Métadonnées dans le contexte d'une cyberinfrastructure de la recherche. "Le cas des thèses françaises"

Jacques Ducloy, Yann Nicolas, Diane Le Henaff, Muriel Foulonneau, Luc Grivel, Jean-Paul Ducasse

Métadonnées dans le contexte d'une cyberinfrastructure de la Recherche

"***Le cas des thèses françaises***"

ARTIST, Jacques Ducloy
INIST / CNRS
{artist, jacques.ducloy}@inist.fr

Yann Nicolas
ABES
nicolas@abes.fr

Diane Le Hénaff
INRA - Centre de Versailles
lehenaff@versailles.inra.fr

Muriel Foulonneau
CCSD
muriel.foulonneau@ccsd.cnrs.fr

Luc Grivel
Université Paris1
luc.grivel@univ-paris1.fr

Jean-Paul Ducasse
Université Lyon 2
ducasse@mail.univ-lyon2.fr

Cet article est la traduction (légèrement modifiée et actualisée) d'un article original soumis et accepté à la conférence DC 2006. Il a été publié avec le tire suivant :

Metadata towards an e-research cyberinfrastructure

The case of francophone PhD theses

Résumé Cet article analyse les pratiques et besoins liés aux métadonnées produites par les communautés de recherche francophones. Il est ciblé sur les thèses dont le cycle de vie est totalement contrôlé par les institutions académiques. Les applications liées au pilotage de la recherche ont également été privilégiés comme étant démonstratives du contexte de *e-science*. Trois études de cas illustrent le rôle fondamental de référentiels complémentaires sur les auteurs, affiliations ou éléments terminologiques. ARTIST, l'auteur collectif de cet article est également présenté.

Mots-clés : métadonnées, terminologie, pilotage de la recherche, thèses

1. Introduction

Cet article est le résultat d'un travail collaboratif réalisé dans le cadre de l'initiative ARTIST (Appropriation par la Recherche des Technologies de l'Information Scientifique et Technique[1]). Il a été écrit par un réseau d'auteurs, d'ingénieurs ou de documentalistes, et travaillant dans différentes institutions. Notre première expérience collective tournait autour d'interventions terminologiques dans le cadre de la traduction[2] d'un article écrit par Carl Lagoze et qui avait pour titre "What Is a Digital Library anyway, anymore ?", (qu'est ce qu'une bibliothèque numérique au juste ?) et qui traitait de la structure profonde de l'architecture des bibliothèques numériques [13]. Pour cette nouvelle expérience, la rédaction de cet article, nous avons choisi de traiter un sujet plus spécialisé : comment les métadonnées

[1] <http://artist.inist.fr/>
[2] <http://artist.inist.fr/article.php3 ?id_article=245 >

Comment citer ce document :
Ducloy, J., Nicolas, Y., Le Hénaff, D., Foulonneau, M., Grivel, L., Ducasse, J.-P. (2007). Métadonnées dans le contexte d'une cyberinfrastructure de la recherche. "Le cas des thèses françaises". Ametist (1), 75-96.

peuvent contribuer à aider la communauté de la recherche à bâtir des bibliothèques numériques.

Nous voudrions dépasser la problématique du rôle traditionnel des métadonnées « contribuer à améliorer la recherche d'information »[3] pour prendre en compte l'ensemble des besoins de la communauté scientifique, et par exemple leur attente de visibilité. A ce sujet, la publication annuelle du classement des meilleures universités au niveau mondial [10] est devenu un indicateur essentiel pour les responsables de la politique scientifique des institutions. Optimiser la qualité des métadonnées, notamment sur les affiliations, est maintenant un élément stratégique répondant à ce besoin de visibilité. Les chercheurs eux-mêmes suivent de très prés cette évolution et manifestent un besoin vital de publier dans des revues à fort facteur d'impact. Le slogan « *publish or perish* » est utilisé par les établissements eux-mêmes comme une incitation à déposer dans des archives ouvertes pour améliorer l'impact des productions scientifiques[8].

Les bibliothécaires et les communautés de chercheurs commencent donc à réaliser que les métadonnées ne sont pas seulement utiles pour rechercher des documents mais peuvent être appelées à jouer une fonction plus stratégique. Cette nouvelle façon de voir est peut être un premier pas vers une analyse plus globale sur le rôle potentiel de la publication scientifique dans ce qui est appelé outre-atlantique par « *cyberinfrastructure for e-science or e-research* » [14][19], autrement dit « comment faire de la science au temps des bibliothèques numériques. En France, le concept d'e-recherche est parfois réduit aux grilles d'ordinateurs en réseau, secteur déjà ancien [6] mais où les références sont particulièrement visibles[4]. En revanche dans d'autres pays européens, comme le Royaume-Uni, l'accent est mis de plus en plus sur tous les aspects liés aux données produites par la recherche avec par exemple la création du DCC[3].

Dans ce contexte, cet article veut approfondir comment les métadonnées peuvent être utilisées pour le pilotage de la recherche en France. Nous avons choisi de spécialiser la discussion sur les thèses dont le cycle de vie est totalement contrôlé dans un cadre académique, sachant qu'une grande partie des discussions peut s'appliquer à l'ensemble des productions de la communication scientifique.

Pour produire des systèmes d'information efficaces, les métadonnées doivent gagner en complexité. Or, dans un contexte d'archive ouverte, les solutions les plus populaires, telles que DSpace [15] ou Eprints[4], ne sont pas du tout contraignantes pour inciter un déposant à élaborer des métadonnées consistantes. La plupart des recommandations se limitent donc souvent à un sous ensemble basique du « Dublin Core » avec le moissonnage comme seule contrainte. Les thèses sont naturellement concernées par cette quête croissante de visibilité[9]. Nous constaterons que le démarrage de leur cycle de vie demande que les métadonnées dépassent le seul cadre descriptif pour inclure des éléments de gestion. En effet, et plus particulièrement dans un contexte français, plusieurs établissements ou institutions sont concernés et doivent coopérer.

Comme pour toutes les productions numériques de la recherche, les métadonnées des thèses doivent pouvoir être utilisées dans tout portail (national, international, thématique) susceptible d'accroître leur visibilité. Dans un contexte de pilotage de la recherche, elles doivent

[3] Data Curation Center
<http://www.dcc.ac.uk/>
[4] <http://www.eprints.org/>

également s'avérer aptes à la manipulation par des outils infométriques en vue de produire des analyses de veille scientifique ou stratégique. A ce niveau, nous verrons l'importance fondamentale des vocabulaires et affiliations.

