Développement de la Veille à l'INRS : Approches et Retours d'Expériences

Françoise Grandjean, Guillaume Moureaux, Michel Servais

Développement de la veille à l'INRS : approches et retours d'expériences

Françoise Grandjean (1)
francoise.grandjean@inrs.fr

Guillaume Moureaux (2)
guillaume.moureaux@inrs.fr

Michel Servais (3)
michel.servais@inrs.fr

(1) INRS, avenue de Bourgogne, 54500 Vandoeuvre 03 83 50 21 56
(2) INRS, avenue de Bourgogne, 54500 Vandoeuvre 03 83 50 20 00
(3) INRS, avenue de Bourgogne, 54500 Vandoeuvre 03 83 50 21 34

Mots-clés : institut recherche, sécurité travail, veille, retour expérience, analyse information, système information, logiciel, prototype, description système, INRS, France, Dilib

Keywords : research institute, work safety, wakefulness, experience feedback, information system, software, system description

Résumé : Depuis 1995, le centre de documentation du site de l'INRS (Institut National de Recherche et de Sécurité) situé à Vandoeuvre, a expérimenté différentes approches de l'infométrie et de la veille dans le but de soutenir les activités de recherches de son site. Cet article présente le cheminement, les réflexions et les difficultés qui ont marqué ce parcours.

Ce parcours a permis d'acquérir de l'expérience et de susciter un intérêt pour la veille. Des projets pilotes ont été menés sur des sujets d'études tels que le stress, la génomique du mésothéliome, les risques biologiques émergeants ou les particules ultrafines. Ils ont été réalisés en collaboration avec le LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications) et l'INIST (INstitut de l'Information Scientifique et Technique) et se sont essentiellement appuyés sur la plate-forme DILIB (Digital Library) développée par ces organismes. Parallèlement, une sensibilisation à la veille a été conduite grâce à un séminaire de réflexion interne, des

stages de formation et la création d'une rubrique de communication interne dédiée à ces sujets sur l'intranet de l'INRS.

Aujourd'hui, la nécessité d'organiser et de structurer une activité de veille au sein de l'INRS s'impose en raison de la quantité accrue d'information disponible, du développement des NTIC (Nouvelles Technologies de l'Information et de la Communication) mais aussi des choix stratégiques dans l'orientation des études et des actions menées en matière de prévention des risques professionnels.

Notre démarche a permis de mettre en évidence la nécessité d'homogénéiser l'accès à l'information et de structurer les informations réunies selon un modèle commun pour une exploitation collective efficace. Mais les aspects humains et relationnels se révèlent problématiques car il s'agit de convaincre et de dépasser les réticences que soulève la mise en place de nouveaux processus.

Introduction

Cet article vise à montrer comment les méthodologies de veille ont fait leur chemin progressivement au sein de l'INRS (Institut National de Recherche et de Sécurité). Il s'agit pour nous de partager notre approche, nos expériences, nos difficultés et nos observations.

L'objectif de cet article est aussi de montrer comment nous avons acquis précocement une connaissance profonde de la discipline et avons préparé la culture d'entreprise afin de permettre l'introduction de la veille à l'INRS non pas comme une révolution mais plutôt en douceur et comme une évidente nécessité.

Ce retour d'expériences mettra en évidence une démarche d'abord de type expérimental avec le développement d'outils d'exploration et d'analyse de l'information faisant appel aux techniques de l'infométrie. Il montrera comment nous avons parallèlement travaillé à une nécessaire sensibilisation des personnels à tous les niveaux de responsabilité de l'institut, du chercheur au directeur en passant par les chefs de projet.

Il montrera également comment nous avons réalisé des prototypes d'outils de collecte, d'analyse et de diffusion et d'analyse de

l'information aussi bien dans le cadre d'un système informatif opérationnel que d'un système informatif de management et de décision. Et nous évoquerons les travaux qui sont aujourd'hui en perspectives ainsi que les difficultés et obstacles que nous avons aujourd'hui à franchir.

1 Importance de l'information dans les missions de l'INRS

Que ce soit à l'échelle des individus ou des sociétés, aujourd'hui, chacun doit être compétitif. De l'invention à la production de masse tout doit aller plus vite dans la course à l'innovation. Cependant, il s'agit de ne pas oublier de préserver les hommes et les femmes acteurs de cette compétition.

C'est le rôle du système français de prévention des risques professionnels. Ce rôle consiste à protéger les individus face aux risques qui apparaissent notamment sous l'effet de cette pression économique. Mieux encore, les acteurs de ce système se doivent, autant que possible, de précéder cette course en proposant des solutions de prévention.

Au sein de ce dispositif l'INRS est dépositaire des connaissances scientifiques et techniques dont le système de prévention a besoin pour mettre en œuvre cette prévention. Le rôle de cet institut se décline en trois missions majeures :

Anticiper : Du risque toxique au bien-être physique et psychologique, l'INRS conduit des programmes d'études et recherches pour améliorer la santé et la sécurité de l'homme au travail. Le bilan de ces actions lui permet également de déterminer les besoins futurs en prévention. Tous les cinq ans, un programme définit son cadre général d'action.

Sensibiliser : L'institut conçoit de nombreux produits d'information : 4 revues, 300 brochures, 150 affiches, 70 vidéos, des cédéroms, un site internet. Ils sont diffusés auprès d'un large public, composé de chargés de sécurité, médecins du travail, ingénieurs, opérateurs,

formateurs... Certaines actions ponctuelles font l'objet de campagnes de prévention auprès du grand public.

Accompagner : L'INRS propose une aide technique aux entreprises : 40 000 demandeurs y font appel chaque année pour résoudre un problème de prévention. L'institut transmet son savoir-faire et ses compétences par 70 offres de formation ou d'aides pédagogiques adaptées aux besoins des animateurs de la prévention en entreprise. Ses experts participent à de nombreux groupes de travail, nationaux, européens ou internationaux, pour la rédaction de textes à caractère réglementaire ou normatif.

L'INRS est réparti sur trois sites, à Paris, à Vandoeuvre et à Neuves-Maisons, chacun disposant d'un centre de documentation adapté à ses activités. Sur le site de l'INRS situé à Vandoeuvre, la documentation traditionnelle est depuis toujours étroitement liée aux travaux d'études et de recherches qui constituent l'essentiel de l'activité scientifique. Elle apporte sa contribution dès l'origine de ces projets et les accompagne tout au long de leur réalisation.

Les trois services de documentation de l'INRS mettent à disposition sur le réseau interne l'ensemble des sources documentaires dont ils disposent. En effet, les 3 centres de documentation de l'INRS bien que de nature et de missions différentes, se sont regroupés pour offrir un accès commun à l'information scientifique et technique sur le site intranet de l'institut (Interligne) (voir illustration 1).

Illustration 1 : Interligne : Le site intranet de l'INRS.

Interligne permet d'accéder à des services de bases de données internes mais aussi externes telles que BiblioSciences de l'INIST, (Pascal, Francis, Inspec, Current Contents…), Kompass, Perinorm. Interligne ouvre également l'accès vers un serveur de cédérom offrant entre autres, l'accès aux bases diffusées par le centre canadien d'hygiène et de sécurité (CCHST).

L'utilisateur d'Interligne a aussi accès aux revues auxquelles les centres INRS de Paris, Vandoeuvre et Neuves Maisons sont abonnés. Elles sont présentées accompagnées de l'état de leur collection, du nom de la personne à contacter pour y accéder en version papier et d'un lien direct sur le site de l'éditeur pour l'accès aux sommaires voire au texte intégral lorsque la revue existe également au format électronique.

En effet depuis 2 ans, l'accès au texte intégral des articles de périodiques s'est beaucoup développé. Après une période expérimentale, où l'accès à la version électronique d'une revue était compris dans l'abonnement à la version papier, une politique d'accès payant s'est mise en place progressivement. Il a donc fallu trouver une solution pour faire face à cette tendance.

C’est dans ce but, que fin 2002, l’INRS a adhéré au consortium COUPERIN qui regroupe la plupart des universités et des organismes de recherche publics français. Ainsi un accord a pu être signé avec l’éditeur Elsevier pour accéder aux 1700 revues électroniques en texte intégral du service ScienceDirect

2 Des approches pour se familiariser avec l’infométrie et la veille

Parallèlement à cette documentation en évolution, depuis 1995, des expériences ont été menées sur le développement d’applications de traitement et d’analyse de l’information grâce à des collaborations informelles établies avec le LORIA et l’INIST. En effet, Jacques Ducloy et ses collaborateurs au LORIA puis à l’INIST, ont développé la plate-forme d’investigation documentaire, DILIB, qui est une bibliothèque de puissantes fonctions de traitement de l’information structurée au format XML. Elle a été implantée sur le serveur du Centre de Services Informatiques de l’INRS et a été utilisée pour la réalisation de nos différents outils de traitement de l’information.

L’objectif initial était de développer un outil muni d’une interface hypertexte permettant un accès intuitif à des fonds documentaires. L’outil a été utilisé pour exploiter les fonds documentaires généraux des centres INRS de Paris et de Lorraine, ainsi que des bases documentaires personnelles de chercheurs de l’institut. Il permettait de visualiser les fréquences et les associations de mots du titre, mots du résumé, descripteurs et auteurs mais aussi de naviguer dans les notices bibliographiques. Au-delà de l’objectif initial, ces interfaces de consultations étaient en fait de véritables serveurs infométriques munis également de représentations graphiques de l’information permettant d'explorer des fonds documentaires avec une optique d'analyse.

Au début des années 2000, cette expérience a été complétée lorsque ces outils ont été mis en œuvre à la demande de chercheurs pour explorer non plus des bases internes mais des fonds résultant de l’interrogation de bases de données externes comme MEDLINE,

NIOSH, PsycINFFO sur des thématiques intéressant l'institut. Ces travaux en collaboration avec le LORIA, ont notamment abouti à la réalisation des applications WebStress qui avait pour objectif d'explorer le vaste fonds des publications concernant les problèmes de stress au travail et Transcriptome/Bibliome (voir illustration 2) en collaboration avec l'INIST qui exploitait des documents traitant de l'expression génétique du mésothéliome, tumeur liée à l'exposition à l'amiante.

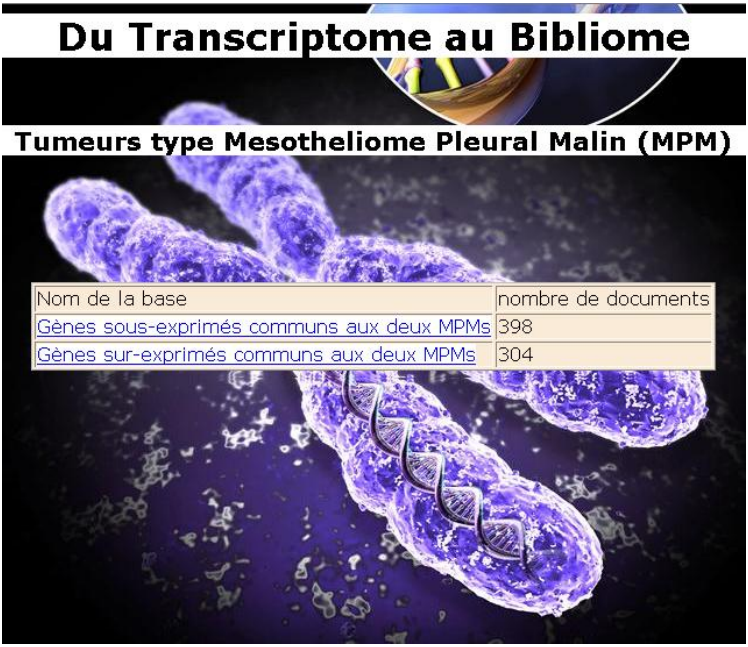


Illustration 2 : Serveur documentaire sur le mesotheliome.

Cependant, après ces premières expériences et avant d'aller plus loin dans le domaine de l'analyse de l'information, l'idée est apparue que l'INRS devait mener une réflexion sur ce qui précède cette analyse : les problèmes de collecte de l'information et donc la veille. De plus, les participants aux premières expériences ayant été enthousiasmés, l'idée était aussi d'organiser une opération visant à sensibiliser l'ensemble du personnel de l'INRS à la veille.

Septembre 2003 a donc vu l'organisation d'un séminaire interne concernant la veille. Une journée de sensibilisation permet alors de faire avancer la réflexion interne du public INRS en situant la veille dans un cadre de prospective stratégique. Tables rondes, retours d'expérience d'organismes externes et expériences internes se déroulent au long de cette journée. Les différentes déclinaisons de la veille sont proposées : veille pour une production de connaissances, veille pour des décisions de société, aspects politico-stratégiques et médiatiques, structures et outils internes.

Parallèlement, la documentation développe alors, sur son site intranet, une nouvelle rubrique consacrée à la veille grâce au concours d'une étudiante (Christelle Martin), du DESS Information Scientifique et Technique et Intelligence Economique (ISTIE) de Nancy. Sa mission était de réaliser une étude comparative des outils existants et de les appliquer à titre d'exemple à la thématique des risques biologiques émergents et plus particulièrement aux risques biologiques dans les métiers du bois. Le site en question répertorie et décrit sources d'informations, méthodes et outils pour la veille, le but étant de mieux appréhender les spécificités de chaque outil et d'aider tout un chacun à les mettre en œuvre pour ses propres besoins. En 2003 et 2004 cette rubrique intranet est accompagnée de stages de formation interne sur le thème de la veille animés par l'INIST.

