvention.

5. Conclusion

Notre analyse reflète l'évolution rapide de l'accès libre à l'information scientifique en France qui se place ainsi au cœur de ce mouvement sur le vieux continent. Les résultats de l'enquête dessinent une image contrastée, avec la particularité d'un système fortement centralisé entouré d'une multitude dynamique d'initiatives locales et sans véritable coordination, malgré plusieurs acteurs qui pourraient jouer ce rôle (CCSD, COUPERIN/GTAO, MESR).

L'usage de ces nouveaux sites fera l'objet de la 2e partie de ce rapport, à paraître au 3e trimestre de 2010.

Nous poursuivrons l'étude empirique du développement des archives ouvertes en France, avec la même méthodologie mais d'une manière simplifiée. Il y aura donc une 3e enquête sur le Web en 2010, puis une 4e en 2011 afin de pouvoir dresser un bilan sur trois ans en 2011 ou 2012. Simplifiée veut dire : la collecte des données se focalisera sur quelques indicateurs importants (institution, type d'archive, contenus...).

De nouveau se posera alors la question de la définition d'une archive ouverte. La part importante des notices bibliographiques nous amène à relire cette définition. L'analyse du contenu des archives révèle une déviation de l'objectif principal d'une archive, la communication directe, rapide et libre entre chercheurs. Préoccupés par un souci d'évaluation de l'impact, certains établissements scientifiques donnent la priorité au signalement plutôt qu'à l'accès direct aux documents.

L'importance des archives à caractère patrimonial est une autre limite au concept d'archive ouverte. Au nombre de 30 en 2009, elles représentent 20% des archives et 46% des documents. Or si initialement les archives ouvertes ont été créées pour faciliter et accélérer la communication scientifique au sein des communautés et disciplines, elles servent aujourd'hui aussi et de plus en plus à la valorisation des collections académiques. Résultat : le mélange entre production scientifique et chargement d'anciens fonds numérisés pose un double problème, celui de la cible et de l'objectif du service, et celui de la pertinence des résultats de recherche dans les archives.

Quelle sera l'évolution des archives ouvertes dans les prochaines années ? Un facteur déterminant sera le développement du système fragmenté des universités et écoles vers des ensembles plus structurés et larges, par le biais des fusions et rapprochements (PRES etc.). A terme, cette évolution pourrait entraîner paradoxalement une diminution du nombre de sites, notamment des archives institutionnelles, mais une augmentation de leur impact et visibilité et aussi, de leur contenu. La mise en place d'une nouvelle infrastructure documentaire, comme le portail national des thèses par l'ABES annoncé pour 2012 ou 2013, contribuera également à cette structuration nécessaire.

Un autre facteur sera la mise en place (ou l'absence) d'une politique contraignante pour le dépôt des publications dans une archive ouverte. Il y a une vive discussion sur le plan international : d'un côté ceux qui plaident pour une « mandatory policy » qui oblige les chercheurs à déposer leurs publications (Stevan Harnad, Hélène Bosc) ou qui la pratiquent déjà (comme Bernard Rentier, le recteur de l'université de Liège) ; de l'autre côté ceux qui (comme Stuart Basefsky de Cornell) considèrent une telle politique improbable et irréaliste, incompatible avec une éthique scientifique, ou qui plaident pour un partenariat avec les éditeurs pour alimenter des archives à vocation nationale (Laurent Romary, Chris

Armbruster). Les résultats du projet PEER apporteront peut-être de nouveaux arguments dans ce débat.

Technologie, marché, choix politiques et investissement sont étroitement liés et créent la dynamique du mouvement de l'accès libre à l'information scientifique. Notre contribution est de fournir quelques éléments d'analyse pour mieux comprendre et agir.

6. Bibliographie

C. Armbruster & L. Romary (2009). `Comparing Repository Types: Challenges and Barriers for Subject-Based Repositories, Research Repositories, National Repository Systems and Institutional Repositories in Serving Scholarly Communication'. *Social Science Research Network Working Paper Series* .