Dans la première partie de cet article nous commencerons par une présentation du contexte francophone. Puis nous décrirons sommairement plusieurs applications structurantes autour de la production des thèses, des catalogues collectifs et des archives institutionnelles. Enfin nous présenterons trois études de cas illustrant trois aspects des métadonnées et des vocabulaires.

2. Les bibliothèques numériques pour la recherche en numérique : un panorama des initiatives européennes et francophones

En matière de coopérations internationales dans le monde de la recherche les acteurs francophones doivent se positionner à travers une multitude d'infrastructures autant nationales, qu'internationales.

Comme tous les autres, ils s'appuient déjà sur les initiatives internationales de normalisation et doivent tenir compte de l'évolution des standards ou des pratiques des Etats-Unis (et dans le monde entier). Ils ont également leurs propres réseaux de coopérations. La France et la Belgique par exemple sont également insérées dans l'Europe d'une part et dans la Francophonie de l'autre. De leur côté, l'Algérie, le Maroc et la Tunisie appartiennent à la communauté de langue arabe.

En conséquence, les acteurs de la recherche doivent coordonner leurs efforts avec de multiples initiatives dans de multiples zones de coopération. Les stratégies de métadonnées définies pour la communication scientifique doivent assurer l'interopérabilité de la production francophone dans tous ces réseaux.

Les acteurs francophones doivent donc faire face à une diversité d'initiatives d'animation, dans de multiples réseaux régionaux, nationaux ou internationaux, et, en terme de métadonnées, à une multitude de contraintes dues à des situations administratives différentes.

2.1 - Le contexte international autour de l'e-Recherche

Le mouvement pour les archives ouvertes et les initiatives sur les archives ouvertes ont encouragé les institutions de recherche à rendre leurs thèses accessibles sur la toile. D'un point de vue technique, le protocole de moissonnage OAI-PMH[5] ouvre la possibilité de partager et d'échanger des métadonnées sur les documents académiques. Tout ceci a permis la création d'une infrastructure ouverte pour publier les thèses et dissertations qui sont déposées dans des référentiels ouverts et partagés sur de vastes réseaux. En France, ceci s'est par exemple traduit par la création du CCSD[6], une initiative majeure pour repenser le processus de la communication scientifique et qui sera présenté plus loin (section 3.2).

Aux Etats-Unis, dans le cadre des programmes « Digital Library Initiative » DLI-I et DLI-II financés par la National Science Foundation, les moyens dégagés pour mettre en place une

[5] Open Archives Initiative Protocol for Metadata Harvesting
< http://openarchives.org>
[6] Centre pour la Communication Scientifique Directe
< http://www.ccsd.cnrs.fr/accueil.php3?lang=efr>

base ouverte sur les bibliothèques numériques ont conduit à la création de projets majeurs tels que le projet « National Science Digital Library »[7]. NSDL a été un vecteur de promotion des normes relatives aux architectures ouvertes pour les bibliothèques électroniques et a contribué au développement de nouveaux services.

Le réseau « Networked Digital Library of Theses and Dissertations » (NDLTD)[8] [18] a mis en place une solution complète, incluant les processus et le contrôle de flux de l'édition électronique des thèses et dissertations. Ce projet a traité les problèmes relatifs aux droits et propriétés intellectuelles des thèses et rapports de recherche. Il a amélioré l'interopérabilité technique entre les référentiels en encourageant l'usage du protocole OAI-PMH et des interfaces SRU. Enfin, il a favorisé l'interopérabilité des contenus en adoptant le jeu de métadonnées de thèses (ETDMS) [5] construit comme un profil d'application Dublin Core. ETDMS est notamment utilisé dans le projet Cyberthèses (portail francophone) qui sera décrit plus loin en section 3.1.

D'autres formats de métadonnées sont également utilisés, tels que ceux qui relèvent de la souche MARC ou MODS (Metadata Object Description Schema, maintenu par la "Library of Congress")[9]. Par exemple le "Metadata Working Group of the Texas Digital Library" vient de développer un profil d'application de métadonnées descriptives pour les thèses et dissertations en MODS[10]. Enfin, de nombreuses bibliothèques insèrent leurs métadonnées descriptives dans un ensemble METS (comme par exemple le Florida Center for Library Automation[11], ou l'université d'Uppsala[12]).

Le projet ARTIST, auteur collectif de cet article, a notamment pour objectif d'assurer une veille sur l'évolution des initiatives autour des métadonnées afin d'en faire bénéficier les acteurs français ou francophones. Il cherche également à fédérer leurs efforts de normalisation.

2.2 - Le contexte européen

Le programme européen IST, comme les DLI aux Etats-Unis, s'est focalisé sur les recherches en technologies de l'information afin de construire un cadre ouvert pour les bibliothèques numériques. La Commission Européenne a ainsi financé plusieurs projets, par exemple le Open Archives Forum[13] [15] en vue d'accroître la compétence des acteurs nationaux et pour approfondir les réflexions technologiques relatives à la communication scientifique.

La Commission soutient et finance également la normalisation de l'Espace Européen de la Recherche. Par exemple le projet EuroCRIS[14] veut « transformer l'information relative à la recherche en savoir » en soutenant et en publiant la recommandation CERIF[15] (Common European Research Information Format).

[7] <http://www.nsdl.org >
[8] < http://www.ndltd.org/index.en.html >
[9] < http://www.loc.gov/standards/mods/ >
[10] < http://www.tdl.org/projects/metadata/tdlappprofile.pdf >
[11] < http://www.fcla.edu/dlini/etd.html >
[12] < http://publications.uu.se/theses/index.xsql?lang=en>
[13] < http://www.oaforum.org/>
[14] < http://www.eurocris.org/en/ >
[15] < http://www.cordis.lu/cerif/home.html >

Néanmoins, les principales initiatives pour construire un cadre pour les bibliothèques numériques de la recherche ont été lancées au niveau national. Le JISC (Joint Information Systems Committee) a financé un projet au nom évocateur, Thesis alive[16] (la thèse vivante) ou Daedalus[17] pour favoriser l'édition électronique des thèses et des dissertations au Royaume-Uni. Il a également favorisé l'insertion des établissements britanniques dans le réseau de NDLTD. Au Pays-Bas, le programme SURF (groupement d'établissements d'enseignement ou de recherche pour les technologies des services réseau de l'information ou de la communication) a soutenu le projet DARE (Digital Academic Repositories)[18] pour modifier l'infrastructure de dissémination de l'information académique. Cependant, tous les pays n'ont pas su, ou pu, mettre en place de telles initiatives pour créer un tel cadre global à la publication scientifique.