Plus récemment, nous avons pu étudier et tester un nouveau produit d'analyse de fonds documentaires, présenté par l'équipe Orpailleur du LORIA (Emmanuel Nauer). Cet outil nommé IntoBib est issu de la technologie DILIB mais complété par les technologies PHP et SQL. Le but est de fournir, au chercheur ou au spécialiste de l'information scientifique et technique, un environnement dans lequel il puisse exploiter les données issues de sa veille, de façon dynamique cette fois, contrairement aux serveurs d'investigation classiques dont les explorations sont prévues dès la construction du serveur.

```
<ref>
  <TITR>Joining the trek with Keith up the Serpentine
Road--the lattice from another perspective.</TITR>
  <AUTE>
    <e>Wolosewick, J J</e>
  </AUTE>
  <SOUR>Biol-Cell. 2002 Dec; 94(9): 557-9</SOUR>
  <JOUR>Biology of the cell under the auspices of the
European Cell Biology Organization</JOUR>
  <ISSN>0248-4900</ISSN>
  <YEAR>2002</YEAR>
  <LANG>
    <e>English</e>
  </LANG>
  <PAYS>France</PAYS>
  <DEEN>
    <e>Cytoplasm chemistry</e>
    <e>Cytoskeleton chemistry</e>
    <e>Organelles chemistry</e>
    <e>Cytoplasm ultrastructure</e>
    <e>Cytoskeletal Proteins analysis</e>
    <e>Cytoskeletal Proteins ultrastructure</e>
    <e>Cytoskeleton ultrastructure</e>
    <e>Microscopy, Electron methods</e>
    <e>Organelles ultrastructure</e>
  </DEEN>
  <TYPE>
    <e>Editorial</e>
  </TYPE>
</ref>
```

Illustration 3 : Données au format XML pour l'application amiante.

Des fonctionnalités de fouille (dénombrements, classifications, extractions de règles, etc.) peuvent être déclenchées à la demande pour analyser plus précisément certains sous-ensembles de données. Le principe technique est que les actions de l'analyste sur l'interface hypertexte sont traduites en requêtes SQL et les résultats de traitement traduits en graphiques et en chiffres. Les temps de traitement sont très courts grâce à la conception du serveur qui pré calcule les résultats lors de sa génération à l'aide de fonctions PHP.

A titre d'essai, nous avons expérimenté l'outil avec un corpus de références bibliographiques sur l'amiante issu de MEDLINE. Là encore le savoir-faire acquis au sein de l'INRS a permis la transformation de ces données au format XML (voir illustration 3) pour l'importation directe dans cette application. Les résultats qui se sont dégagés se sont avérés intéressants dans la mesure où ils indiquaient des tendances en présentant des pics d'intérêt et de publications pour le sujet, liés aux dates des décisions politiques ou aux échos médiatiques.

3 Des prototypes de systèmes informatifs

En 2004 une étude d'instruction de projet sur les risques professionnels liés aux particules ultrafines est initiée par un laboratoire de l'INRS. Le chargé de projet (Olivier Witschger) dont la mission est de faire le point sur ces risques et sur l'intérêt de lancer une étude sur ce sujet, est intéressé par la mise en place d'un processus de veille pour soutenir l'étude. Sensibilisé par le précédent séminaire de veille il avait constaté que la documentation traditionnelle et les outils mis à sa disposition ne suffisaient pas pour répondre à son besoin. Nous décidons alors d'avoir recours aux services d'un veilleur, étudiant du DESS ISTIE (Guillaume Moureaux) dans le cadre de son stage d'étude en collaboration avec les membres de la cellule veille de l'INIST (Catherine Czysz, François Parmentier, Philippe Houdry et Solveig Vidal) qui lui transmettent leur savoir faire et nous donnent accès à certains de leurs moyens d'investigation.

La première tâche consiste alors à réunir un fonds de références bibliographiques sur le sujet des particules ultrafines. Pour cela les bases documentaires mises à disposition par l'INIST au travers du service BiblioSciences sont employées. Ce service présente l'intérêt de permettre l'interrogation de 11 bases de données externes avec une seule interface d'interrogation et de permettre le téléchargement des résultats sous un format unique. Un fonds documentaire sera donc effectivement constitué après plusieurs réunions avec les chercheurs impliqués dans le projet pour définir le besoin et préciser le vocabulaire requis pour l'interrogation des sources.

Illustration 4 : La base de donnée documentaire particules ultrafines par l'interface CinDoc Web

Mais à cette étape on n'a encore qu'un résultat brut puisqu'il reste encore à traiter le fonds collecté afin de le diffuser et de l'exploiter. Une rapide analyse de l'existant permet de faire le point sur les outils logiciels disponibles à l'INRS et utilisables à cet effet. Le choix est fait de créer une base de notices bibliographiques à l'aide du logiciel documentaire CINDOC récemment acquis par l'INRS. Cette base sera accessible grâce à l'interface Web de CINDOC (voir illustration 4) depuis les trois centres de l'INRS constituant ainsi un moyen de diffusion idéal. De plus, le logiciel est muni d'index interrogeables qui permettront une exploitation des données recueillies par les personnes qui devront analyser ce fonds.

Mais avant cela il s'agissait de préparer les données afin de réaliser ce produit. Les données sont dans un format propriétaire incompatible avec une importation immédiate dans une base CINDOC. Il convient donc de transformer les données dans un format approprié. Une phase de formatage et de traitements divers, effectués sous Unix à l'aide des outils de la plate-forme DILIB est donc engagée. La plate-forme DILIB est d'abord mise à jour et les outils de développement installés avec le concours du Centre de Services Informatiques de l'INRS (Michel Servais).

Le développement des programmes de formatage commence alors. Il s'agit de passer par un format XML qui servira de format transitoire. Le format propriétaire de BiblioSciences est donc transformé en format XML qui sera lui même ensuite transformé en format Ajout Piloté. L'Ajout Piloté est en effet le format d'importation des données de CINDOC qui permettra de constituer la base documentaire.

Dans le même temps, d'autres traitements intermédiaires sont ajoutés à la chaîne de formatage. En effet, à l'étape du format XML il devient possible d'appliquer des traitements supplémentaires comme le dédoublonnage et l'indexation semi-automatique des références bibliographiques.

L'objectif de ces outils complémentaires est d'effectuer un retour vers des problématiques d'analyse en se proposant d'explorer des fonds issus des bases de données externes.

Le dédoublonnage est en effet un pré requis pour pouvoir réaliser ensuite une analyse infométrique du fonds. Quant à l'indexation semi-automatique elle permet de marquer les références bibliographiques avec les termes que l'on veut analyser. Le principe étant de détecter des termes ou expressions spécifiques présents dans les notices bibliographiques et de les représenter par un terme générique qui sera ensuite analysé en termes de fréquence.

Ces travaux permettent de produire des historiques des publications sur tel ou tel sujet sous forme de représentations graphiques. Grâce à cette application on peut montrer par exemple que le nombre de publications concernant un type de risque ne progresse plus alors que l'industrie correspondante est florissante. On peut ainsi se poser la

question de l'intérêt de la proposition d'autres études dans ce domaine. Cette application constitue donc un système informatif décisionnel à destination de la direction scientifique de l'institut. Afin de tester ce système d'analyse, de tels travaux ont été réalisés à titre d'exemple sur des thématiques connues de longue date ou bien nouvelles comme les nanoparticules, les éthers de glycol, les troubles musculo-squelettiques, le stress, les fibres céramiques, le traitement au niveau de l'indexation permettant de différencier les aspects prévention, épidémiologiques et toxicologiques.

L'objectif ici est évidemment de susciter l'intérêt et de sensibiliser les directions en montrant l'apport d'une veille stratégique et prospective. Cet outil pourrait en effet contribuer à la prise de décision concernant les orientations scientifiques de l'institut dans le cadre du Plan à moyen terme quinquennal 2008-2012.

4 Perspectives

Où en sommes-nous aujourd'hui ? Actuellement, nous commençons à développer, en collaboration avec Michel Servais, un système automatisé d'alerte, basé sur les profils proposés par ScienceDirect de Elsevier. Il s'agit de construire des équations de recherche pour chacun des sujets des 7 projets transversaux en cours au sein de l'INRS. Les alertes envoyées par ScienceDirect seront redirigées vers le serveur de messagerie de l'INRS, basculées dans une base de données de type SQL et après validations sélectives des résultats, diffusées dans une rubrique « Veille PTI » d'Interligne (voir illustration 5).


Robot de Surveillance Documentaire (SERVEUR : 192.168.2.21)
139 documents sélectionné(s) au 19/04/2006

SCIENCE DIRECT Alerte(s) : stress+auteurs+psychosocial

Page: 2 / 20 de la liste des 139 document(s) [début | suite | précédent | fin]

2006-03-28 **Work related post-traumatic stress as described by Jordanian emergency nurses**
Accident and Emergency Nursing, In Press, Corrected Proof, Available online 27 March 2006
Anders Jonsson and Jehad Halabi

2006-03-25 **Observational Stress Factors and Musculoskeletal Disorders in Urban Transit Operators**
Journal of Occupational Health Psychology, Volume 11, Issue 1, January 2006, Pages 38-51
Birgit A. Greiner and Niklas Krause

2006-03-25 **Relationships Among Organizational Family Support, Job Autonomy, Perceived Control, and Employee Well-Being**
Journal of Occupational Health Psychology, Volume 11, Issue 1, January 2006, Pages 100-118
Cynthia A. Thompson and David J. Prottas

2006-03-23 **Rapports au travail, contrôle et santé dans les centres de gestion de la relation–client**
Psychologie du Travail et des Organisations, In Press, Corrected Proof, Available online 22 March 2006
M. Lourel

2006-03-23 **Confirmatory factor analysis of posttraumatic stress symptoms in emergency personnel: An examination of seven alternative models**
Personality and Individual Differences, In Press, Corrected Proof, Available online 22 March 2006
Leanne Andrews, Stephen Joseph, Mark Shevlin and Nick Troop

2006-03-21 **Musculoskeletal complaints and psychosocial risk factors among physicians in mainland China**
International Journal of Industrial Ergonomics, In Press, Corrected Proof, Available online 20 March 2006
Derek R. Smith, Ning Wei, Yi-Jie Zhang and Rui-Sheng Wang

Illustration 5 : Les alertes ScienceDirect accessibles sur le site intranet

Que ressort-il du parcours effectué ? Au niveau de la Documentation, l'intérêt est évident, dans la mesure où c'est le virage qu'elle doit prendre rapidement pour évoluer vers d'autres missions.

Le rôle traditionnel de recherche et de fourniture d'informations bibliographiques a été remis en cause par l'explosion de sources documentaires électroniques à disposition des chercheurs via l'intranet. La Documentation s'est aussi attachée à les aider à les utiliser et les exploiter (formation des utilisateurs sur les sources, aide au démarrage d'une recherche bibliographique, constitution de bases de données, mise en place d'alertes..). Elle se doit de jouer désormais un rôle moteur dans l'analyse et l'exploitation des données bibliographiques dans le cadre de la veille.

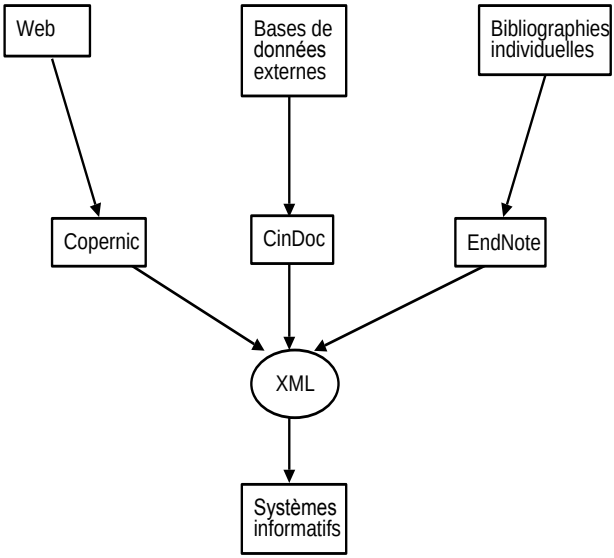


Illustration 6 : Modélisation de prototypes de systèmes informatifs avec les outils de l'INRS

Tandis que les technologies de l'information font leur progrès nous tentons autant que possible d'apporter des innovations dans notre manière de pratiquer la gestion de l'information à l'INRS. Nous suivons pour cela une démarche expérimentale qui nous permet de nous former aux méthodes et aux technologies qui apparaissent. Cette acquisition d'expérience nous permet d'aller chaque fois plus loin et de rendre l'INRS progressivement autonome dans sa gestion de l'information sur la base des expériences et savoir-faire acquis dans le cadre de collaborations.

Par exemple, nous avons commencé à acquérir une certaine pratique des technologies XML. Cette maîtrise nous permet aujourd'hui et de plus en plus, d'échanger des données entres différentes applications. (voir illustration 6) Dans ce domaine nous avons notamment développé des passerelles logicielles faisant appel à la technologie XML. Elles permettent dans un cas de regrouper des bibliographies

constituées sous EndNote en une base documentaire CINDOC. Cela constitue une sorte d'outil de veille collective rassemblant les recherches bibliographiques des membres d'un groupe de travail donné. Nous avons également constitué des outils destinés à monter un système de veille orienté web.

Ces expériences nous ont permis d'apprendre à réaliser différents prototypes à toutes les étapes du processus de veille et de tester sur des groupes d'utilisateurs les solutions les plus appropriées à la culture d'entreprise de l'INRS. De même, nous aurions pu faire le choix d'investir dans des solutions commerciales clés en main. Mais nous avons préféré faire d'abord nos expériences afin de pouvoir le moment venu être capables d'une véritable maîtrise des outils et de l'exploitation de ces outils à leur plein potentiel.

Dans le même temps, nous nous sommes employés à faire progresser l'idée qu'une gestion moderne de l'information à l'INRS qui pourrait profiter non seulement à chacun, mais encore plus à l'ensemble et qu'il devient nécessaire d'avoir une politique globale dans ce domaine. De plus, chacun commence à réaliser par lui même qu'il ne peut pas maîtriser à lui seul la masse d'information qui se déverse sur le web, dans l'intranet ou bien envahit sa boite de messagerie et son bureau.