C. Aubry & J. Janik (2005). *Les archives ouvertes, enjeux et pratiques*. Tec Doc.

P. Baruch (2007). `La diffusion libre du savoir. Accès libre et Archives ouvertes'. *L'Archicube* (3):77-95.

M. H. Bath (2009). `Open access repositories in computer science and information technology: an evaluation'. *IFLA Journal* **35**(3):243-257.

B. Bauer (2009). `It's economy stupid! – Anmerkungen zu ökonomischen Aspekten des goldenen und des grünen Weges beim Open Access Publishing'. *Information Wissenschaft & Praxis* **60**(5):271-278.

C. Bertignac & D. Gac (2009). `La voie verte : les archives ouvertes'. In *Open Access Week. 19 October 2009*. Institut Universitaire Européen de la Mer.

H. Bosc (2008). `L'auto-archivage en France : deux exemples de politiques différentes et leurs résultats'. *Liinc em Revista* **4**(2):196-217.

C. Bruley, et al. (2007). `Résultats de l'enquête sur les projets d'archives ouvertes de la recherche dans les établissements du consortium Couperin'. Tech. rep., GTAO Couperin.

G. Bueno-de-la Fuente, et al. (2009). `Study on the use of metadata for digital learning objects in university institutional repositories (MODERI). 2009, 47, 3-4, 262-285.'. *Cataloging & Classification Quaterly* **47**(3-4):262-285.

L. Carr, et al. (2008). `Repository Statistics: What Do We Want to Know?'. In *Third International Conference on Open Repositories 2008, 1-4 April 2008*.

S. Henneberger (2009). `Standardisierte Nutzungsanalysen als alternative Impact-Messungen wissenschaftlicher Publikationen'. In *3. Open-Access-Tage, Konstanz, 7 - 8 October 2009*.

J.-F. Lutz (2009). `Le mouvement pour le libre accès aux publications scientifiques'. In P. Carbone & F. Cavalier (eds.), *Les collections électroniques, une nouvelle politique documentaire*, pp. 75-85. Electre Edition du Cercle de la Librairie.

MESR (2010). `Open Access in France. A state of the Art Report - April 2010'. Tech. rep., Ministère de l'Enseignement Supérieur et de la Recherche.

P. Paillassard, et al. (2007). `Dissemination and preservation of French print and electronic theses'. *The Grey Journal* **3**(2):77-93.

RLG-NARA (2005). `RLG-NARA Audit Checklist for Certifying Digital Repositories'. Tech. rep., Research Library Group - National Archives and Records Administration.

A. Sale (2006). `The impact of mandatory policies on ETD acquisition'. *D-Lib Magazine* **12**(4).

J. Schöpfel & C. Stock (2009a). `Grey literature in French digital repositories: a survey'. *The Grey Journal* **5**(3):147-161.

J. Schöpfel & C. Stock (2009b). `Les archives ouvertes en France – Un potentiel documentaire pour la formation à distance'. *Distances et Savoirs* **7**(4):443-456.

K. Smith (2008). `Institutional Repositories and E-Journal Archiving: What Are We Learning?'. *Journal of Electronic Publishing* **11**(1).

J. Willinsky (2005). *The Access Principle: The Case for Open Access to Research and Scholarship (Digital Libraries and Electronic Publishing)*. The MIT Press.

Annexe A – Publications

J. Schöpfel & H. Prost (2009). `Les statistiques d'utilisation d'archives ouvertes. Etat de l'art'. In *Ressources électroniques académiques: mesures et usages. Colloque international. Lille, 26-27 novembre 2009*.

J. Schöpfel, et al. (2009). `Usage of grey literature in open archives'. In *Eleventh International Conference on Grey Literature: The Grey Mosaic: Piecing It All Together. Washington D.C., 14-15 December 2009.*

J. Schöpfel & C. Boukacem-Zeghmouri (2010). `Assessing online usage'. *Research Information* (47):25.

J. Schöpfel & C. Boukacem-Zeghmouri (2010 forthcoming). `Assessing the Return on Investments in GL for Institutional Repositories'. In D. Farace & J. Schöpfel (eds.), *Grey Literature in Library and Information Studies*. De Gruyter Saur.