Actuellement la priorité du programme européen IST sur la mise en réseau de la recherche (IST 2.5.6) s'attaque aux enjeux de la construction d'un cadre européen pour la publication de résultats académiques. Dans ce contexte, le projet DRIVER (2006-2008), coordonné par l'université d'Athènes va contribuer à bâtir cette infrastructure pour la recherche. Il sera basé sur les concepts d'infrastructure ouverte proposés dans le cadre du réseau d'excellence DELOS (network of Excellence for digital libraries)[19].

En pratique les acteurs européens doivent faire face à une extrême variété de situations administratives issues du passé. L'interopérabilité entre les systèmes nationaux doit donc tenir compte de l'hétérogénéité des structures académiques ou de recherche, leurs dépendances et leurs relations (voir paragraphe 4.2). De plus la construction d'une infrastructure européenne doit répondre aux enjeux du multilinguisme qui entraîne des conséquences significatives sur la création de schémas de métadonnées et la gestion des terminologies.

2.3 - Le contexte francophone

Les réseaux de communautés de recherche francophones doivent également maîtriser des contraintes organisationnelles et linguistiques. La plupart des pays francophones (plus de 50 pays sur 5 continents) sont extérieurs à l'Europe et ont des infrastructures de soutien de la recherche très diversifiées.

Plusieurs initiatives contribuent à structurer cet ensemble et notamment par le biais de la standardisation. Par exemple, et plus précisément sur les thèses et dissertations, l'OIF (Organisation Internationale de la Francophonie)[20] a financé le montage du projet Cyberthèses (voir section 3.1). D'autres institutions comme l'Agence Universitaire de la Francophonie (AUF)[21] ont soutenu d'autres programmes le "Système d'Information Scientifique et Technique" (SIST)[22] en Afrique francophone .

De nombreux pays francophones sont en réalité des espaces ou cohabitent des langues différentes. Ils doivent alors mettre en service des systèmes multilingues avec toutes les contraintes bien connues pour les langues latines et qui deviennent plus complexes dans le cas

[16] . < http://www.thesesalive.ac.uk/ >
[17] < http://www.lib.gla.ac.uk/daedalus/ >
[18] < http://www.darenet.nl/ >
[19] < http://delos-noe.iei.pi.cnr.it/ >
[20] < http://www.francophonie.org/ >
[21] < http://www.auf.org>
[22] < http://www.sist-sciencesdev.net/ >

de l'arabe. L'IMIST (Institut Marocain pour l'Information Scientifique et Technique)[23] est justement en train de mettre en en service un catalogue bilingue pour les thèses[24].

2.4 - Le contexte français

La situation française est encore plus complexe en raison de la multiplicité des cadres administratifs (un exemple sera donné en section 4.2). Dans les dix dernières années, aucun programme réellement ambitieux pour structurer la communication scientifique n'a été lancé. Les opérateurs publics en charge des bibliothèques universitaires, l'IST ou la Communication Scientifique tels que l'ABES (Agence Bibliographique de l'Enseignement Supérieur)[25] ou l'INIST (INstitut de l'Information Scientifique et Technique)[26] ont surtout mis en place des projets intégrant, dans une approche centralisée, toute la chaîne documentaire du dépôt à la fourniture d'indicateurs. Les porteurs d'initiatives locales se sentent alors complètement déconnectés des opérations lancées au niveau national.

Autrement dit, et de façon paradoxale, les opérateurs nationaux ont eu tendance à mettre en place des opérations « trop bien cadrées » en fonction de leurs propres contraintes et qui ne peuvent pas servir de cadre pour des opérations de terrain voulant s'intégrer dans une bibliothèque numérique « fédérative » de la recherche.

3. Quelques inititiatives structurantes

3.1 Cyberthèses

Cyberthèses est issu d'un programme francophone qui s'est élargi à l'Amérique du sud. Cyberdocs, la plateforme opérationnelle associée est portable, libre et propose une chaîne de production allant de la rédaction du document à sa diffusion en passant par le stockage.

Les principaux acteurs du réseau Cyberthèses dans la francophonie sont l'Université de Dakar au Sénégal, l'Université d'Antananarivo à Madagascar et l'Institut National d'Agronomie à Alger [1] (signalons également l'Université du Chili à Santiago[27].

Dans le projet Cyberthèses, chaque université est en charge de la conversion de ses thèses dans un format d'archivage en XML basé sur le schéma TEI-lite[28]. A l'Université Lyon 2, l'enregistrement et le dépôt électroniques font maintenant explicitement partie de la "charte des thèses" qui définit la relation entre le doctorant et son établissement. Le dépôt d'une version complète de la thèse est obligatoire. L'enregistrement est encore assuré par l'administration mais les modules logiciels de gestion des flux on été conçus pour soutenir

[23] < http://www.imist.ma/ >
[24] Outre les problèmes liés à la dualité des alphabets, rappelons que les sens d'écriture eux-mêmes sont différents. De plus, dans plusieurs éléments de métadonnées comme par exemple dc:descrition, une phrase en arabe peut comporter une expression (ou une formule) en Français.
[25] < http://www.abes.fr/ >
[26] < http://www.inist.fr/ >
[27] < http://www.cybertesis.cl/,universities/ >
[28] < http://www.tei-c.org/Lite/teiu5_fr.html >

simultanément le dépôt et l'enregistrement du document et de ses métadonnées (Dublin Core[29] , ETDMS, OAI-PMH).

3.2 CCSD : Une archive ouverte avec vue institutionnelle

L'acronyme CCSD signifie « Centre de la Communication Scientifique Directe » et précise sa vocation à promouvoir la communication scientifique directe entre les chercheurs. Très proche de la philosophie d'ArXiv la plate-forme logicielle HAL[30] fournit une interface pour que les auteurs puissent charger dans la base du CCSD leurs manuscrits d'articles scientifiques dans n'importe quel domaine. La plupart des établissements de recherche français ont signé un accord pour un pilotage global du CCSD. HAL peut offrir une vue personnalisée pour chaque organisme participant.

Une interface spécifique appelée TEL (thèses-EN-ligne) est dédiée à l'auto-archivage des thèses qui sont également considérées comme des documents important pour la communication directe entre scientifiques. TEL peut être moissonné via le protocole OAI-PMH avec deux formats de métadonnées : du Dublin Core non qualifié[31] et un schéma spécifique au CCSD.

Une caractéristique de celui-ci est la définition formelle et précise des relations entre auteurs et affiliations qui doivent être correctement déclarées dans la procédure de dépôt. Ce dispositif facilite les vues institutionnelles et illustre l'objectif premier du CCSD : une archive ouverte conçue pour faciliter le pilotage de la recherche par les institutions concernées.