En fait de société de l'information, nous serions plutôt actuellement dans une société de la surinformation ou chacun a la responsabilité ou parfois le besoin vital de ne rien ignorer mais n'a pas le temps matériel de traiter la quantité d'information qui lui incombe. C'est ainsi que certains chercheurs de l'INRS cherchent à monter des groupes de travail non pas pour mettre en commun l'information mais pour en partager l'analyse dans des processus d'intelligence collective. Car le problème n'est plus de trouver l'information, mais de savoir laquelle est l'information adéquate et pertinente.

Si pendant un temps, à l'INRS, la tendance a consisté à donner à l'utilisateur tous les moyens de s'informer par lui-même, aujourd'hui, cette situation tend à s'inverser. Des chercheurs commencent en effet à revenir chercher de l'aide auprès des spécialistes pour trouver l'information ou plutôt pour la trier puis pour l'analyser lorsqu'elle est trop abondante. Cependant l'utilisateur conserve le désir d'être autonome et a parfois du mal à admettre que

des médiateurs sont nécessaires, à l'interface, entre lui et l'information.

Pour ce qui concerne l'avenir de l'INRS, 2005 voit se définir des projets à l'échelle de chaque département scientifique mais aussi des projets transversaux rassemblant les moyens de plusieurs départements qui ont des besoins croissants de collecte, de tri, de validation, d'analyse et de diffusion de l'information. Pour 2005-2008, l'INRS a été retenu comme maître d'œuvre du projet européen d'observatoire des risques professionnels dont le principe même consiste à surveiller la littérature scientifique et toutes les autres sources d'information.

Cette offre nouvelle en matière d'outils et de méthodes, va pouvoir aider à faire des choix stratégiques au moment où se met en place la préparation du prochain Plan à Moyen Terme et sans doute permettre à chacun de mieux repérer et exploiter l'information utile pour les missions de l'INRS.

Bibliographie

[1] Rihn b., Mohr s., Grandjean f., Nemurat c.,From transcriptomics to bibliomics.,Medical sciences monitor, vol.9, no.8, 2003, pp.89-95.

[2] Grandjean f., Mur d., Puzin m., Ciccotelli j., Falcy m.,Séminaire interne "La veille", Institut National de Recherche et de Sécurité, INRS, Vandoeuvre, 18 septembre 2003.,Vandoeuvre, Institut National de Recherche et de Sécurité, INRS, 2003, 47p.

[3] Jolibois s., Mouze-Amady m., Chouaniere d., Grandjean f., Ducloy j.,WEBSTRESS : a web-interface to explore a multidatabase bibliographic corpus on occupational stress,Work and stress, vol. 14, no 4, octobre 2000 pp.283-296.

[4] Jolibois s., Chouaniere d., Ducloy j., Grandjean f., Mouze-Amady m.,Un exemple d'utilisation de l'ULMS, base multilingue de connaissances biomédicales.,Documentaliste, vol.37, no.2, juin 2000, PP.94-103.

[5] Jolibois s., Nauer e., Chouaniere d., Mouze-Amady m., Ducloy j., Grandjean f.,Standardisation of a multidatabase bibliographic corpus.,Consensus Workshop on "stress at work" organise par l'AMI (UK),

Copenhague, 21-22 juin 1999,Vandoeuvre, Institut National de Recherche et de Sécurité, INRS, Service Epidémiologie en Entreprise, EE, 1999

[6] Ducloy j., Nauer e., Jolibois s., Grandjean f., Chouaniere d., Mouze-Amady m.,DILIB, how to use an XML technology to build Intranet or Internet services oriented towards scientific survey,Consensus Workshop on "stress at work" organise par l'AMI (UK), Copenhague, 21-22 juin 1999,Vandoeuvre, Institut National de Recherche et de Sécurité, INRS, Service Epidémiologie en Entreprise, EE, 1999

[7] Jolibois s., Chouaniere d., Ducloy j., Grandjean f., Mouze-Amady m.,La gestion informatisée de corpus bibliographiques. Adaptation des normes et formats documentaires.,Bulletin des bibliothèques de France, vol. 45, no. 1, 2000, pp. 998-108.

[8] Jolibois s., Chouaniere d., Mouze-Amady m., Grandjean f., Servais m.,Standardisation of a multibase bibliographic corpus.,Journal of the American Society for Information Service,Vandoeuvre, Institut National de Recherche et de Sécurité, INRS, Service Epidémiologie en Entreprise, EE, 1999

[9] Chouaniere d., Jolibois s. ; Mouze-Amady m., Grandjean f., Francois m.,Une base documentaire sur le stress professionnel.,Travail et sécurité, n° 579, décembre 1998, pp. 8-11, ill.

αMETIST

Appropriation
Mutualisation
Expérimentations
des technologies de l'IST

Numéro 0 - Septembre 2006

SPECIAL SDN

http://ametist.inist.fr

AMETIST http://ametist.inist.fr

Appropriation, Mutualisation, Expérimentations des Technologies de l'Information Scientifique et Technique

---

Equipe technique

DUCLOY Jacques
jacques.ducloy@inist.fr

ROUSSEL Clotilde
clotilde.roussel@inist.fr

GAUTIER Patricia
patricia.gautier@inist.fr

SAFA Djamila
djamila.safa@inist.fr

RASOLOMANANA Magali
magali.rasolomanana@inist.fr

WIRTZ Pierre
pierre.wirtz@inist.fr

(1) 2 allée du parc de Brabois CS 10310 F-54519, Vandoeuvre lès Nancy

Sont également remerciés pour leur aide et leurs conseils GRESILLAUD Sylvie (INIST), NOMINE Jean-François (INIST), HAMEAU Thérèse (INIST).

ISSN PAPIER ISSN WEB

TABLE DES MATIERES

PARTIE 1 : Appropriation : besoins, conditions

Appropriation, mutualisation, expérimentations des technologies de l'information scientifique et technique 11

Sylvie Lainé-Cruzel

1 Positionnement 11

2 AMETIST : une expérience éditoriale 16

3 Quelques thématiques 18

4 Contenu 20

5 Un numéro zéro, prototype… 22

Qu'est-ce qu'une bibliothèque numérique, au juste ? : Au-delà des fonctions recherche et accès dans la National Science Digital Library 25

Carl Lagoze, Dean B. Krafft, Sandy Payette, Susan Jesuroga

1 Construire une bibliothèque numérique avec un entrepôt de métadonnées : phase I de la NSDL 30

2 Utilité de l'entrepôt de métadonnées en tant qu'architecture de bibliothèque numérique 32

3 Modélisation informationnelle pour gérer la complexité et le contexte 35

4 Le Réseau d'Information Superposé (RIS) 42

5 L'entrepôt de données de la NSDL : NSDL phase 2 46

Conclusion 50

Bibliographie 52

PARTIE 2 : Capitalisation/Mutualisation

D'un thésaurus vers une ontologie de domaine pour l'exploration d'un corpus 59

Claude Chrisment, Françoise Genova, Nathalie Hernandez, Josiane Mothe

Introduction 59

1 Présentation de la méthode 62

2 Conceptualisation du lexique du thésaurus 71

3 Construction de la structure de l'ontologie 75

4 Détection des relations associatives 84

Conclusion 89

Bibliographie 91

Développement de la veille à l'INRS : approches et retours d'expériences 95

Françoise Grandjean, Guillaume Moureaux, Michel Servais

Introduction 96

1 Importance de l'information dans les missions de l'INRS 97

2 Des approches pour se familiariser avec l'infométrie et la veille 100

3 Des prototypes de systèmes informatifs 104

4 Perspectives 107

Bibliographie 111

PARTIE 3 : Coups de flash

INCISO : Elaboration automatique d'un index de citations des revues espagnoles en sciences sociales 113

José M. Barrueco, Julia Osca-Lluch, Thomas Krichel, Pedro Blesa, Elena Velasco, Leonardo Salom

Introduction 114

1 Autres travaux sur le sujet 117

2 Méthodologie et déroulement du projet 120

3 Architecture du système 124

Conclusions 127

Bibliographie 128

PARTIE 4 : ARTIST, un lieu d'expérimentations

A propos du numéro zéro d'AMETIST
Rapport sur une expérience d'appropriation 133

Jacques Ducloy, Patricia Gautier, Magali Rasolomanana, Clotilde Roussel, Djamila Safa, Pierre Wirtz

Introduction 133

1 Appropriation des techniques éditoriales 135

2 Autour de l'écriture scientifique technique et numérique 139

3 Autour des traductions : travail coopératif et mise en ligne spécialisée 141

4 Mutations liées à l'appropriation des pratiques éditoriales 146

Conclusion 149

Bibliographie 150

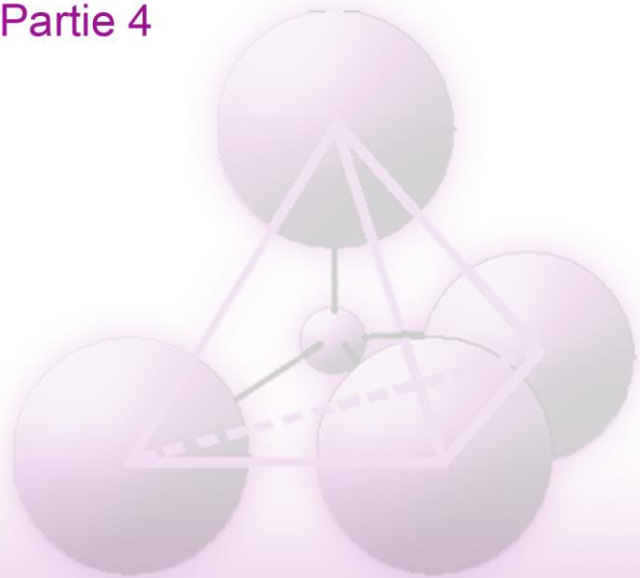
αMETIST

Partie 1 :
Appropriation : besoins, conditions

Partie 2

Partie 3

Partie 4



Appropriation, mutualisation, expérimentations des technologies de l'information scientifique et technique

Sylvie Lainé-Cruzel – laine@univ-lyon3.fr

ERSICOM – Université Jean Moulin Lyon 3

5 Positionnement

Pourquoi une revue sur la question de l'appropriation des Nouvelles Technologies de l'Information et de la Communication (NTIC) par les chercheurs ? Parce que les enjeux sont énormes. L'enjeu réel n'est rien moins que la capacité d'un pays à faire de la science : c'est-à-dire produire de la science, de la recherche et des chercheurs. Or les chercheurs ne sont plus depuis longtemps des savants isolés dans leurs bureaux, consignant à la plume sur leurs cahiers le récit de leurs expérimentations solitaires, envoyant leurs manuscrits à des sociétés savantes et à quelques confrères. La recherche se construit dans un contexte de compétitivité internationale. Les hypothèses et expériences des chercheurs s'expriment dorénavant dans des modes complexes, qu'il faut impérativement maîtriser : une culture technologique est devenue indispensable aux chercheurs, autant que l'est la maîtrise de leur domaine de recherche propre.

5.1 Communiquer la science

Faire de la science, c'est la concrétiser dans des productions destinées à être validées, puis rendues visibles et partagées. Or les formes de ces productions se diversifient, à la fois du point de vue des supports (l'électronique complète le papier, sans l'avoir fait disparaître), du point de vue des médias (les modes de diffusion possibles sont nombreux), du point de vue des formes rédactionnelles et du point de vue des publics concernés. Il existe

donc actuellement de multiples modes de rédaction de la science, comme il existe de multiples modes de diffusion électronique et de multiples diffuseurs individuels, institutionnels ou commerciaux. Les modes de validation, la visibilité et l'exploitation des documents électroniques sont régis par des systèmes qui offrent des solutions extrêmement variées, entre lesquelles le choix devient souvent complexe, à la fois pour les producteurs et pour les utilisateurs. Les utilisations et réagencements seront parfois sous le contrôle direct des futurs lecteurs et exploitants, lesquels souvent sont eux-mêmes chercheurs et donc producteurs d'information scientifique.

Et faire de la recherche implique de connaître l'état de la recherche, de ses productions et ses acteurs. Pour identifier les travaux, les projets, les thématiques, les concurrents ou les partenaires potentiels dans un domaine de recherche, il est nécessaire de savoir où et comment les rechercher et comment obtenir les informations utiles, dans les formes les plus adaptées aux besoins et aux usages auxquels on les destine.

5.2 Evolution des outils, évolution des usages

Pendant plusieurs décennies, les formes de production scientifique liées au papier ont conservé des formes stables. On a vu coexister la production académique (comme celle des thèses : validation institutionnelle, tirages limités et diffusion restreinte), la production éditoriale (dans des revues ou des collections, avec une plus ou moins stricte validation par les pairs et un indice d'audience assez facilement quantifiable, ne serait-ce que par le tirage) et une production plus souterraine, constituée de rapports de projets, d'actes de colloques, de publications de sociétés savantes, etc.

La mise en visibilité de ces différentes productions a été améliorée par l'élaboration de systèmes d'information répertoriant les références d'une partie de ces publications et permettant de connaître leur existence.

Des attentes majeures ont été alors exprimées vis-à-vis de ces systèmes, tandis que de nombreux facteurs limitaient encore leur évolution et en particulier leur capacité de stockage et le nombre restreint de points d'accès à l'information.

Mais en même temps que ces verrous disparaissaient, la créativité des concepteurs de ces systèmes commençait à bouillonner. Aujourd'hui, de nombreuses possibilités techniques sont rendues disponibles avant même que les usages ne les réclament. De multiples choix sont possibles, avant que la réflexion n'ait mûri pour guider leur évolution et surtout leur prise en mains.