C. Boukacem-Zeghmouri, et al. (forthcoming). `Mesures d'usage. Usage effectif et utilisabilité des sources d'information'. In E. Delamotte & G. Lallich-Boidin (eds.), *Mesure de la science.* CNRS Editions.

Annexe B – Répertoires de référence

BASE
Bielefeld Academic Search Engine.
http://base.ub.uni-bielefeld.de/index.html

DSpace
Repositories using Dspace – Alphabetical.
http://www.dspace.org/index.php?option=com_content&task=view&id=596&Itemid=182

Eprints
Sites Powered by Eprints.
http://www.eprints.org/software/archives/

OpenDOAR
Directory for Open Access Repositories.
http://www.opendoar.org/

ROAR
Registry of Open Access Repositories.
http://roar.eprints.org/

Scientific Commons
Register URL
http://en.scientificcommons.org/register-repository

University of Illinois OAI-PMH Data Provider Registry.
http://gita.grainger.uiuc.edu/registry/

Webometrics
Ranking Web of World Repositories.
http://repositories.webometrics.info/

Ministère de l'Enseignement Supérieur et de la Recherche - Mission IST
Liste d'archives ouvertes et institutionnelles d'enseignement supérieur et de recherche françaises
http://www.sup.adc.education.fr/bib/

CCSD
Centre pour la Communication Scientifique Directe
http://www.ccsd.cnrs.fr/

ESUP Portail
Projet ESUP ENT
http://www.esup-portail.org/

Les Archives Ouvertes
COUPERIN Les Archives Ouvertes pour les Etablissements d'Enseignement Supérieur et de Recherche
http://www.couperin.org/archivesouvertes/

Wiki des Archives Ouvertes
Wiki des URFIST
http://urfist.enc.sorbonne.fr/ArchiveOuverte/OA.html

NUMES
Inventaire des fonds numérisés du ministère de l'enseignement supérieur et de la recherche
http://www.tge-adonis.fr/?Les-objectifs-du-projet-NUMES

ABHATOO
Portail de recherche documentaire du Centre National de Documentation du Haut Commissariat au Plan du Royaume du Maroc
http://www.abhatoo.net.ma/index.php/fre

THOTCURSUS
"Le monde de la formation à distance"
Répertoire des archives ouvertes de documents académiques (thèses, mémoires, rapports)
http://www.cursus.edu/?module=directory&subMod=PROD&action=getMod&pclass=6&uid=10693

OAIster
University of Michigan
http://www.oaister.org/

Celestial
Université de Southampton
http://celestial.eprints.org

SICD Grenoble
Service Interétablissements de Coopération Documentaire
http://bibliotheques.upmf-grenoble.fr/php/sicd2-ressources-categorie.php?categorie=2&txt=Archives+ouvertes