3.3 STAR : un médiateur entre des acteurs locaux mondialisés.

Le Ministère de l'Education Nationale, en charge des thèses, propose, par l'intermédiaire de son agence bibliographique ABES, un nouveau service appelé STAR qui fonctionne comme un échangeur au sein des circuits de communication des thèses.

En entrée, STAR recueille les thèses et leurs métadonnées auprès des établissements de soutenance, seuls habilités à garantir que le document transmis est bien conforme à la version validée par le jury.

En sortie, la thèse sera systématiquement orientée vers un système national de conservation numérique, géré par le CINES (Centre Informatique National de l'Enseignement Supérieur)[32]. De même, les métadonnées seront traitées pour alimenter la bibliographie nationale des thèses, au sein du catalogue collectif Sudoc (Système Universitaire de DOCumentation)[33] qui héberge le référentiel national des thèses avec un catalogage en UNIMARC.

A côté de ces débouchés réglementaires, STAR proposera aux établissements de soutenance des services complémentaires (voir la figure 1) :

[29] Un atelier du réseau TEI-FR, sur les adaptations de la Text Encoding Inititiative au monde francophone, est dédié à la conformité de l'en-tête de la TEI aux recommandations du DCMI. < http://listserv.inist.fr/wwsympa.fcgi/info/tei-fr >
[30] < http://hal.ccsd.cnrs.fr/ ?langue=en >
[31] < http://www.openarchives.org/OAI/2.0/oai_dc.xsd >
[32] < http://www.cines.fr/ >
[33] < http://www.sudoc.abes.fr >

- diffusion par HAL ou d'autres bases documentaires ;
- signalement et indexation plein texte dans le portail documentaire Sudoc ;
- assignation d'un identifiant pérenne (URI) et service de résolution pour donner accès à la thèse, quelle que soit sa localisation en ligne.

En un seul dépôt, chaque établissement sera en mesure d'assurer la conservation à long terme de ses thèses, leur diffusion par de multiples canaux, en ayant la certitude qu'il s'agit de la bonne version, validée scientifiquement et administrativement. Le seul fait qu'une thèse française soit accédée via l'identifiant pérenne national témoignera qu'il s'agit d'une version officielle, garantie par l'établissement. L'URI servira d'identifiant unique, d'URL d'accès en ligne et de tampon de validité.

STAR ne se substitue par aux outils ou processus que les établissements ont pu constituer localement. Certes, pour ceux qui ne disposent pas d'outil local pour gérer les thèses, STAR offrira une interface Web pour déposer les fichiers et saisir les métadonnées. Pour les autres, STAR pourra importer les fichiers et les métadonnées produits localement. Ces métadonnées seront encodées selon le format d'échange national TEF.

TEF (Thèses Electroniques Françaises) est une recommandation produite par un groupe de travail AFNOR[34] (AFNOR CG46/CN357/GE5). Elle vise à organiser de manière cohérente et souple différentes catégories de métadonnées de thèse : métadonnées bibliographiques (DC), métadonnées de droits (METS Rights), métadonnées administratives relatives au diplôme délivré et métadonnées de conservation. Dans TEF, FRBR[35] sert de modèle conceptuel pour débrouiller la notion de thèse, METS d'enveloppe XML pour lier ces différents modules de métadonnées et enfin Schematron[36] d'outil de validation à la fois précis et souple pour préciser les contraintes métiers propres au contexte français.

L'outil STAR et la structure de données TEF sont des intermédiaires au service de ceux qui produisent et valident les thèses et leurs métadonnées et de ceux qui les exploitent.

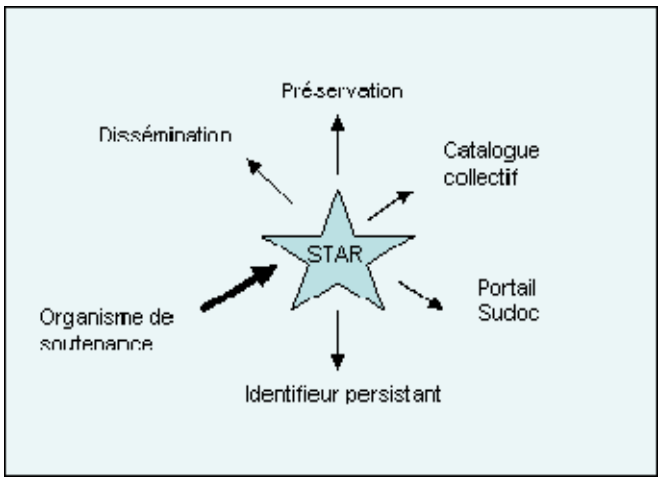

Figure 1.

[34] L'AFNOR (Association Française de NORmalisation) est un chapitre français de l'ISO et du CEN.
[35] < http://www.ifla.org/VII/s13/frbr/frbr.htm >
[36] Schematron est un langage de schéma XML décrivant un ensemble de règles à respecter. <http://xmlfr.org/index/object.title/schematron/>

3.4 INIST

L'INIST (INstitut de l'Information Scientifique et Technique) est une centrale documentaire qui produit des bases bibliographiques (Pascal and Francis).

Cette activité est en évolution permanente. Il y une quinzaine d'années les notices bibliographiques étaient fabriquées manuellement pour être codées dans un format ISO 2709. Dans une première étape, le procédé de production a été modernisé en utilisant un format équivalent dans une codification SGML. Maintenant l'INIST a introduit des mécanismes d'indexation automatique et vise à homogénéiser ses métadonnées d'échanges autour d'un schéma XML nommée Exodic[37], compatible avec le Dublin Core.

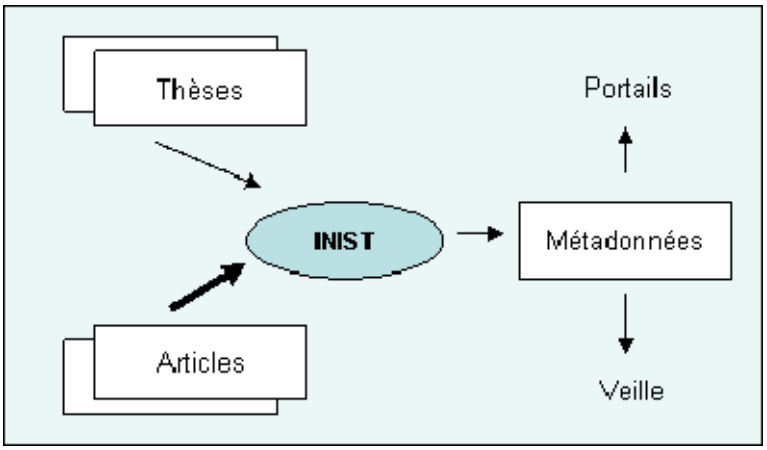

Figure 2.