Il est maintenant nécessaire d'inventer cette exploitation et de construire des usages intelligents. La technique ne peut plus prendre seule le contrôle des évolutions.

L'abondance technologique peut d'ailleurs contribuer au vertige et à la sensation d'éparpillement. Il ne suffit pas de connaître la technique pour se l'approprier, c'est-à-dire l'utiliser efficacement. Il ne suffit pas de connaître les flux RSS, les podcasts et autres folksonomies pour surveiller l'évolution d'un champ scientifique, comme il ne suffit pas de participer à un blog collaboratif ou de tenir à jour son site pour avoir une activité de production scientifique légitimée.

5.3 Les frontières s'estompent

En cette période riche en évolutions des outils et des pratiques, où tous les positionnements sont à reconstruire, les anciennes frontières sont remises en question. Les territoires se chevauchent et les définitions s'entremêlent. D'anciennes distinctions bien établies méritent d'être réétudiées.

- Distinction entre **auteurs et médiateurs** : il est logique que la répartition des rôles entre les producteurs de l'information soit remise en cause, dans la mesure où elle a été longtemps associée à la distinction entre fond et forme dans les productions documentaires. Les médiateurs ont longtemps « mis en forme » les textes des chercheurs : ce fut une activité importante des éditeurs traditionnels, avant l'ère du document numérique. Dorénavant les auteurs sont souvent amenés à composer eux-mêmes leurs documents, parfois même à les structurer, à les décrire, etc. Lorsque des professionnels de l'information interviennent, c'est pour apporter une plus-value au texte initial : délinéarisation, recomposition, associations, enrichissements. Et dans un tel travail de « redocumentarisation », pour utiliser un terme cher à JM Salaün, la distinction entre fond et

forme ne semble plus si pertinente. La réflexion, dont les bases ont été très clairement posées par le RTP-[1]doc, doit se poursuivre et aider à cerner les compétences et les nouveaux métiers de la production scientifique. Car il semble évident que le travail des médiateurs s'est déplacé : si par le passé ils intervenaient sur des documents pour en faciliter l'accès à des lecteurs dont ils étaient les interlocuteurs, dorénavant ils travaillent davantage avec les auteurs pour élaborer du « matériau numérique », selon la traduction longuement discutée du terme « electronic stuff » de Carl Lagoze…

- Distinction entre les questions d'informations qui intéresseraient les **« sciences dures ou les sciences expérimentales »** et l'information **« sciences humaines et sociales »** : ces questions sont-elles si fondamentalement différentes ? Probablement non, en tout cas si elles diffèrent, ce n'est pas sur des caractéristiques globalement triviales. Nous avons choisi pour parler de nos activités éditoriales le terme vague d'IST, dont nous ne sommes pas nous-mêmes réellement satisfaits et qui a déjà suscité débat. Pour la vague de fond portée par le passage au numérique de l'information et tout ce qu'elle recompose, ce n'est pas ce clivage qui nous paraît pertinent : toutes les disciplines nous intéressent. Il faudrait sans doute trouver un meilleur terme pour parler d'informations produites par les chercheurs et diversement validées par les communautés scientifiques concernées, qui lui accordent un statut et une légitimité…

- Ne nous attardons pas sur les multiples clarifications à reconstruire, entre documents (plus ou moins structurés, composites ou complexes), ressources, données ou bases de données, matériaux… : le débat est essentiel et il est mené en divers lieux. Mais évoquons aussi tous les glissements et jeux de vocabulaire associant à l'information, de manière plus ou moins rigoureuse et avec des acceptions diverses, les termes de « connaissance » (extraction de connaissance, ou bases de connaissances, ou gestion des connaissances…) ou l'idée de sémantique, omniprésente dans la vision du web du futur et à laquelle nous aimerions bien sûr consacrer un numéro *2.0*…

Et pourtant, s'il y a partout glissement et déplacement des frontières, cela ne signifie pas qu'il y ait nécessairement uniformisa-

[1] http://rtp-doc.enssib.fr/

tion des pratiques. D'une communauté à l'autre, on aura des formes différentes de production de la science (revues, ouvrages, rapports, colloques, brevets, bases de données, etc.), comme on aura des complémentarités différentes entre les différents supports (mise en ligne de prépublications ou de postpublications), ou différentes manières d'envisager les répartitions des fonctions entre les acteurs concernés (ex : création d'archives ouvertes). Ainsi, si les « bonnes pratiques » sont à définir au sein des communautés, rien n'indique à l'heure actuelle que cette définition doit se faire de manière homogène et cohérente entre toutes les disciplines ou tous les chercheurs. Nous sommes tous en phase d'expérimentations.

5.4 Appropriation, Mutualisation, Expérimentations : l'AME de notre projet éditorial

Il nous semble nécessaire que les chercheurs s'approprient leur système de production scientifique. Nous entendons par ***s'approprier*** : connaître et comprendre l'ensemble de ce système dans toute sa complexité, pour pouvoir l'utiliser au mieux ; et, dans ce but, en assurer le contrôle.

Il s'agit donc essentiellement d'acquérir la maîtrise d'un dispositif actuellement complexe, hétérogène et fragmenté. Maîtriser, c'est-à-dire, dans une situation donnée ou par rapport à un objectif donné, utiliser plus efficacement, plus intelligemment, plus facilement, plus économiquement ce dispositif. Il faut pour cela adopter les bonnes pratiques et adapter l'outil à l'usage.

Pour adopter les « bonnes pratiques », il faut d'abord les identifier : c'est-à-dire permettre la capitalisation et la transmission des expériences et des savoir-faire acquis.

Par ***mutualiser***, nous entendons la prise en compte de la dimension collective. Il s'agit à la fois de partager les outils et les savoir-faire : nous portons donc un intérêt particulier aux plates-formes partageables, à la normalisation ou aux logiciels libres. Nous offrirons une tribune privilégiée à tous ceux qui se donnent, à travers la prise en mains de leurs outils, un objectif de meilleure concertation dans l'action (coordination, partenariats, travail collaboratif). Enfin, nous souhaitons mettre en lumière les initiatives qui ont pour but de nous aider à mieux nous connaître et nous

identifier (nous, acteurs de l'IST) et de nous rendre visibles collectivement : archives ouvertes, référencement, élaboration d'indicateurs…

Enfin, par ***expérimenter***, nous souhaitons à la fois manifester :

– notre volonté de mettre en valeur les témoignages concernant des expériences pilotes, novatrices, originales, ou concernant des communautés spécifiques,
– notre souhait d'apporter, ne serait-ce qu'à titre expérimental, notre soutien à l'expérimentation en divers lieux et à favoriser l'incubation,
– notre propre volonté d'expérimentation dans la conception de formes rédactionnelles, rendues possibles par l'environnement numérique (voir partie 4).

6 AMETIST : une expérience éditoriale

AMETIST a pour vocation d'être une revue **scientifique**. Elle souhaite accueillir dans ses pages des contributions originales et inédites, proposées par des chercheurs et spécialistes de l'IST, mais aussi par des chercheurs d'autres communautés scientifiques qui portent un intérêt particulier à la question de la communication scientifique et qui trouveront là une tribune où témoigner de leurs pratiques, de leurs dispositifs ou de leur organisation et de leurs réflexions.

Tous les textes envoyés à la revue seront soumis à notre **comité de rédaction**.

Lorsque le comité de rédaction ne disposera pas de suffisamment de propositions de communications originales soumises par leurs auteurs, il effectuera une veille dans un certain nombre de manifestations scientifiques, pour identifier des travaux intéressants sur la question de l'appropriation, qui auraient pu être présentés en divers lieux (colloques, séminaires, éventuellement dans des revues non francophones) mais qui n'auraient pas encore fait l'objet d'une publication dans une revue francophone. Il pourra alors se mettre en rapport avec les auteurs pour leur proposer d'en construire une

rédaction adaptée aux attentes de la revue et éventuellement les y aider.

La revue AMETIST sera intégralement **francophone**. Aucune culture, aucune langue ne doit être dépossédée du savoir scientifique ni de la maîtrise technologique : le monde francophone doit défendre son identité et sa richesse culturelle et s'approprier tout ce qui favorisera le développement de sa production scientifique. L'uniformisation de la langue scientifique ne peut conduire qu'à un appauvrissement culturel. Or la science a besoin d'humanisme, de culture et d'une réflexion sur ses productions et ses enjeux : elle aurait beaucoup à perdre si elle n'existait plus que dans une seule langue. S'il en était besoin, le travail que nous avons réalisé sur ce numéro nous en aurait encore davantage convaincus. Mais nous reparlerons de l'intérêt d'une réflexion collective sur la terminologie et la traduction.

Enfin, la revue AMETIST que vous avez en mains – ou que vous consultez sur votre écran – est en elle-même une expérimentation. Nous souhaitons en effet la faire exister sous une double forme : celle d'une revue scientifique « traditionnelle », c'est-à-dire sur **papier** et celle d'une **revue électronique**, qui ne sera pas simplement l'image de la revue papier.

Le contenu sera globalement similaire dans les deux formes, mais la version électronique sera parfois plus riche que la revue papier. Dans chaque numéro, nous travaillerons particulièrement certains articles, pour en permettre d'autres types de lecture qu'un parcours linéaire et pour enrichir le contenu avec des notes, des liens, des commentaires, des renvois, des extraits de discussion … que nous vous laisserons découvrir.

L'équipe technique, basée à l'INIST, qui réalise la revue effectue ici un travail important et original, avec une conception qui sera différente pour chaque article. Nous sommes dans l'expérimentation, nous souhaitons inventer !

La revue papier, comme la revue électronique, paraîtra à un rythme **semestriel**. (Septembre et Mars de chaque année).

7 Quelques thématiques

3.1 Production

Les NTIC sont utilisées (ou peuvent l'être) à tous les niveaux de la production de la science et pour toutes ses formes d'écriture et de publication électronique.

Toutes les communautés les emploient pour élaborer leurs connaissances et les mettre en forme, mais souvent de manière très différenciée d'une communauté à une autre. Les astronomes matérialisent le résultat de leurs observations en liant des bases de données (telles que des catalogues d'objets stellaires) avec des données bibliographiques ou des documents. De même, les généticiens relient les séquences d'ADN localisées et exprimées factuellement, avec les productions discursives qui les commentent. Les archéologues photographient leurs chantiers de fouille (qu'ils doivent détruire au fur et à mesure de leur progression pour accéder aux couches les plus profondes) et les artefacts qu'ils y découvrent, mais la valeur de ces photographies réside essentiellement dans le discours signifiant qui leur est associé. Les mathématiciens utilisent des outils spécialisés dans la gestion et la mise en forme du discours mathématique, donnant accès à des fonctions de calcul et d'aide aux démonstrations. Les chercheurs en lettres relient les œuvres sur lesquelles ils travaillent aux commentaires critiques qui les accompagnent et pour les historiens, les textes médiévistes deviennent accessibles sur tous les écrans, associés à leurs traductions et à des glossaires.

Mais ces communautés ont du mal à stabiliser leurs pratiques et la forme de leurs productions numériques, confrontées à la diversité des outils, des standards et de la manière de les utiliser. Les avancées techniques se multiplient, progressant sans toujours converger.

3.2 Accessibilité et visibilité

La question de l'accessibilité renvoie dans un premier temps à la question du dépôt : où déposer ? dans quel(s) format(s), avec quelles métadonnées, sous quelle affiliation ? quand déposer ? et à quel stade relativement à une éventuelle publication papier ?

Comment gérer les différentes versions d'un document ? : doivent-elles coexister ? comment articuler une logique de conservation qui veille à une garantie de la pérennité dans une vision patrimoniale et une logique d'usage qui garantit l'accès à des documents actualisés ? Ces questions classiques relevant traditionnellement de la bibliothéconomie doivent être repensées entièrement, car les facteurs qui les régulaient « naturellement » changent de nature dans le passage au numérique (du côté de la production : contraintes éditoriales, tirages…et du côté de la mise à disposition : utilisation d'un espace physique limité en mètres linéaires de rayonnage – ouvrages ou tirés à part se détériorant au fil du temps – budget d'acquisition limité, etc.).

Mais surtout, ces questions ne sont plus sous l'unique responsabilité des médiateurs, documentalistes ou bibliothécaires professionnels. De même que le médiateur intervient de plus en plus très en amont dans la conception et la structuration des documents, en élaborant des cadres rédactionnels facilitant le travail des auteurs, l'auteur quant à lui est amené à faire des choix devant lesquels il peut se retrouver d'autant plus désarmé que les outils qu'on lui met à disposition l'autorisent à (presque) tout faire. C'est le cas de la plate-forme HAL, principale plate-forme de dépôt de la communauté scientifique française, comme c'est le cas pour bien d'autres outils.

Elle renvoie ensuite à la question du recensement et du référencement : comment améliorer le recensement dans les réservoirs bibliographiques, quels réservoirs ou flux faut-il alimenter ? Avec quels langages de référence ou en suivant quelles recommandations les auteurs doivent-ils décrire leurs documents électroniques ou leurs affiliations et comment ajuster les métadonnées ? Comment une équipe, une institution ou un groupe de projet peut-il rendre, mettre en lumière, sa production scientifique visible, de manière à s'assurer une meilleure visibilité et un positionnement international ? Là encore, les choix stratégiques sont de plus en plus souvent sous la responsabilité au moins partielle de leurs auteurs et une très bonne coordination et connaissance mutuelle des acteurs de la production (auteurs et médiateurs) est nécessaire. Cette connaissance mutuelle (en particulier en termes de répartition des compétences et des responsabilités) est partout en cours de reconstruction.

3.3 Valeur ajoutée

L'évolution des outils permet d'envisager de nouvelles perspectives d'exploitation de la production scientifique, telles que l'élaboration de connaissances à partir de documents, l'intégration d'enrichissements collaboratifs (commentaires, compléments, appréciations…), la personnalisation des outils d'accès à l'information, etc. Comment adapter ces nouvelles fonctionnalités aux besoins des chercheurs et des structures ? Comment les partager dans un cadre communautaire ?