Annexe C – Sites analysés

Archives Audiovisuelles de la Recherche	AAR
Archive sur l'histoire et Mémoires de l'Académie royale des sciences	Académie des Sciences
Bibliothèque municipale de Toulouse : collection musicale	Agatange
Accès Libre aux Archives du Dépôt Institutionnel Numérique de la Maison des Sciences de l'Homme-Alpes	ALADIN
AnimalPhysiology-LivestockSystems	AnimalPhysiology-LivestockSystems
Institutional Archive of IFREMER	ArchiMer
ECOLE NORMALE SUPERIEURE HUMAN SCIENCES ARCHIVE	Archive ENS LSH
Archive ouverte INSEP	Archive ouverte INSEP
Archives EIAH	Archives EIAH
UNIVERSITE LUMIERE LYON 2 EPRINTS = Archives Lyon2 en SHS	Archives Lyon 2 en SHS
Archive Ouverte en Sciences de l'Information et de la Communication	ArchiveSIC
Université de Franche-Comté: ARTUR-FC - ARchive des Travaux Universitaires et de la Recherche	ARTUR
ArtXiker - ©HAL (artxiker.ccsd.cnrs.fr)	ArtXiker
Association des Technologies de l'Information pour l'Education et la Formation	ATIEF
Ressources linguistiques informatisées du laboratoire Analyse et Traitement Informatisé de la Langue Française	ATILF
Documents numérisés de la Bibliothèque de documentation internationale contemporaine	BDIC Archives et images
Différents fonds de la Bibliothèque Nationale de Strasbourg	BNUS
Bibliothèques Virtuelles Humanistes (BVH)	BVH
Vidéothèque numérique de l'enseignement supérieur	Canal U
Archives de l'Ecole des Mines de Nantes	Castore
Archive du Cemagref	Cemadoc
Correspondances Scientifiques	CHRST Correspondances
Consultation en texte intégral des Thèses en Réseau	CITHER
Le Conservatoire numérique des Arts & Métiers	CNUM-CNAM
Cristallography Open Database	COD
Colloques & Conférences de l'Université Lyon 2	Colloques & Conferences de l'Universite Lyon 2
Centre de Ressources pour la Description de l'Oral (CRDO) (crdo.vjf.cnrs.fr)	CRDO
Archive numérique d'Objets et de Matériaux iconographiques scientifiques	CRHST-AMOS
Histoire de l'électricité	CRHST-Ampère
Buffon et l'histoire naturelle	CRHST-Buffon
Correspondances Scientifiques	CRHST-Correspondances
Histoire de la justice	CRHST-Criminocorpus
Réseau européen pour l'histoire des cartes géologiques	CRHST-HistMap
Œuvres et rayonnement de Jean-Baptiste Lamarck	CRHST-Lamarck
Histoire et Sociologie des Sciences	CRHST-Nadirane
Sciences et savants en révolution (1780-1820)	CRHST-Révolution
UNIVERSITE LUMIERE LYON 2 CYBERTHESES	Cyberthèses

Dspace Avignon	Dspace Avignon
Archive EduTice	EduTice
ELEC - Éditions en ligne de l'École des chartes	ELEC
Bibliothèque de l'Ecole Nationale d'Administration	ENA
ENSSIB Bibliothèque Numérique	ENSSIB Bibliothèque Numérique
ENSTA Ecole Nationale Supérieure de Techniques Avancees	ENSTA
Eurecom Publications	Eurecom Publications
Fonds de la Bu de Dijon	Fonds Raymond Queneau (1903-1976)
Gallica	Gallica
Get Savoirs partagés	Get Savoirs partagés
Hyperarticle en Ligne	HAL
Archive pluridisciplinaire de l'Université d'Artois	HAL-Artois
Archive du Laboratoire "Biogéochimie et écologie des milieux continentaux"	HAL-BioEMCO
Archive du Cea	HAL-CEA
Cours en ligne C.E.L.	HAL-CEL
Archive du Cirad	HAL-CIRAD
Archive de Complex Systems Society	HAL-CSS
DECOR 04 - 1ère Conférence Francophone sur le Déploiement et la (Re) Configuration de Logiciels	HAL-DECOR 04
Archive de l''université Paris Descartes	HAL-Descartes
Dépôt Universitaire de Mémoires Après Soutenance (DUMAS)	HAL-DUMAS
Archive de l'Ecole Centrale de Lyon	HAL-EC-Lyon
Archive de l'École Nationale Supérieure des Mines de Saint-Etienne	HAL-EMSE
Archive de l'Ecole Normale Supérieure de Cachan	HAL-ENS Cachan
Archive ouvertes de l'Ecole de management de Grenoble	HAL-Grenoble-EM
The Nordic arts and humanities e-print archive	HAL-Hprints.org
ICPP 2004 - 12th International Congress on Plasma Physics	HAL-ICPP 2004
INSTITUT NATIONAL DE PHYSIQUE NUCLEAIRE ET DE PHYSIQUE DES PARTICULESHYPER ARTICLE EN LIGNE	HAL-IN2P3
Archive de l'Institut National de l'environnement industriel et des risques	HAL-INERIS
INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE ARCHIVE OUVERTE	HAL-INRIA
Archive de L'Institut National de la Santé et de la Recherche	HAL-INSERM
Archive de l'Institut National des Sciences de l'Univers	HAL-INSU
Archive de l'Institut de Recherche pour le Développement	HAL-IRD
Archive de l'Institut de Recherche Mathématique Avancée	HAL-IRMA
Archive de l'Institut de Radioprotection et de Sûreté Nucléaire	HAL-IRSN
IWFHR10 - International Workshop on Frontiers in Handwriting Recognition	HAL-IWFHR10
JITH2007 - 13èmes Journées Internationales de Thermique	HAL-JITH2007
Archive du Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier	HAL-LIRMM
Archive de l'Université Jean Moulin Lyon 3	HAL-Lyon 3
Archive de l'Université de Provence Aix-Marseille 1	HAL-Marseille 1
Archive de Météo-France	HAL-Météo France
Archive du Muséum National d'Histoire Naturelle	HAL-MNHN
Archive de l'Observatoire de Paris	HAL-ObsPM
Archive de l'Université Paris 1 Panthéon-Sorbonne	HAL-Paris 1
Archive de l'Institut Pasteur	HAL-Pasteur