Un des départements de l'INIST est spécialisé dans la réalisation de portails bibliographiques ou d'études de veille (voir figure 2). Il est de plus en plus impliqué dans la définition d'indicateurs à caractère institutionnel. Ceci a amené l'INIST, comme le CCSD, a améliorer la qualité des métadonnées relatives aux relations entre auteurs et affiliations.

4. Trois études de cas

Nous venons de présenter un ensemble d'opérateurs qui semblent offrir un ensemble complet de services. Malheureusement, pour des raisons historiques, ils ont été mis en place indépendamment les uns des autres. En pratique les choses sont donc souvent moins idylliques qu'il n'y paraît. Cet article est rédigé par une équipe de personnes venant précisément de ces institutions, qui ont formalisé le caractère crucial de l'interopérabilité et qui travaillent sur l'échange de formats basés sur le Dublin Core qualifié.

Est-ce suffisant ?

Pour le vérifier, nous allons maintenant aborder trois études de cas dont les besoins vont au delà du simple niveau bibliographique (dépôt et recherche associée) pour aborder la problématique du pilotage. Dans le premier cas nous allons analyser le déroulement de la vie d'une thèse, depuis sont dépôt jusqu'à son accessibilité en OAI-PMH. Le "flux théorique" qui se dessine dans la figure 3 semble assez simple : une thèse est gérée au départ par

[37] < http://international.inist.fr/article159.html >

Cyberthèses, puis est intégrée à la plate-forme STAR et finit enfin par rejoindre le flux des articles au CCSD. Nous nous situerons dans un cas où deux institutions, un EPST (Etablissement Public à caractère Scientifique et Technique) et une université sont impliqués.

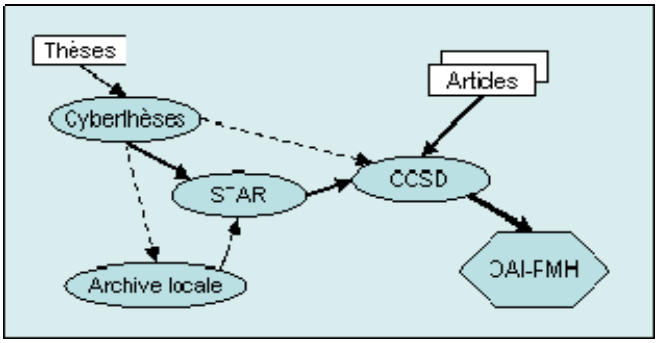

Figure 3.

Ensuite nous aborderons deux cas d'utilisation des métadonnées pour des analyses de veille ou pour la création de portail dans des cadres institutionnels ou thématiques. Nous supposerons que les métadonnées auront été au préalable indexées et homogénéisées dans un centre de documentation tel que l'INIST.

4.1 La réutilisabilité des métadonnées

Comme indiqué précédemment, les institutions de recherche françaises encouragent les chercheurs à déposer leurs travaux dans une archive ouverte à des fins de visibilité et de valorisation de leur production scientifique. Celles ci comprennent différents types de documents ou données : des articles publiés mais également des rapports d'expertise, des cours, des communications de congrès, des thèses, de la documentation de logiciels ou même la production de données primaires (démographiques ou autres). L'ensemble de ces productions sert à l'évaluation du chercheur et de son unité de recherche.

Actuellement, au sein des organismes publics de recherche, les chercheurs doivent fournir l'information sur leur production scientifique au service de l'évaluation tous les quatre ans. Pourquoi obliger le chercheur ou l'unité à produire une information qui existe déjà dans l'entrepôt documentaire ?

L'INRA, comme tout autre organisme public de recherche, pourrait souhaiter transférer les données de l'entrepôt documentaire vers l'application qui gère les dossiers d'évaluation des chercheurs et des unités de recherche. Explicitons le cas des thèses en rappelant que le doctorant effectue son travail de recherche dans une unité de recherche dépendant d'une (ou plusieurs) institution(s) de recherche ; le doctorant est inscrit dans une université de rattachement.

Maintenant, étudions les données exportables les plus fondamentales qui sont nécessaires à l'application « évaluation » :

- **Les personnes**

Le chercheur qui aura effectué sa thèse dans une institution de recherche devra pouvoir être identifié par le service d'évaluation comme chercheur, anciennement doctorant et rédacteur

d’une thèse présente dans l’entrepôt. Le chercheur pourra également avoir changé de nom. Identifier les différents statuts ou noms d’une personne, détecter les homonymies pour éviter les erreurs, nécessitent d’enrichir les métadonnées ou croiser les informations relatives aux personnes.

- **Les structures**

L’unité de recherche dans laquelle le doctorant a effectué son travail de recherche peut être différente de l’unité dont dépend le chercheur qu’il est devenu. Les deux structures sont légitimes à revendiquer les travaux décrits dans la thèse, l’une pour avoir consacré des moyens pour le travail de recherche et l’autre en tant qu’affiliation du chercheur. L’application « évaluation » a besoin de pouvoir identifier la structure telle qu’elle a été mentionnée dans la thèse avec son équivalent dans le référentiel des structures de l’institution. L'évaluation concernant les 4 dernières années d'activité, le référentiel "structure" doit pouvoir « historiser » les changements de structure et suivre les affectations des personnes dans ces structures.

- **Les partenaires**

La diversité des établissements de recherche en France incite à créer une liste de partenaires scientifiques et à qualifier les différentes collaborations. Ainsi, l’université de rattachement du doctorant est mentionnée dans la thèse. Il s’agira donc de compléter la liste des partenaires ou/et d’y ajouter le type de collaboration.

4.2 : Etude à caractère institutionnel

L’expérience des demandes traitées par l’INIST montre que la détection des relations entre le communautés de chercheurs est un outil très important pour le pilotage de la recherche[7]. Concernant les thèses, le traitement informatique des affiliations des membres d’un Jury peut fournir des indicateurs pertinents pour découvrir les "co-laboratoires cachés".

D’un point de vue technique, il s’agit d’extraire des métadonnées les éléments relatifs aux auteurs et affiliations. Ceci serait assez élémentaire dans un monde normalisé. Mais en réalité une même institution peut apparaître sous des formes très diversifiées. Dans ce contexte, les fichiers d’autorités complétés par des éléments terminologiques jouent un rôle important dans la normalisation des données bibliographiques avant les traitements infométriques proprement dits.