3.4 Exploitation riche – aide à la décision

Comment mener une activité de veille et comment mutualiser cette activité ? Quels indicateurs peut-on élaborer pour évaluer l'audience ou la productivité, ou repérer les thématiques les plus porteuses et les secteurs les plus dynamiques ? Quelles interprétations peuvent être associées aux différentes mesures, comment identifier les biais ou les silences inhérents à la construction de ces indicateurs ? Comment les exploiter pour favoriser les collaborations émergentes, renforcer les secteurs en difficulté ou soutenir les plus actifs ? Toute question liée à l'interprétation ne peut être menée qu'avec une excellente connaissance des domaines concernés, mais également des outils qui collectent les informations exploitées. Là encore, de nouvelles coordinations sont à trouver entre gestionnaires décisionnaires, experts du domaine, et spécialistes de l'information.

8 Contenu

La revue sera structurée en 4 parties, dont chacune a un objectif différent :

Partie 1 : l'appropriation : pour quels besoins, à quelles conditions ?
Les articles qui seront publiés sous cette rubrique ont une dimension plus globale et théorique. Ils analysent les attentes des chercheurs, de manière globale ou dans des communautés ciblées. Ils dressent le bilan d'un certain nombre d'actions et proposent des réflexions sur les conditions qui permettraient une meilleure appropriation des

outils : conditions techniques, sociales, culturelles, organisationnelles, langagières, etc., en tentant d'identifier les facteurs de blocage ou d'incohérence (verrous) ; enfin, ils brossent les contours des nouvelles pratiques et des nouveaux métiers.

Partie 2 : capitalisation et mutualisation

Partager des bilans à portée généralisable, témoigner d'expériences concluantes avec un certain recul, exprimer des recommandations dans la prise en main et l'exploitation d'outils, accompagner les outils existants de prescriptions ou de recommandations… De nombreux outils, plates-formes ou langages de représentation sont souples et évolutifs ; mais le fait qu'ils permettent de tout faire ne guide guère leur prise en main et leur usage. Quantité d'indices ou de mesures peuvent être mis en place, mais de quoi témoignent-ils réellement et quelle signification leur associer ? Ce sont les communautés concernées qui doivent accompagner ces outils par la proposition de recommandations sur la manière de les utiliser et élaborer procédures et spécifications. Les articles publiés dans cette deuxième partie présenteront donc des bilans et des témoignages relatifs à des expériences significatives, susceptibles d'être intéressantes pour d'autres communautés que celles au sein desquelles elles auront été menées.

Partie 3 : Coups de flash

Expérimentations ponctuelles ou originales : l'innovation est testée en divers lieux. Il ne faut pas toujours attendre qu'elle ait pris une ampleur et une portée significative pour la porter en lumière. Par ailleurs, les difficultés ou limites d'une expérimentation ont une valeur de témoignage, autant que les réussites, et méritent parfois tout autant d'être partagées. Dans cette partie seront publiés des textes courts. Les contributions des jeunes chercheurs y seront volontiers accueillies.

Partie 4 : ARTIST, un lieu d'expérimentations

AMETIST est une revue portée par le collectif ARTIST (Appropriation par la Recherche des Technologies de l'Information Scientifique et Technique). ARTIST, créé en 2005, rassemble un réseau de praticiens de l'IST et de chercheurs en sciences de l'information ou en NTIC et est soutenu par l'INIST, qui lui a affecté une équipe de cinq personnes. L'activité actuelle d'ARTIST se

concentre autour de différents projets : outre la revue AMETIST, sont développés :

- le forum ARTIST : espace de travail pour chercheurs et praticiens, il s'agit d'un site web interactif complété par des listes de diffusion. Sa fréquentation actuelle est d'environ 1000 visites par mois.
- le banc d'essai ARTIST/AMETIST, qui développe des expérimentations collectives : rédaction collective dans un forum public destinée à produire des publications ou communications, traductions avec forums terminologiques, échanges sur divers aspects techniques liés à l'édition scientifique : métadonnées, formats, vocabulaire, ontologies…
- un ensemble d'initiatives destinées à favoriser l'insertion internationale, l'ouverture vers d'autres communautés scientifiques, l'élaboration d'observatoires ou de portails.

Nous vous invitons à en découvrir les évolutions sur le site Artist : http://artist.inist.fr

Dans cette partie 4, nous tiendrons une chronique des projets les plus significatifs du groupe ARTIST et des bilans que nous en tirons.

9 Un numéro zéro, prototype…

Enfin, quelques mots sur ce numéro zéro. Nous sommes fiers d'y présenter des réflexions et des travaux intéressants et pertinents, émanant de chercheurs confirmés. Nous ne sommes pas tout à fait aussi sûrs de nos compétences éditoriales, compte-tenu du pari difficile que nous avons souhaité relever, avec l'objectif d'une réalisation parallèle de deux formes différentes liées à des supports différents et dans des conditions un brin acrobatiques (voir partie 4).

Ce numéro est donc sans doute entaché de multiples coquilles, imperfections et maladresses pour lesquelles nous demandons l'indulgence de nos lecteurs et le pardon de nos auteurs. Nous tenons à remercier l'INIST qui a donné à cette revue la possibilité d'exister en lui affectant une équipe et qui a par ailleurs intégralement financé la réalisation de ce premier numéro. Nous vous donnons rendez-vous

avec le numéro 1 dans quelques mois, numéro pour lequel nous espérons déjà recevoir de multiples propositions de contributions. Pour connaître l'évolution de nos projets éditoriaux et tous les détails pratiques liés à la soumission d'articles ou l'accès à notre revue, rendez-vous bien sûr sur le site http://artist.inist.fr. Et bonne lecture !

Sylvie Lainé-Cruzel, Présidente du Comité de rédaction d'AMETIST

Pour le comité de rédaction

Qu'est-ce qu'une bibliothèque numérique, au juste ?
Au-delà des fonctions recherche et accès dans la National Science Digital Library

Carl Lagoze (1)
lagoze@cs.cornell.edu

Sandy Payette (1)
payette@cs.cornell.edu

Dean B. Krafft (1)
dean@cs.cornell.edu

Susan Jesuroga (2)
jesuroga@ucar.edu

(1) Computing and Information Science, Cornell University, Ithaca, NY
(2) UCAR-NSDL, Boulder, CO

Cet article est la traduction par Frédéric MARTIN (BnF) d'un article publié dans le numéro de novembre de la revue D-Lib Magazine. Il a été révisé par Catherine GUNET (INIST-CNRS), mis en ligne par l'équipe ARTIST et a bénéficié du soutien du RTP-DOC.

Il est paru sous la référence originale :
What Is a Digital Library anyway, anymore ?
Beyond Search and Access in the NSDL

Les bibliothèques numériques, bien que de tailles différentes[1], vivent à présent leur adolescence. Comme pour toute adolescence, il y a de quoi s'enthousiasmer et se préoccuper.

Les succès rencontrés pendant une décennie de recherche, de développement, de déploiement sont source d'enthousiasme. Toute liste en serait nécessairement incomplète, mais inclurait sans nul doute Google[2], le Handle System®[3], le Dublin Core[4], l'OAI-PMH[5]

[1] Bien que les catalogues de ressources numériques aient été introduits de bonne heure dans l'histoire de l'informatique [51] l'emploi généralisé du terme « bibliothèque numérique » remonte au début des années 90 [16] [20].
[2] <http://www.google.com>
[3] <http://www.handle.net>
[4] <http://www.dublincore.org>

(protocole pour la collecte des métadonnées créé par l'Initiative des Archives Ouvertes), l'Open URL [40], arXiv[6], DSpace [53] et LOCKSS [50]. Ces réalisations, parmi d'autres, sont à mettre en relation avec l'explosion généralisée du web lui-même, durant les quinze dernières années[7]. Elles tendent vers cette vision de la bibliothèque numérique comme « accès universel au savoir humain » exprimée dans le rapport du President information Technology Advisory Committee en 2001 (PAC)[45].

Les préoccupations proviennent en partie de problèmes soulevés lors des premiers ateliers de bibliothèque numérique [8, 34] et pour lesquels des solutions pratiques restent à mettre en oeuvre. Quelques exemples le montrent. Tandis que Handle et DOI[8] ont été déployés avec succès dans les communautés des bibliothèques et de l'édition, l'objectif visant à disposer d'identifiants universels et pérennes n'est toujours pas atteint. L'adoption largement répandue du Dublin Core et de l'OAI-PMH semble répondre aux objectifs initiaux d'une description des ressources qui soit interopérable. Pourtant, des problèmes liés à la qualité des métadonnées [58] compromettent l'utilité des standards. Les logiciels médiateurs d'identité fédérée comme Shibboleth[9] commencent à répondre aux questions d'autorisation et d'identification, mais l'infrastructure à clé publique, considérée comme « essentielle à l'émergence des bibliothèques numériques » [34], n'est pas encore développée. En dépit des efforts produits par l'initiative pour le Web sémantique du W3C [13], le Saint Graal de l'interopérabilité sémantique [42] reste hors de portée. Enfin, face aux volumes croissants d'information sous forme numérique native et stockée dans des entrepôts institutionnels, il nous manque toujours des techniques extensibles et normalisées pour préserver pleinement cette information.

Ces inconvénients techniques se situent dans un contexte institutionnel plus large et plus inquiétant, que certains[10] ont

---

[5] <http://openarchives.org>
[6]<http://arxiv.org>
[7]Dont la richesse sera significativement accrue par des efforts de numérisation massifs tels que Google Print <http://print.google.com>
[8]<http://www.doi.org>
[9]<http://shibboleth.internet2.edu/>
[10] Une recherche a montré qu'il existe plus de 13 000 occurrences de ce terme sur le web, dont une webémission de Clifford Lynch et Michael Keller sur le sujet [35].

caractérisé comme la « google-isation » des bibliothèques numériques et de l'information en général. Comme tout néologisme, « google-isation » a plusieurs sens. Ici, il réfère à l'idée fausse et agaçante selon laquelle Google représente l'apothéose de l'information numérique et que les problèmes restant dans ce domaine ont déjà été résolus - ou vont l'être (peut-être même par Yahoo !, MSN, etc.). Suite à des discussions informelles avec des collègues de la communauté des bibliothèques numériques de recherche, il ressort que la « google-isation » a contaminé les organismes de financement, à la fois publics et privés. Si l'absence de financement important pour un programme de bibliothèque numérique au sein de la National Science Foundation est imputable à de nombreuses causes, l'idée que « Google a résolu le problème » y est certainement pour quelque chose.

Les réalisations de Google sont certes frappantes. Mais cette vision réductrice d'une « Fin de l'Histoire »[11]apparaît comme le résultat d'une confusion sur « ce qu'est une bibliothèque (numérique) ». Peut-être sous l'influence de visions utopiques et trompeuses, comme les commentaires d'Al Gore sur « l'écolier de Carthage, Tenessee » [9], il existe semble-t-il une croyance selon laquelle une bibliothèque numérique ne concerne que la recherche d'information (« est-ce que je peux la trouver ? ») et l'accès (« est-ce que je peux l'obtenir ? »). Certes, ces fonctions sont essentielles (et demeurent des défis), mais elles ne sont que la partie d'un environnement informationnel. Les bibliothèques traditionnelles sont bien plus que des entrepôts bien organisés de livres, de cartes, de périodiques, etc. Elles sont par nature des lieux où des personnes se rencontrent pour accéder à un savoir qu'ils partagent et qu'ils échangent. Les ressources que les bibliothèques sélectionnent et les services qu'elles offrent devraient refléter l'identité des communautés qu'elles servent [31].

Comme le suggère Borgman [14-16], les bibliothèques numériques devraient non seulement ressembler aux bibliothèques traditionnelles mais encore aller beaucoup plus loin qu'elles. Elles ne doivent pas se limiter à de simples moteurs de recherche. Comme toutes les

[11]Il s'agit d'une référence à un livre de Francis Fukuyama publié en 1992 [21] remarquant une semblable myopie euphorique dans le domaine de l'économie politique.

bibliothèques, il faut qu'elles procèdent à un haut niveau de sélection des ressources qui répondent aux critères de leur mission. Il est également nécessaire qu'elles fournissent des services, comme la recherche, qui facilitent l'utilisation des ressources par la communauté ciblée. Mais, libérées des contraintes physiques d'espace et de support, les bibliothèques numériques peuvent mieux s'adapter aux communautés qu'elles servent et mieux les refléter. Elles doivent être collaboratives, en permettant aux utilisateurs de contribuer et d'apporter du savoir, soit de façon active par des annotations, des comptes rendus de lecture, etc., soit de façon passive au travers de leurs profils d'utilisateurs. En outre, il faudrait qu'elles soient contextuelles, illustrant ainsi le réseau extensible des relations et des couches de savoir qui se tisse autour des ressources sélectionnées. De la sorte, le noyau de la bibliothèque numérique devrait être une base d'information évolutive, en entrelaçant ainsi dans un même tissu « sélection professionnelle » et « sagesse des peuples » [54].

Cette vision élargie du rôle des bibliothèques numériques implique de repenser les modèles informationnels sur lesquels elles reposent. Le poids de l'héritage légué par le catalogue collectif dans les bibliothèques traditionnelles, ajouté parfois à la disproportion prise par la fonction recherche, a conduit à l'utilisation répandue d'un modèle informationnel bâti sur un entrepôt de métadonnées. Même si souvent les bibliothèques numériques sont implémentées différemment, on constate généralement qu'à la base elles compilent, indexent et fournissent des requêtes sur un catalogue composé de notices de métadonnées. Comme nous le montrerons plus tard, ce modèle de catalogue simpliste est nettement insuffisant dans le cadre d'une vision plus étendue de ce qu'est une bibliothèque numérique.