PRASAC - Colloque Savanes africaines Prasac-Cirad	HAL-PRASAC
Archive institutionnelle de l'Ecole Normale Supérieure de Lyon	HAL-Prunel
Archive ouverte de l'Université de Savoie	Hal-Savoie
Archive ouverte en Sciences de l'environnement	HAL-SDE
Archive ouverte pour l'enseignement de la Société Française d'Optique	HAL-SFO
HAL-SHS : Social Sciences and Humanities	HAL-SHS
Archive du centre Alexandre Koyré	HAL-SHS-Koyre
Archive ouverte du Service de Santé des Armées	HAL-SSA
Archive de l'Ecole Supérieur d'Electricité	HAL-SUPELEC
Thèses de l'Inrp	HAL-TEL-INRP
Archive de l'Université Jean Monnet - Saint-Etienne	HAL-Ujm
Archive de l'Université de Nice Sophia Antipolis	HAL-Unice
Archive de l'Université de Limoges	HAL-UniLim
Archive de l'Université de Bretagne Occidentale	HAL-Univ-Brest
Archive de l'Université de Rennes 2 Haute Bretagne	HAL-Université Nantes
Archive de l'Université de Lyon 2	HAL-Univ-Lyon 2
Archive de l'Université de Nantes	HAL-Univ-Rennes 2
Horizon Pleins Textes	Horizon Pleins Textes
Sources numérisées en histoire des sciences et des techniques : l'entrepôt OAI-PMH du CRHST	HSTL
HSTL CRHST	HSTL CRHST
Travaux de l'Institut Français d'Etudes Andines	IFEA
Institut national de l'audiovisuel	INA
Différents fonds de l'Institut National d'Histoire de l'Art	INHA
Les thèses en ligne de l'Institut National Polytechnique de Toulouse	INPT ethesis
INSTITUT JEAN NICOD ARCHIVE ELECTRONIQUE	Institut Nicod Archive Electronique
I-Revues	I-Revues
IRIS	IRIS
Revue de l'Institut de Recherche en Informatique de Toulouse	IRIT
Fonds patrimonial de la Bu Pierre et Marie Curie	Jubilothèque
LACITO Archive	LACITO Archive
Libre Accès aux Rapports Scientifiques et Techniques	LARA
Publications du Laboratoire Biométrie et Biologie Evolutive	LBBE
Les Livres en Lignes de Presses Universitaires de Lyon	Les Livres en Lignes de Presses Universitaires de Lyon
Les mémoires en lignes de l'Institut d'Etudes Politiques de Lyon	Les mémoires en lignes de l'Institut d'Etudes Politiques de Lyon
Fonds de la Bu de Poitiers	Les premiers socialismes
Association "Librapport"	Librapport
Médiathèque de l'agglomération Troyenne Patrimoine	Médiathèque Troyes
Medic@Bibliothèque Numérique	MEDICa
Travaux universitaire de l'Université de Renne 2	Memorable
Mémoires de 3e cycle en Sciences de l'Information et de la Communication	MemSIC
Numérisation de Documents Anciens Mathématiques	NUMDAM
Publications l'IMFT : OAIMF	OAIMF
OATAO (Open Archive Toulouse Archive Ouverte)	OATAO
PARIS INSTITUTE OF TECHOLOGY PASTEL THESES	PASTEL
SICD Universités de Strasbourg - Patrimoine numérisé	Patrimoine numérisé ULP