Dans un premier temps, les fichiers d’autorité permettent l’établissement de listes de correspondance, par exemple, pour les noms de pays. La technique généralement utilisée pour établir des équivalences et uniformiser les champs de données présentant des variations essentiellement typographiques (majuscule, minuscule, etc.) ou flexionnelles (pluriels, singuliers) est d’aboutir à une convergence par rapport à une forme appauvrie, analogue à une clé à laquelle est associée sa forme attestée.

Dans la plupart des indicateurs infométriques, l’unité d’analyse (l’objet d’étude) est une entité géographique ou institutionnelle. Les publications sont assignées à ces unités sur la base d’une analyse des adresses des auteurs. Au sein de données bibliographiques, les variations de noms de pays sont limitées en nombre. Mettre en correspondance publications et institutions de recherche est une tâche beaucoup plus délicate qui ne peut être effectuée directement et

simplement en se basant sur les adresses des auteurs des publications. Très fréquemment, il arrive de rencontrer de nombreuses formes lexicographiques pour la même donnée.

Ceci suppose l'existence de fichiers d'autorité géographiques (codes postaux, villes, régions, pays) et institutionnels (code d'institution, classification sectorielle des organismes, ...).

Tant que l'on reste sur un niveau statistique approximatif ce type de post-traitement est suffisant. Mais dès que l'on cherche à améliorer la précision des études, et donc des calculs associés, des situations très complexes apparaissent immédiatement.

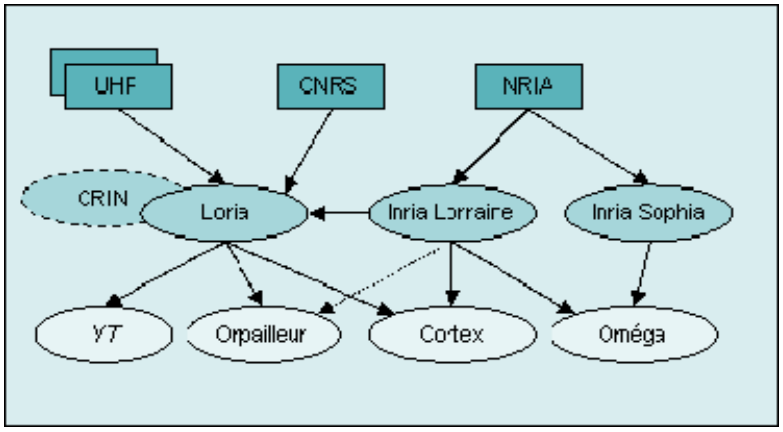

Figure 4.

La figure 4 montre un exemple concret de la situation d'un laboratoire de Nancy, il y a deux ans. Il veut illustrer la complexité de ce que l'on peut rencontrer dans la réalité. Une simple liste d'affiliations n'est pas suffisante et des outils plus sophistiqués sont nécessaires pour la représenter.

Au plus haut niveau, nous trouvons deux EPST (Le CNRS et l'INRIA) et une université (UHP – Université Henri Poincaré)[38].

Au niveau intermédiaire (laboratoire), le LORIA est une UMR (Unité Mixte de Recherche) entre l'INRIA, l'UHP, le CNRS et deux autres universités. L'INRIA Lorraine est le nom de l'Unité (avec sa dimension administrative) INRIA de Lorraine. Quelques années plus tôt le CRIN était le nom d'un laboratoire mixte entre le CNRS et les universités, mais sans l'INRIA. Le CRIN n'existe plus à ce jour, mais de nombreux articles ou thèses sont indexés par CRIN et doivent être pris en compte dans un traitement historique.

Au niveau le plus bas, celui des équipes, la plupart des entités, comme Cortex, sont reconnues par toutes les organisations. Mais les choses peuvent être plus complexes !

Par exemple, une jeune équipe (appelée YT pour Young Team) est seulement reconnue par les universités et pas encore par l'INRIA (et donc par l'INRIA Lorraine). Orpailleur était en cours d'habilitation au moment de l'étude. Oméga est une équipe commune entre « l'Inria Loraine » et « l'INRIA Sophia » mais qui ne fait pas partie du LORIA...

[38] La Figure 4 est en réalité une vue simplifiée de la situation réelle, et, par exemple, deux autres universités (Nancy II and INPL) sont également associées.

Seul un solide format de métadonnées d'affiliation tel que LEAF[12], peut offrir une base solide qui devra être complétée par une solide étude sur les liens d'affiliation. Nous aurions voulu lancer un groupe de travail au sein d'ARTIST pour étudier ce sujet en implémentant plusieurs types de liens dans ce qui s'apparente à une taxonomie des affiliations.

4.3 : Etude à caractère thématique : la biodiversité

Pour cette dernière étude de cas, nous allons aborder un aspect plus thématique de la veille. La biodiversité est un sujet transversal qui mobilise fortement les instances de décisions. La Commission Européenne a par exemple lancé un projet nommé BiodivERsA[39] qui vise à mettre en place une coopération transnationale et efficace dans le cadre des financements de programmes liés à la biodiversité.

Nous allons aborder, à titre d'exemple, un sujet qui n'est pas inscrit dans ce projet mais qui s'inscrit dans son cadre de réflexions : « comment se situent les travaux des laboratoires publics par rapport aux axes des programmes de soutien à la biodiversité ? ». L'information disponible est contenue dans les publications et, par exemple, dans les métadonnées associées aux thèses.

Le programme BiodivErSA évalue ainsi les chiffres liés aux financements des actions de R&D sur la BioDiversité :

- plus d'une centaine d'agence de financement[2] ;
- plusieurs programmes par agences, soit des centaines de programmes ;
- plusieurs projets par programme, soit des milliers de projets ;
- plusieurs résultats plus ou moins directs par projet, soit des dizaines de milliers de résultats !

En pratique BiodiverSA veut créer un inventaire de tous les programmes de financement de recherche en biodiversité (sous forme d'une "métadatabase" compatible avec les recommandations CERIF[40]. Les aspects terminologiques jouent ici un rôle fondamental. Plus précisément, des outils de classification (ou taxonomies) vont servir, avec des contraintes de calcul numérique, à produire un ensemble d'indicateurs.

Au moment où cet article est rédigé le système de classification envisagé comporte trois volets :

- une classification scientifique basée sur ASRC (Australian Standard Research Classification). Elle est assez proche des standards proposés par l'OCDE (Organisation pour la Coopération et le Développement Economique) pour l'analyse des développements en Recherche et Développement ;
- une classification spécifique à la biodiversité, basée sur plusieurs systèmes existants ;
- une indexation complémentaire à base de mots-clés issus de la CBD[41] (Convention on Biological Diversity).