Le présent article décrit un modèle informationnel pour les bibliothèques numériques qui va délibérément "au-delà de la recherche et de l'accès", sans pour autant ignorer ces fonctions de base et qui facilite la création d'environnements de savoir collaboratifs et contextuels. Ce modèle est un réseau d'information superposé (information network overlay) qui représente la bibliothèque numérique sous la forme d'un graphe. Ce graphe comporte des nœuds typés, qui correspondent aux unités d'information (documents, données, services, acteurs) au sein de la

bibliothèque et des arêtes représentant les relations contextuelles qui se nouent entre ces unités. Ce modèle informationnel incorpore de l'information locale et distribuée intégrée aux web services, autorisant la création de documents enrichis (par ex., des objets d'apprentissage, des publications pour l'e-science, etc.). Il exprime les relations complexes entre les objets d'information, les acteurs, les services et la méta-information (comme les ontologies) et représente ainsi les ressources dans leur contexte, plutôt que comme le résultat d'un accès web isolé. Il facilite les pratiques collaboratives, fermant ainsi la boucle entre les utilisateurs-consommateurs et les utilisateurs-contributeurs.

Nous nous proposons de décrire comment ce modèle de données est implémenté dans Fedora, [27] logiciel libre de gestion d'entrepôt. Fedora est particulièrement adapté à cette tâche, grâce à sa manière unique de combiner un modèle d'objet flexible, l'intégration des web services, une gestion de l'accès permettant une fine granularité et l'incorporation de l'expressivité du web sémantique.

Ces travaux se situent dans le cadre du projet de la National Science Digital Library (NDSL) [61]. Les conditions posées par la NDSL, en termes d'échelle et de contraintes, nécessitent une approche aussi poussée. Franck Wattenberg en a décrit la vision originelle comme suit :

« A bien des égards, la NDSL pourrait aller bien plus loin que l'image traditionnelle de la bibliothèque. En plus de fournir un accès large et actualisé à des ressources à jour et de grande qualité destinées à la formation scientifique, la NDSL pourrait profiter de la connectivité apportée par internet et du potentiel des technologies interactives pour créer un lieu de travail riche et asynchrone : une salle de séminaire, une salle de lecture et un laboratoire où partager et construire la connaissance. Ainsi, la NDSL pourrait fournir un cadre qui, au travers d'un ensemble de ressources diversifié et puissant, faciliterait le travail des utilisateurs dans des environnements différents » [59].

Nous pensons que cette vision élargie de la bibliothèque numérique n'est pas propre à la NSDL. Bien que les communautés qui recherchent et partagent de l'information aient besoin de trouver des aiguilles dans des bottes de foin [28] - un créneau occupé par Google

et ses concurrents - elles ont aussi besoin de fonctionnalité « au-delà des fonctions recherche et accès » , où les bibliothèques numériques « créent un lieu de travail riche et asynchrone » dans lequel l'information est partagée, agrégée, manipulée et affinée.

1 Construire une bibliothèque numérique avec un entrepôt de métadonnées : phase I de la NSDL

Les lecteurs de D-Lib Magazine et la communauté des bibliothèques numériques connaissent sans doute le projet de la NSDL. Aussi, cette section ne présente-t-elle que brièvement le contexte dans lequel s'inscrit le travail que nous décrirons dans le reste de l'article. Nous suggérons donc à ceux qui souhaitent davantage d'information de lire les articles déjà rédigés à ce sujet [6, 7, 25, 61] et de consulter la page « about NSDL » à l'adresse <http://nsdl.org/about>.

L'idée d'une NSDL est née en 1998 au cours d'un atelier [3] financé par la National Science Foundation (NSF). Cet atelier devait étudier les problèmes concernant l'état de l'enseignement en science, technologie, ingénierie et mathématiques (STEM en anglais) aux Etats-Unis et a mis en lumière les opportunités offertes par Internet et les technologies du web pour l'améliorer. En s'appuyant sur les résultats de cet atelier, la NSF commença en 2000 à subventionner des projets NSDL et à ce jour, elle a accordé plus de 180 bourses. Ces aides couvrent un grand nombre de domaines comme le développement des collections, les services et la recherche fondamentale. Les travaux décrits dans le présent article ont bénéficié de subventions de la part des universités Cornell et Columbia ainsi que l'université Corporation for Atmospheric Research (UCAR), pour la partie « noyau intégré », qui comporte la coordination de l'architecture, de l'organisation et de la stratégie pour la NSDL.

Les premiers travaux techniques de l'équipe « noyau intégré » (NI) ont abouti à une architecture et à un modèle de données ayant ces trois fonctions de base : sélectionner des ressources web en STEM, les interroger transversalement et en faciliter l'accès. Le paradigme

architectural pour réaliser ces trois fonctions est essentiellement le catalogue collectif et un entrepôt de métadonnées [EM] en Dublin Core. Ce dernier correspond aux ressources développées et gérées par les projets de collections de la NSDL et par d'autres organismes participants. L'EM est implémenté sous la forme d'une base de données relationnelle Oracle™, dans laquelle les notices de métadonnées individuelles sont stockées dans des séries de tables.

Les notices de métadonnées en Dublin Core, qui contiennent des URL pointant vers les ressources numériques correspondantes, sont absorbées dans l'EM via l'OAI-PMH [29]. Au cours de ce processus d'alimentation, les dates et différents éléments de vocabulaire contrôlé dans ces notices sont normalisées. D'autres services, comme la recherche gérée par le "noyau intégré" et l'archivage des ressources, utilisent un serveur OAI-PMH[12] pour collecter ces notices normalisées et obtenir ainsi l'information nécessaire (par ex., pour construire des index de recherche à partir des métadonnées).

La fonction « recherche » utilise le système d'indexation en texte intégral Lucene[13]pour indexer à la fois les métadonnées collectées décrivant la ressource et le contenu textuel de la première page HTML ainsi référencée. La fonction « archivage » utilise le Storage Resource Broker [10] développé par le San Diego Supercomputing Center. Elle parcourt chaque mois le web à la recherche de toutes les ressources numériques identifiées dans les notices de métadonnées collectées à partir des EM. La fonction « archivage » identifie une collection de pages reliées entre elles, considérée comme la plus représentative de la ressource elle-même et effectue une capture d'archive de ces pages.

Du point de vue de l'utilisateur, les ressources dans le catalogue de la NSDL et les services sous-jacents sont disponibles par le biais d'un portail central disponible à <http://www.nsdl.org>. Celui-ci sera bientôt complété par des portails spécifiques à des communautés éducatives et soutenus par le programme NSDL Pathways [2].

Le portail central de la NSDL et son architecture fondée sur des entrepôts de métadonnées ont été déployés pour la première fois en

[12] <http://services.nsdl.org:8080/nsdloai/OAI>
[13] <http://jakarta.apache.org/lucene/docs/index.html>

décembre 2003. En deux ans, la collection s'est enrichie jusqu'à atteindre plus de 1,1 million de ressources, avec des notices de métadonnées collectées à partir de plus de 80 fournisseurs de données OAI-PMH.

2 Utilité de l'entrepôt de métadonnées en tant qu'architecture de bibliothèque numérique

En règle générale, l'usage à grande échelle du Dublin Core et de l'OAI-PMH dans l'EM de la NSDL a prouvé son utilité pour fournir les services de base d'une bibliothèque numérique, mais il a également révélé de nombreux problèmes d'implémentation. Le plus sérieux concerne la qualité des métadonnées [6] et la validité au regard de l'OAI-PMH, en particulier la conformité au schéma XML. Les coûts administratifs de maintenance de l'EM ont atteint ainsi des niveaux auxquels on ne s'attendait pas. Ces difficultés techniques feront l'objet d'un prochain article.

Cependant, notre sujet ici est d'examiner d'un point de vue plus large l'architecture de la NSDL existante et les bibliothèques numériques en général. Dans cette partie, nous passons en revue des travaux de recherche émanant de la communauté de l'enseignement qui étudient les conditions que doit remplir une bibliothèque numérique axée sur l'enseignement et les fonctionnalités nécessaires pour y répondre.

Les bibliothèques numériques ont une réelle valeur pour le monde de l'enseignement car elles offrent l'accès en ligne à des ressources primaires et des moyens de les utiliser. Mais, pour être vraiment efficaces en tant qu'outils didactiques, elles ne doivent pas se limiter au seul accès à des ressources de qualité. Selon Reeves, « les médias et les technologies ne réaliseront pleinement leur véritable pouvoir d'améliorer l'enseignement que lorsque les étudiants les utiliseront de façon active comme des outils cognitifs au lieu de ne voir en eux que de simples tuteurs ou stocks d'information avec lesquels ils peuvent interagir » [49]. Marshall constate aussi que les bibliothèques numériques doivent être plus que des entrepôts et accompagner la totalité du cycle de vie des données, de

l'information, du savoir et de la construction du savoir en général [36].

Cette fonctionnalité plus large requiert un modèle informationnel pour les bibliothèques numériques qui soit plus riche qu'une collection de simples pages web ou de documents statiques. Wiley [60], entre autres, utilise la notion d'objets d'apprentissage pour indiquer une collection d'informations, qui comprend non seulement une ou plusieurs ressources primaires, mais aussi le contexte pédagogique dynamique de cette information. Ce contexte inclut des informations culturelles et sociales, les profils d'utilisation, les objectifs pédagogiques, la nature des systèmes éducatifs des apprenants, les capacités des apprenants, leurs profils individuels et leurs connaissances antérieures [37]. Le contexte informationnel peut être vraiment complexe, reflétant la diversité des publics desservis et les différences dans la façon qu'ont ces publics d'utiliser et de manipuler l'information.

Certains chercheurs ont examiné les différentes facettes de cette information contextuelle. Elles consistent notamment à :

– recueillir des opinions, des commentaires, des comptes rendus portant sur les ressources de la bibliothèque [39] et l'historique de leur utilisation [43],

– décrire la communauté des utilisateurs impliqués dans la création d'un objet d'apprentissage [48],

– cerner les interactions des utilisateurs et mettre en relation leurs profils avec les objets d'apprentissage [38],

– intégrer les recommandations des enseignants et les corrélations qui existent avec les programmes éducatifs [47],

– repérer et stocker des mots-clés utilisés pour l'interrogation qui conduisent à une utilisation éventuelle de la ressource [4].

Le modèle primaire de données et de métadonnées, orienté « notices », qu'utilisent la plupart des bibliothèques numériques (et traditionnelles), possède une capacité limitée à modéliser pleinement ce contexte informationnel multidimensionnel.

Premièrement, les notices de métadonnées et les entrepôts de métadonnées représentent principalement les propriétés d'un item individuel. Elles ne permettent souvent pas de modéliser complètement les relations contextuelles [43] qui entourent les ressources et n'opèrent aucune distinction entre les multiples entités - ressources, métadonnées, acteurs, ontologies - qui font partie de cette structure relationnelle. De plus, parce qu'elles reposent fréquemment sur des schémas ou des modèles figés, elles sont difficiles à adapter à des besoins en information évolutifs. L'entrepôt de métadonnées de la NSDL, par exemple, ne reconnaît que les collections et les items et ne représente que leur relation d'appartenance. Parce que l'EM est stocké dans une base de données relationnelle, chaque relation nouvelle nécessite une redéfinition du schéma. Ce manque de souplesse s'est avéré problématique à cause de l'évolution des contraintes au cours des activités de la NSDL.

Deuxièmement, la nature statique des notices de métadonnées, qui sont en général créées une fois pour toutes par les créateurs de ressources ou les catalogueurs, pose problème. Le contexte des ressources est dynamique, car il exprime les changements de profils d'utilisation, de préférences individuelles et d'environnement culturel. Selon Recker et Wiley, « un objet d'apprentissage appartient à un réseau complexe de relations sociales et de valeurs touchant l'apprentissage et la pratique. Ainsi la question se pose-telle de savoir si de telles notions contextuelles et mouvantes peuvent être représentées et regroupées dans une notice de métadonnées unique et figée » [48].

Troisièmement, un modèle informationnel centré sur les métadonnées se heurte inévitablement à la distinction floue entre « données et métadonnées »[14] [19]. Par exemple, nous avons remarqué plus haut que les annotations sont une des formes utiles de l'information contextuelle. Les annotations sont-elles des métadonnées (portant sur quelque chose) ou des données à part entière ? Il n'y a pas de réponse univoque, mais une architecture qui marque nettement la distinction entre données et métadonnées rend difficile le traitement de telles ambiguïtés.

[14] Remarquons que c'est cette distinction problématique qui a été l'une des motivations premières [44] pour l'architecture Fedora, utilisée pour implémenter le modèle décrit plus loin.

Enfin, nous avons aussi remarqué l'importance de la réutilisation de l'information - c'est-à-dire la capacité de prendre des ressources primaires et de les combiner dans des objets d'apprentissage agrégés ou des plans de cours [46], puis de les réutiliser pour fabriquer de nouveaux objets. Parce que les unités d'information sous forme physique, dans les bibliothèques traditionnelles, ne peuvent faire l'objet d'un tel réemploi, une approche centrée sur les métadonnées, avec des notices descriptives, était possible. Mais une bibliothèque numérique doit être centrée sur les ressources et fournir un cadre pour gérer, manipuler et traiter le contenu et les métadonnées tout comme la ligne continue qui les sépare.

3 Modélisation informationnelle pour gérer la complexité et le contexte

Quel est donc le modèle informationnel approprié pour dépasser les limites de l'approche "métadonnées" ? En cherchant une réponse à cette question, il faut veiller à ne pas se débarrasser trop vite des notices catalographiques ni ignorer la valeur des métadonnées uniformes, qui mettent de l'ordre dans une information hétérogène [10]. Il est nécessaire pourtant d'incorporer ces notices de catalogue dans une fondation plus riche, de nature à représenter des descriptions structurées et non-structurées, l'hétérogénéité et l'homogénéité, les métadonnées et le contenu, l'information statique et l'information dynamique, les relations complexes et toute une multitude d'autres complexités.