Revues Scientifiques en Sciences Humaines et Sociale	PERSEE
ProdINRA - Base de données des publications de l'INRA	ProdINRA
UNIVERSITE DE PARIS X NANTERRE PUBLICATIONS	Publications de Paris X
Publications des laboratoires ENS	PubliENS
RevEl@Nice - Revues électroniques de l'université de Nice	REVEL
Revues.org - Fédération de revues en ligne en sciences humaines et sociales	Revues.org
Thèses de Sciences Po	Sciences Po - Thèses
Sciences Po Institutional Repository	SPIRE
Colloques en ligne - Cirad, Ifremer, Cemagref, Inra	Symposcience
Thèses en ligne	TEL
Open Archive in Technology Enhanced Learning	TeLearn
Thèses de Bordeaux 1	Thèses de Bordeaux 1
Thèses de l'IRISA	Thèses de l'IRISA
Thèses de l'ULP	Thèses de l'ULP
Thèses de l'UMB	Thèses de l'UMB
Thèses de l'université de Limoges	Thèses de l'université de Limoges
Archive institutionnelle des thèses de lUniversité de Toulouse I Sciences sociales	Thèses de Toulouse 1
Thèses de l'Université Nancy 2	Thèses en ligne de l'université Nancy 2
Thèses en ligne de l'INSA de Toulouse	Thèses en ligne INSA Toulouse
Thèses et Mémoires numérisés des universités de Clermont-Ferrand	Thèses et Mémoires numérisés des universités de Clermont-Ferrand
UNIVERSITE REIMS CHAMPAGNE ARDENNE Portail BU	Thèses URCA
Site des thèses de l'Université Toulouse III - Paul Sabatier - thesesups	THESESUPS
Thèses et mémoires de l'URS	UNERA-URS
Université Numérique Ingénierie et Technologie	UNIT
VizieR Catalogue Service	VizieR

Annexe D – Données d'analyse

Category	Number	Data	Description	Values
General information	1	Name	Nom de l'archive	Libre
General information	2	Acronym	Abréviation ou sigle	libre
General information	3	URL	Adresse électronique	libre
General information	4	Type institution	Indexation de l'établissement	3 catégories : Higher Education, Public Research Organization, Other
General information	5	Institution	Etablissement	libre
General information	6	Host	Etablissement hébergeur (si différent)	libre
General information	7	Description	Description de l'archive	libre
General information	8	Creation (yr)	Année de création de l'archive	Année ou n.d.
Specific information	9	Repository type	Indexation du type d'archive	4 valeurs : Doc-type, Institutional, Subject-based, Other
Specific information	10	Repository type deposit	Possibilité d'un dépôt par l'auteur	2 valeurs : yes, no
Specific information	11	Repository type heritage	Caractère patrimonial (bibliothèque numérique)	2 valeurs : yes, no
Specific information	12	Content	Description du contenu	libre
Specific information	13	Subjects	Indexation des domaines scientifiques	libre (multidisciplinary, SS&H, STM ...)
Specific information	14	Software	Logiciel de l'archive	libre
Specific information	15	Language	Langage(s) de l'interface	fre, eng, ger, spa
Specific information	16	Size (items) 2008	Nombre de dépôts en 2008, au moment de l'enquête	
Specific information	17	Size (items) 2009	Nombre de dépôts en 2009, au moment de l'enquête	
Specific information	18	Size 2005	Nombre de dépôts fin 2005	
Specific information	19	Size 2006	Nombre de dépôts fin 2006	
Specific information	20	Size 2007	Nombre de dépôts fin 2007	
Specific information	21	Size 2008	Nombre de dépôts fin 2008	
Content information	22	Presence GL 2008	Présence de la littérature grise dans archive	2 valeurs : yes, no
Content information	23	ETD 2008	Nb de thèses et dissertations en 2008	
Content information	24	Reports 2008	Nb de rapports en 2008	