Maintenant, comment aborder le problème posé par cette étude, et, par exemple, comment construire un indicateur basé sur la production de thèses ?

[39] < http://www.eurobiodiversa.org/ >
[40] < http://www.cordis.lu/cerif/ >
[41] <http://www.biodiv.org/doc/lists/cbd-voc.pdf>

Un premier aspect est l'alimentation d'une application compatible avec la norme CERIF (qui devrait être utilisée par BiodivERsA avec quelque chose de proche du Dublin Core qualifié. Mais le point probablement le plus crucial sera relatif au système de classification. On peut imaginer que des laboratoires fortement concernés utilisent le système de classification de BiodivERsA pour être plus visibles des agences de financement. Dans ce cas il convient simplement d'être vigilant dans l'indexation proprement dite. En effet, celle-ci va être utilisée dans une logique de comptabilisation (ce qui est différent d'un simple classement pour rangement ou pour faciliter la navigation).

Mais l'immense majorité des thèses relatives à la Biodiversité ne vont pas utiliser ce système. Il faudra donc recourir à des adaptations terminologiques. Ceci peut encore être relatvement simple si les thèses sont indexées avec un plan de classement reconnu (MeSH par exemple).

Dans les autres cas, une analyse linguistique du contenu du document devient nécessaire demandant une fois encore de disposer de plusieurs types de ressources terminologiques. Les référentiels terminologiques permettent alors de prendre en compte les termes complexes comme les groupes nominaux, et de nombreux autres phénomènes linguistiques comme, par exemple, la variation des mots dans leurs formes graphiques, le rôle qu'ils jouent dans la phrase, leur sens dans un contexte donné. En utilisant cet arsenal de techniques on peut ainsi utiliser des thesaurus ou vocabulaires pour réaliser des analyses à caractère numérique.

Dans les deux premières études de cas, nous avions montré que, dès que les métadonnées devaient être utilisées pour des activités de pilotage de la recherche, la connexion avec des référentiels à caractère administratif devenait critique. Ce dernier exemple étend la réflexion aux référentiels terminologiques.

5. La conclusion initiale

Nous reprenons ici la conclusion qui figurait dans l'article original.

Pour cette nouvelle « expérience rédactionnelle », l'écriture de cette article à la suite de la traduction de « *What is a Digital Library anyway anymore ?* », nous avons choisi de travailler avec un angle plus technique. Dans ce cadre, nous avons identifié un large spectre d'ingrédients (*stuff*)[42], tels que les thèses, les relations d'affiliation, les éléments de vocabulaire qui peuvent enrichir nos services. Nous avons cherché à montrer l'importance des référentiels de tous ordres (terminologie, affiliations auteurs) qui complètent les métadonnées bibliographiques.

Mais, en réalité, que voulons-nous faire au juste ? (“*what do we really want to do anyway, anymore* ?”)

Notre objectif commun est de nous inscrire dans la dynamique de l'e-recherche et d'analyser l'information scientifique et technique dans la globalité des besoins de la recherche.

Comme nous travaillons dans des institutions séparées qui ne partagent pas tout à fait les mêmes objectifs ou priorités, l'écriture de cet article n'a pas été une tâche facile. Notre

[42] La traduction de stuff, terme souvent utilisé dans l'article de Carl Lagoze, a fait l'objet d'une discussion argumentée dans un forum :
< http://artist.inist.fr/article.php3 ?id_article=250 >

résultat le plus intéressant a été peut-être d'identifier les compromis que nous devons assumer :

1. compromis entre le contexte national des thèses et l'échelle internationale ;
2. compromis entre les métiers pour rendre les métadonnées réutilisables à travers leurs applications respectives ;
3. compromis entre les besoins : bibliographiques, infométriques, pilotage de la recherche, réseaux sociaux de l'activité scientifique. (avec un regard particulier sur la question de l'évaluation : en principe, le statut de la thèse témoigne d'une validation, et donc d'une confiance, qui se situe au dernier étage du web sémantique) ;
4. compromis entre un regard focalisé sur les thèses et leur intégration dans un contexte plus large, qui dépasse même la notion de bibliothèque – même qualifiée de numérique.

En résumé, les thèses sont un nœud dans une vaste constellation, qui contient bien les "articles, thèses, affiliations, vocabulaires", mais aussi les projets, les données, les brevets..., autrement dit tous les composants d'un Système d'Information sur les Recherche en Cours[43] qui croise le mouvement du libre accès comme le montre un article de K. Jeffery [11]. L'idée est que la recherche (comme toute activité d'ailleurs) doit être vue à travers ses produits (documents, brevets, données) comme à travers son activité vivante (réseaux sociaux, projets, évaluation, colloques...).

En l'absence de projets transversaux fédérateurs (type NSDL), ARTIST essaye de donner un cadre aux acteurs de terrain pour échanger et expérimenter sur les nouvelles pratiques de production d'Information Scientifique et Technique.

Cet article est d'ailleurs un exemple de ce que nous voulons expérimenter de façon plus permanente. ARTIST a démarré comme un « blog scientifique et coopératif ». Nous venons de lancer AMETIST[44], une revue électronique, francophone, dotée d'une sélection par les pairs, cadre d'expérimentations pour l'écriture numérique. La pratique du français ne doit pas être considérée comme une limitation car nous pensons que les nouveaux concepts peuvent trouver un avantage à progresser en profondeur dans un cadre de langue naturelle avant une confrontation internationale.

Les métadonnées sont un cadre naturel de partage d'expérience et une voie de collaboration que nous allons poursuivre. L'adaptation des travaux du DCMI au contexte de l'intégration des éléments francophones dans un espace mondial de la recherche est une excellente base de travail pour poursuivre la réflexion initialisée dans cet article.

6. Une suite à la conclusion

Cet article a donc été accepté et présenté à la conférence DC 2006 à Manzanillo (Mexique), ce qui nous a permis de confronter la réalité française avec la situation internationale. Nous avons été interpelés par la faiblesse de la présence « officiellement francophone » (une seule personne au milieu de 180 participants) et par la consistance de la création d'un véritable réseau de l'e-Recherche.

[43] CRIS : Current Research Information System
[44] < http://ametist.inist.fr/>

Comment citer ce document :
Ducloy, J., Nicolas, Y., Le Hénaff, D., Foulonneau, M., Grivel, L., Ducasse, J.-P. (2007). Métadonnées dans le contexte d'une cyberinfrastructure de la recherche. "Le cas des thèses françaises". Ametist (1), 75-96.