Cette partie décrit le cadre d'un modèle informationnel plus riche qui concilie la complexité et le contexte. Nous développons ce modèle en décrivant « le problème de l'item »[15], avec comme point de départ les fonctions basiques de recherche et d'accès à des éléments homogènes (ressources), auxquelles nous ajoutons progressivement de la complexité. Nous soutenons que, bien que le contexte de cette

[15] Merci, pour cette expression, à un ancien collègue qui travaille maintenant chez Amazon (et qui restera anonyme). La coïncidence entre les problèmes de modélisation de l'information chez Amazon et dans les bibliothèques numériques n'est pas fortuite. Amazon est peut-être le meilleur exemple d'environnement informationnel qui offre aux utilisateurs une information riche et contextuelle construite à partir d'une couche basique de données (ses produits).

description soit la NSDL, le problème envisagé ici peut être généralisé à toute une variété de bibliothèques numériques et d'environnements d'information.

3.1 Représenter des matériaux numériques [16]

Figure 1

Comme évoqué plus haut, le but initial de la NSDL était de proposer des fonctions de sélection, de recherche et d'accès portant sur des ressources en STEM accessibles par URL. Cet objectif limité était atteint par le modèle bien connu du catalogue collectif, où la bibliothèque est représentée comme un ensemble de métadonnées uniformes (Dublin Core) qui référencent les ressources via leurs URL. Il faut noter que dans ce modèle, la représentation des ressources passe au second plan. Elles ne sont pas représentées elles-mêmes mais n'existent qu'indirectement, par le biais de références (URLs) issues des métadonnées.

3.2 Décrire des matériaux numériques de plusieurs manières, structurées ou non-structurées

Figure 2

[16] L'emploi de ce terme est emprunté à Godfrey Rust [11]. "People Make Stuff, People Use Stuff, and People Do Deals About Stuff"

Si le Dublin Core permet une interopérabilité descriptive minimale, il lui faudra coexister avec d'autres formats, plus riches, spécifiques aux disciplines ou aux objectifs visés [24]. Par ailleurs, comme nous l'avons vu précédemment, des descriptions non structurées telles que des commentaires ou des annotations sont souvent aussi utiles que des notices de métadonnées structurées. Sans compter que ces descriptions, structurées ou non, proviennent de multiples contributeurs. Cette complexité supplémentaire met à mal les fondements du modèle de catalogue collectif, qui repose sur un ensemble unique de producteurs (les catalogueurs) qui créent et gèrent un ensemble uniforme de descriptions. Deux nouvelles difficultés de modélisation apparaissent. Tout d'abord, les ressources doivent être modélisées parallèlement aux notices descriptives, puisque les ressources constituent le point d'ancrage pour relier entre elles de multiples descriptions[17] . Ensuite, la modélisation des agents et des producteurs se révèle importante pour représenter le marquage (branding) des ressources (qui a sélectionné ou créé la ressource ?) et le distinguer du marquage des métadonnées (qui a fourni les métadonnées ?). Le marquage est un outil utile pour aider les utilisateurs à connaître la qualité des ressources numériques.

3.3 Ajouter d'autres types de matériaux numériques

Figure 3

[17]Collecter des métadonnées provenant de plusieurs fournisseurs soulève d'intéressants problèmes d'équivalences. La capacité à déterminer que deux descriptions concernent la même ressource se fonde sur l'heuristique et la subjectivité.

Comme nous l'avons vu, le modèle nécessite déjà de représenter différents types de descriptions, les agents qui les produisent et les ressources qu'elles décrivent. Mais les ressources elles-mêmes ne sont pas homogènes. Les bibliothèques numériques collectent une variété grandissante de ressources - images, fichiers audio, textes - et s'ouvrent à des types de ressources bien plus complexes comme les données, les simulations, les objets d'apprentissage multimédia et autres. C'est une source de complexité supplémentaire dans la modélisation - notamment en ce qui concerne la meilleure façon d'associer simultanément l'uniformité au niveau de l'interface utilisateur et la représentation des caractéristiques propres à chaque type de ressource. En plus des problèmes liés à la description (métadonnées), il existe des difficultés concernant l'accès et la présentation, puisque différents types d'informations peuvent requérir différents protocoles d'accès et applications d'aide, qui doivent tous être représentés dans le modèle informationnel.

3.4 Les matériaux numériques sont parfois durs à définir

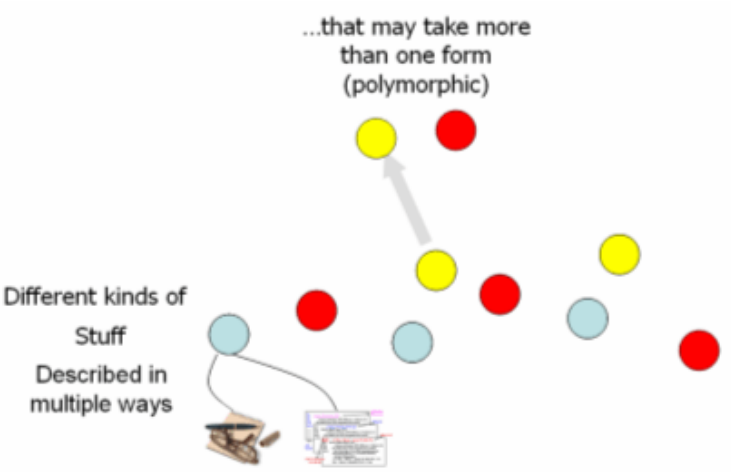


Figure 4

Les ressources dans une bibliothèque numérique ne sont pas toujours simples à caractériser. Par exemple, un livre électronique (e-book) est-il un livre ou un logiciel [33] ? Une information est-elle nécessairement soit une donnée soit une métadonnée ? Un acteur ne peut-il pas être aussi une ressource d'information ? Ce ne sont là que quelques exemples des difficultés que l'on rencontre lorsqu'on modélise l'information. Plutôt que de forcer les matériaux numériques à entrer dans des cases inadaptées, la structure des types du modèle informationnel doit être polymorphe. Toute entité doit pouvoir adopter différentes caractéristiques et différents comportements, en fonction du contexte d'accès ou d'utilisation.

3.5 Permettre aux utilisateurs de personnaliser les matériaux numériques

Figure 5

A l'origine, les bibliothèques numériques ont eu recours à la notion d'objets numériques, qui sont des paquets d'information avec de

multiples diffusions disponibles par le biais de demandes de service [22, 26]. La plupart des systèmes de bibliothèques numériques modernes implémentent cette fonctionnalité en utilisant des standards comme les conteneurs [12, 32] d'objets complexes qui encapsulent les flux de données et de métadonnées associées à un objet numérique. Une demande de service peut alors inclure un paramètre qui spécifie la nature de la diffusion demandée - par exemple, une requête pour une diffusion en PDF ou en LaTex d'un document scientifique.

Dans une architecture orientée "services", ces diffusions peuvent être produites aussi bien sous une forme dynamique que statique. Par exemple, plutôt que de stocker une image en plusieurs formats et résolutions, il est possible de répondre à la requête d'un utilisateur (par ex., 300 dpi, jpeg) en utilisant une seule forme d'archive (TIFF) qui sera traitée par un web service de manipulation d'images. Cette fonctionnalité est particulièrement attrayante dans une bibliothèque numérique à vocation éducative où la personnalisation du contenu, en fonction des divers besoins de l'utilisateur (par ex., la langue) est souhaitable.

C'est pourquoi le modèle informationnel doit modéliser les services parallèlement au texte, aux données, aux images et à toute autre information et doit caractériser les interactions de ces services avec les autres unités d'information dans la bibliothèque.

3.6 Exprimer les relations entre les matériaux numériques

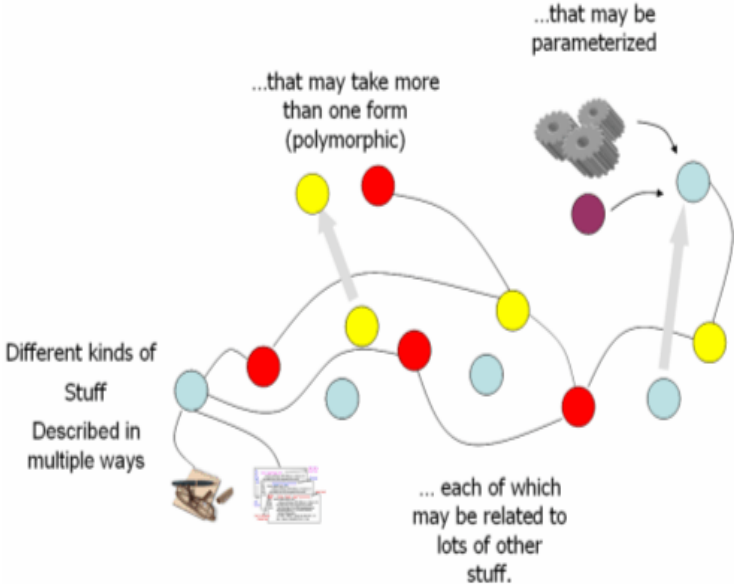


Figure 6

Dans le but de fournir une meilleure prise en compte de « l'objectif de collocation » [55], la communauté des bibliothèques a expérimenté plusieurs modèles informationnels pour modéliser les relations bibliographiques. Les Functional Requirements for Bibliographic Records (FRBR) [1] ou " Spécifications fonctionnelles des notices bibliographiques" en sont un exemple. Notre expérience au sein de la NSDL a montré que les relations bibliographiques ne sont qu'un aspect du problème. Il existe d'autres relations : entre les métadonnées et les producteurs, entre les ressources et leurs gestionnaires, entre les ressources et les taxonomies thématiques, entre les objets d'apprentissage et les programmes d'enseignement, etc. Au fur et à mesure de l'expansion de la bibliothèque, nous anticipons des besoins de modélisation pour d'autres relations

spécifiques à des communautés. Le modèle informationnel doit représenter ce graphe de nœuds d'information interconnectés et les ontologies qui fournissent la méta-information sur ces relations. En outre, ces relations doivent pouvoir évoluer sans être contraintes par des schémas statiques.

4 Le Réseau d'Information Superposé (RIS)

La partie précédente proposait un modèle informationnel sous forme de graphe, avec des arêtes reliées sémantiquement et des nœuds qui soient typés de façon souple et compatible avec les web services. Nous utilisons la notion de réseau d'information superposé (RIS) [1] pour représenter ce modèle.

Le RIS emprunte à deux corpus préexistants. Les réseaux superposés ont été utilisés dans de nombreuses applications pour représenter un ensemble d'arêtes ou de connexions projeté au niveau supérieur d'un ensemble de nœuds qui existe dans d'autres environnements réseaux comme Internet, par exemple. Il existe deux domaines d'application particulièrement connus, le routage réseau [5] et la recherche sur les réseaux P2P [18]. Les graphes sémantiques, qui expriment les relations entre des ressources web, sont consubstantiels au web sémantique [13] et ont été utilisés dans des applications destinées au monde éducatif comme Edutella [41]. En fait, notre application de réseaux d'information superposés recourt aux technologies du web sémantique intégrées dans Fedora.

Les concepts qui sous-tendent le RIS sont illustrés par la figure ci-dessous, avec les couches suivantes :

– Les ressources primaires ou données brutes sélectionnées par la bibliothèque figurent au niveau inférieur. Dans la NSDL, ce sont les ressources en STEM accessibles par le web. Mais, comme nous l'avons vu, ces matériaux bruts consistent également en ensembles de données, d'agents et d'organismes qui contribuent à la bibliothèque et à ses services.
– Le réseau d'information superposé, qui se situe au niveau immédiatement supérieur, est la zone où sont modélisées les ressources de la bibliothèque, leurs descriptions et la toile

d'informations tissée autour d'elles. Il est d'abord alimenté par les ressources primaires, ou les références à ces ressources via les métadonnées, qui sont représentées par des nœuds rouges. L'association et la dérivation de ces nœuds avec le niveau des ressources primaires sont matérialisées par des flèches rouges. Les tirets rouges dans le RIS indiquent les relations initiales entre ces nœuds, telles que les relations entre l'item et la collection dans la NSDL. Ici, l'alimentation du RIS est effectuée par le biais d'une collecte de métadonnées à partir des producteurs de collections, essentiellement en reprenant la fonctionnalité de l'entrepôt de métadonnées (phase I).

– L'API de contrôle d'accès, illustrée à un niveau supérieur, fournit l'accès programmatique total au RIS. Cela inclut l'accès en lecture et en écriture aux composants du modèle de données - documents, données, métadonnées, acteurs, relations, etc. - et la recherche au sein des relations (par ex., "trouver toutes les ressources impliquant une contribution de DLESE"[18]).

– L'API peut alors être utilisée par des contributeurs externes - par ex., des utilisateurs, des services, des fonctions de classification par ontologies, etc. - pour enrichir l'information dans le RIS. Ces requêtes effectuées à travers les API, représentées par des flèches vertes, ajoutent à la fois des nœuds supplémentaires (comme les objets d'apprentissage qui combinent des ressources existantes), qui apparaissent en vert dans la figure et de nouvelles relations entre ces nœuds, notées par des tirets verts.

Ce mouvement bidirectionnel (la représentation des ressources primaires à partir de la couche de données brutes / la représentation de l'information contextuelle à partir de la couche supérieure) permet au RIS d'évoluer à travers le temps vers un espace d'information de plus en plus riche. De la même façon que Amazon.com est une source d'information qui dépasse de loin le simple catalogue de produits, nous espérons que les bibliothèques numériques fondées sur le modèle du RIS reflèteront les communautés de savoir qui se construisent à partir des ressources de la bibliothèque.