Content information	25	Conferences 2008	Nb de conférences en 2008 (fusion des deux données C-Paper et Proceedings)	
Content information	26	Working papers 2008	Nb de travaux non publiés en 2008	
Content information	27	Courseware 2008	Nb de ressources pédagogiques en 2008	
Content information	28	Other GL 2008	Nb d'autres documents gris en 2008	
Content information	29	Datasets 2008	Nb de données scientifiques en 2008	
Content information	30	Multimedia 2008	Nb de dépôts multimédia en 2008	
Content information	31	Total nb GL 2008	Somme des documents gris en 2008 (sans "datasets" et "multimedia")	
Content information	32	% GL 2008	Part des documents gris par rapport au nombre total des dépôts en 2008, au moment de l'enquête	
Content information	33	Presence GL 2009	Présence de la littérature grise dans archive	
Content information	34	ETD 2009	Nb de thèses et dissertations en 2009	
Content information	35	Reports 2009	Nb de rapports en 2009	
Content information	36	Conferences 2009	Nb de conférences en 2009	
Content information	37	Working papers 2009	Nb de travaux non publiés en 2009	
Content information	38	Courseware 2009	Nb de ressources pédagogiques en 2009	
Content information	39	Other GL 2009	Nb d'autres documents gris en 2009	
Content information	40	Datasets 2009	Nb de données scientifiques en 2009	
Content information	41	Multimedia 2009	Nb de dépôts multimédia en 2009	
Content information	42	Total nb GL 2009		
Content information	43	% GL 2009		
Content information	44	Nb records 2009	Nb de notices bibliographiques dans texte intégral, au moment de l'enquête en 2009	
Content information	45	Articles 2009	Nb d'articles de revues au moment de l'enquête en 2009	

Content information	46	Articles publ 2007	Nb d'articles de revues avec année de publication 2007	
Content information	47	Articles publ 2008	Nb d'articles de revues avec année de publication 2008	
Qualitative data	48	Policies	Politique de dépôt explicite (cf. charte COUPERIN)	libre ou no
Qualitative data	49	Specific metadata GL	Présence de métadonnées spécifiques pour les documents gris	libre ou no
Qualitative data	50	Validation	Est-ce qu'il existe une fonction éditoriale pour le dépôt (vérification du contenu etc) ?	libre ou no
Qualitative data	51	Limited access fulltext	Accès au texte intégral de certain documents protégé, réservé à certains utilisateurs?	2 valeurs : yes, no
Qualitative data	52	Harvesting	Est-ce que l'archive moissonne les métadonnées d'autres archives ?	2 valeurs : yes, no
Qualitative data	53	Harvested	Est-ce que les métadonnées de l'archive sont-elles moissonnées par d'autres sites ?	2 valeurs : yes, no
Qualitative data	54	Other	D'autres éléments sur l'archive (projets etc.)	libre ; par exemple, si intégration dans HAL ou part d'OATAO etc.
Comments	55	Comments	Commentaires concernant la collecte des données	libre
Comments	56	Date 2008	Date de l'enquête en 2008	Date format personnalisée jj-mm-aaaa
Comments	57	Date 2009	Date de l'enquête en 2009	Date format personnalisée jj-mm-aaaa
Comments	58	Signature	Signature de la personne qui a effectué la collecte des données	Sur 3 caractères