En effet, nous avons pu constater une participation très significative de trois réseaux d'acteurs :

- aux Etats-Unis, NSDL, souvent cité dans cet article,
- au Royaume-Uni, UKOLN[45] une initiative soutenue notamment par le JISC,
- dans le monde hispanisant, le réseau Scielo qui constitue une large bibliothèque numérique à partir de publications en libre accès.

A travers divers ateliers ils construisaient les bases d'une fédération de bibliothèque numérique dont la francophonie est un peu en retrait, sinon absente. Nous nous attentions donc à une audience assez critique et nous avons été surpris par des réactions plutôt encourageantes qui se résument en : « vous avez entre les mains un magnifique projet potentiel ».

Autrement dit, nous avons les moyens et les compétences. Nous disposons de spécialistes motivés (par exemple les rédacteurs de cet article). Il manque une volonté nationale telle que celle qui se manifeste au sein du JISC ou du NSDL. A court terme, le véritable enjeu de l'appropriation n'est donc pas seulement celui des technologies ou des pratiques individuelles ou collectives. Il se situe au niveau des politiques de la recherche qui doivent s'approprier l'Information Scientifique et Technique, et la considérer comme un carburant essentiel de la machine à produire de la recherche dans un contexte mondialisé.

Remerciements

Seuls les principaux contributeurs sont cités comme auteurs de cet article.

Nous voudrions remercier ici tous ceux qui ont participés par leurs avis, leurs informations et leurs conseils (Francis ANDRE, Catherine MOREL-PAIR, Clotilde ROUSSEL et Pierrette PAILLASSARD de l'INIST ; Amos DAVID du LORIA, Daniel CHARNAY du CCSD ; Ghalia MRAHI de l'IMIST ; Estelle BALIAN de "BiodivErSA - Belgium Biodiversity Platform") ou qui nous ont aidés dans les phases de traduction ou de révision (Marc RUBIO et Catherine GUNET de l'INIST).

Bibliographie

[1] Y. Bakelli et S. Benrahmoun. Long-term preser-vation of ETDs in Algeria : discussion through the CERIST Deposit system. In Proceedings of ETD2003. Berlin 2003.
< http://edoc.hu-berlin.de/conferences/etd2003/bakelli-yahia/HTML/bakelli.html >

[2] BiodivERsA. Compendium of Biodiversity Re-search Funding Agencies in Europe <
http://www.eurobiodiversa.org/rich_files/attachments/Compendium%201%20Feb%202006rev.doc >

[3] S. Brin et L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, in Proceedings of the seventh international conference on World Wide Web 7, Brisbane, Australia, 1998

[4] F. Cappello, E. Caron, M. Dayde, F. Desprez, E. Jeannot, Y. Jegou, S. Lanteri, J. Leduc, N. Melab, G. Mornet, R. Namyst, P. Primet et O. Richard. Grid'5000: a large

[45] UK Office for Library Networking
<http://www.ukoln.ac.uk/>

scale, recongurable, controlable and monitorable Grid platform. In Grid'2005 Workshop, Seattle, USA, November 13-14 2005. IEEE/ACM.

[5] ETD-MS : an Interoperability Metadata Standard for Electronic Theses and Dissertations < http://www.thesis.org/standards/metadata/current.html >

[6] David J. Farber; K. Larson (Sept 1970). "The Architecture of a Distributed Computer System - An Informal Description". *Technical Report Number 11*, University of California, Irvine.

[7] L. Grivel, H. Fagherazzi, P. Fourneret and A. Zerouki. La conception de bases de données infométri-ques hybrides : analyse de la pratique de trois observa-toires européens. In Journées SFBA proceedings Ile Rousse 99. < http://archivesic.ccsd.cnrs.fr/sic_00000464.html >

[8] S Harnad. Publish or Perish - Self-Archive to Flourish : The Green Route to Open Access. In *ERCIM News* January 2006 <http://www.ercim.org/publication/Ercim_News/enw64/harnad.html>

[9] D. Le Henaff and C. Thiolon. Gérer et diffuser des thèses électroniques : un choix politique pour un enjeu scientifique. In Documentaliste - Sciences de l'information. 42(4-5):272-280. October 2005.

[10] Institute of Higher Education. Academic Ranking of World Universities. Shanghai Jiao Tong University, 2005 < http://ed.sjtu.edu.cn/ranking.htm >

[11] K. Jeffery. CRIS + open access = the route to research knowledge on the GRID. In 71st IFLA General Conference and Council proceedings, Oslo, Norway, 2005 <http://www.ifla.org/IV/ifla71/papers/007e-Jeffery.pdf>

[12] M. Kaiser. New Ways of Sharing and Using Au-thority Information. In D-lib Magazine, September 2001 < [http://www.dlib.org/dlib/november03/lieder/11lieder.html]>

[13] C. Lagoze, D. Krafft, S. Payette and S. Jesuroga. What Is a Digital Library anyway, anymore ? In D-lib Magazine. November 2005. < [http://dx.doi.org/10.1045/november2005-lagoze]>

[14] [11] C. Lynch. Where Do We Go From Here ? The Next Decade for Digital Libraries. In *D-lib Magazine*, July 2005 < doi:10.1045/july2005-lynch >

[15] M. Patel, Fourth Open Archives Forum Workshop In Practice, Good Practice : The Future of Open Archives, Ariadne Issue 37, Oct 2003, < [http://www.ariadne.ac.uk/issue37/oa-forum-ws-rpt/#6]>

[16] [13] X. Polanco, L. Grivel, J. Royauté -'How to do things with terms in informetrics : terminological variation and stabilization as science watch indicators'- 5th Int. Conf. of the International Society for Scientometrics and Informetrics -, Chicago, Illinois, pp.435-444, 1995.

[17] [14] M. Smith, M. Bass, G. McClellan, R. Tansley, M. Barton, M. Branschofsky, D. Stuve, and J. H. Walker, DSpace : An Open Source Dynamic Digital Repository, D-Lib Magazine, 9 (1), 2003. < [doi:10.1045/january2003-smith]>

[18] [15] H. Suleman, A. Atkins, M. Gonçalves, R. France and E. Fox. Networked Digital Library of Theses and Dissertations, Bridging the Gaps for Global Access - Part 1 : Mission and Progress. In D-lib Magazine, September 2001 < [doi:10.1045/september2001-suleman-pt1]>

[19] M. Welshons. "Our Cultural Commonwealth" The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities

and Social Sciences. Connexions. 15 Dec. 2006 <http://cnx.org/content/col10391/1.2/>.