[18] <http://dlese.org>.

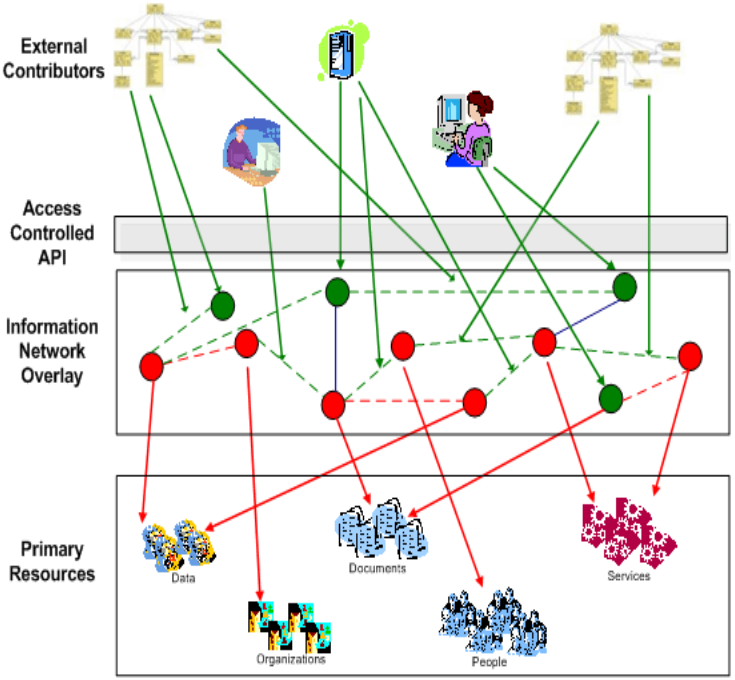


Figure 7

La plate-forme que nous utilisons pour implémenter le RIS est Fedora, un logiciel libre de gestion d'entrepôt[19]. Fedora a été déployé dans une variété d'applications incluant des entrepôts institutionnels, des archives, des musées, des projets de bibliothèques commerciales. Le modèle d'objets riche sur lequel repose Fedora et l'exploitation de ce modèle à travers une interface de service web font de ce logiciel un cadre idéal pour implémenter le RIS.

Chaque nœud dans le RIS correspond à un objet numérique dans Fedora. Le modèle d'objets numériques de Fedora offre des fonctionnalités considérables, combinant la gestion de contenu

[19] <http://www.fedora.info>.

traditionnelle, des architectures orientées service et des technologies du web sémantique. Le modèle permet d'agréger des données locales ou distantes dans de multiples formats. Des services accessibles sur le web peuvent alors être associés aux données agrégées dans un objet numérique. Celui-ci devient alors accessible dans de multiples représentations, certaines étant des transcriptions directes des données agrégrées et d'autres étant produites dynamiquement par les services web associés. Dans le contexte de la NSDL, c'est ainsi que se constitue le socle technique nécessaire au réemploi et à la construction d'objets d'apprentissage complexes [46] ; on mêle les ressources primaires et les commentaires de l'enseignant qui peuvent être présentés dynamiquement dans de multiples formats (par ex., comme des présentations Power Point ou Flash).

Chaque arête dans le RIS correspond à une relation sémantique exprimée à l'intérieur du modèle d'objets numériques de Fedora. On peut citer, comme exemples de relations entre les objets numériques dans le RIS, des relations de gestion bien connues (du type organisation des items dans une collection), des relations de structure (liens de la partie au tout entre des chapitres et un livre), des relations sémantiques utiles dans une organisation de bibliothèque numérique éducative comme la pertinence des sujets, des niveaux d'études, des programmes d'enseignement... Fedora définit une ontologie relationnelle de base en utilisant RDFS [17] et fournit un emplacement dans l'objet numérique pour exprimer des relations fondées sur cette ontologie. Des déclarations provenant d'autres ontologies peuvent aussi être incluses en complément des relations de base de Fedora. Toutes les relations exprimées dans les objets numériques sont converties dans le format Kowari [57], un triplet en RDF natif. L'interface de recherche RDQL [52] et ITQL [56] pour ce triplet est exposée en tant que service web. Comme tout service web, il peut être associé à un objet numérique, autorisant des diffusions à partir d'objets numériques qui sont paramétrés par leur contexte sémantique.

5 L'entrepôt de données de la NSDL : NSDL phase 2

Pour distinguer notre travail de l'entrepôt de métadonnées (EM) de la première phase, nous appellerons notre implémentation du RIS "l'entrepôt de données NSDL" (EDN). La totalité des détails techniques du modèle de données implémenté dans l'EDN dépasse le cadre de cette publication. Les trois exemples de fonctionnalité suivants illustrent quelques unes des caractéristiques du modèle. L'EDN complet consiste en une multitude d'instances des éléments du modèle combinés à d'autres éléments. Par exemple, le modèle-type de métadonnées, décrit dans la partie 5.1 est répété pour chacune des 1,1 million de ressources contenues dans la NSDL.

L'EDN est actuellement implémenté sous la forme d'un entrepôt Fedora unique géré par l'équipe « noyau intégré ». A l'avenir, nous pensons implémenter l'EDN sous la forme d'un ensemble d'entrepôts fédérés.

Chacun des exemples suivants s'accompagne d'une illustration, dans laquelle les cercles représentent des nœuds dans le réseau d'information, implémentés en tant qu'objets numériques Fedora. Chaque cercle a la couleur qui correspond au type d'information qu'il représente dans le contexte de l'exemple de modélisation. Les lignes représentent des relations sémantiquement chargées entre les unités d'information. Comme toutes les autres dans Fedora, ces relations sont stockées à l'intérieur de l'objet numérique et ensuite indexées dans le triplet Kowari. On peut alors rechercher ces relations à travers des requêtes de graphes.

5.1 Des métadonnées multi-sources et multi-formats grâce au marquage

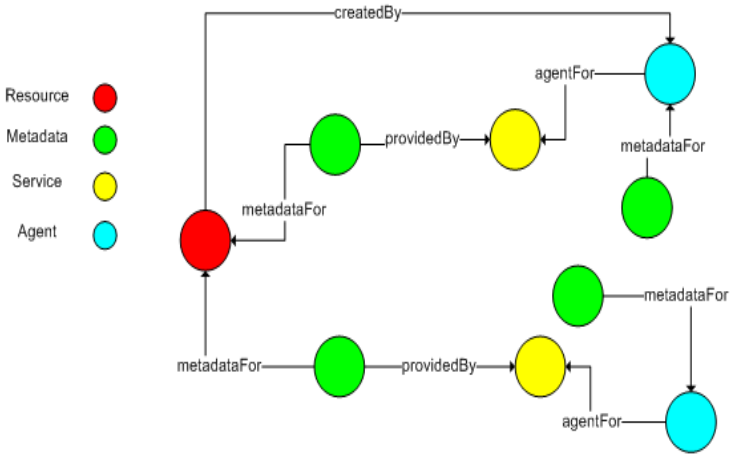


Figure 8

L'illustration ci-dessus montre le modèle EDN qui associe des métadonnées multi-sources et multi-formats à une ressource. Chaque objet numérique de métadonnées agrège plusieurs formats provenant d'un seul fournisseur de métadonnées. Grâce à la capacité de diffusion dynamique de Fedora, certains de ces formats sont générés de façon automatique à partir d'un format de base. Le lien entre l'objet numérique de métadonnées et son fournisseur et celui entre la ressource et son créateur ou sélectionneur fournit une information de marquage. Le marquage est important pour toute bibliothèque dont les données et les métadonnées proviennent de plusieurs sources. Relié à un référentiel de notoriété, le marquage permet de déterminer la qualité des ressources et de leurs descriptions.

5.2 Comptes rendus et annotations non structurés

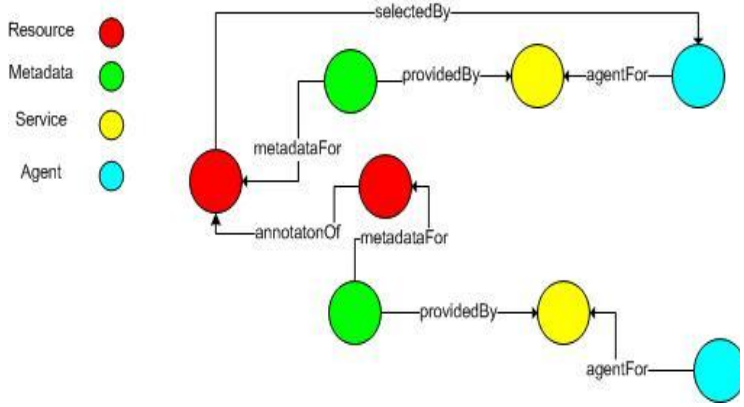


Figure 9

Bien que les métadonnées structurées soient utiles pour répondre à de multiples besoins, les annotations et les comptes rendus non-structurés ont eux aussi leur importance. Le modèle représente ces annotations et comptes rendus comme des ressources à part entière - leur statut d'annotation dépend de leur association à une ressource-cible par le biais d'une relation "annotationFor". C'est un exemple de polymorphisme au sein du RIS, selon lequel un nœud peut endosser plusieurs caractéristiques. Autres exemples : une ressource peut aussi être un "agent", une "collection" peut aussi être un "item" qui peut être agrégé dans d'autres collections. L'objet numérique de Fedora rend tout cela possible, sans les contraintes des architectures orientées "objet" avec relation d'héritage unique. Par essence, un objet numérique peut revêtir toute combinaison d'identités types.

5.3 Collections et agrégations

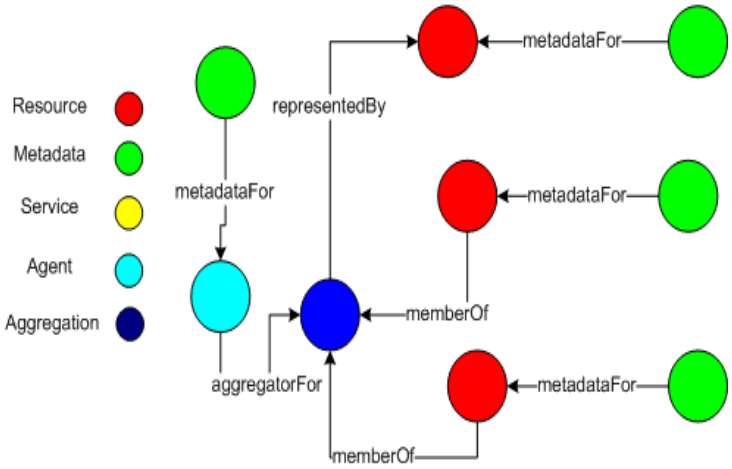


Figure 10

La phase I de l'architecture de la NSDL ne permettait qu'une seule forme d'agrégation, qui exprime les relations entre un fournisseur de métadonnées et l'ensemble des métadonnées collectées chez ce fournisseur. L'implémentation de l'EDN inclut un modèle d'agrégation fondé sur des ensembles, qui permet à un nombre quelconque de ressources d'être contenues dans un nombre quelconque d'ensembles. Comme le montre l'illustration, une agrégation est représentée par ("representedBy") une ressource. Cette ressource fournit la sémantique pour l'agrégation. Par exemple, une agrégation peut indiquer un ensemble de ressources qui se rapportent à un standard pédagogique officiel. Dans ce cas, la ressource au bout du lien "representedBy" exprime ce standard. Les agrégations sont elles-mêmes des "ressources", qui peuvent être imbriquées dans des agrégations additionnelles. Disposant d'une sémantique, ces agrégations constituent, dans le modèle de données, les fondations pour une contextualisation riche des ressources de la bibliothèque numérique.

5.4 Avancement de l'EDN

Au moment de la rédaction de cet article, le chargement initial de l'EDN, à partir des données de l'entrepôt préexistant, est presque terminé (plus de 1,1 million de notices). Le graphe RIS qui en résulte contient environ 1,5 million de nœuds et 10 millions d'arêtes explicites (d'autres arêtes implicites étant propres à Fedora). Dans un article à venir, nous rendrons compte de nos expériences concernant les triplets RDF et de la démarche d'élever le RIS à un niveau supérieur. En particulier, nous avons remarqué qu'un grand nombre d'indicateurs de performance sur les triplets s'appliquent à d'autres applications du web sémantique.

Une fois achevé le chargement de données, nous communiquerons la spécification de l'API de l'EDN à la communauté de la NSDL. Ce sera le début du processus de déploiement et d'approfondissement du RIS, qui s'appuiera sur l'information contextuelle ajoutée par la communauté de la NSDL. Des résultats intéressants concernant la nature du RIS sont à prévoir, au fur et à mesure qu'il se développera.

Conclusion

A l'heure de Google, qu'est-ce qu'une bibliothèque numérique, au juste ? Une telle question ne peut qu'enflammer les passions. Nous avons ardemment défendu les succès accomplis, en une décennie, par la communauté des bibliothèques numériques. Mais la stupéfiante réussite des moteurs de recherche commerciaux a changé la donne. Les fonctions de recherche et d'accès sur un ensemble de ressources, en dépit de leur importance, ne suffisent pas. Les bibliothèques numériques ont besoin de se distinguer des moteurs de recherche par la façon dont elles ajoutent de la valeur aux ressources internet. Cette valeur ajoutée consiste à mettre ces ressources en contexte, à les enrichir par de nouvelles informations et des relations qui expriment les modèles d'usage et le savoir de la communauté servie par la bibliothèque. La bibliothèque numérique devient alors un espace pour l'information collaborative et l'enrichissement - bien plus qu'un simple endroit où trouver de l'information et y accéder.

Le travail que nous avons mené au sein de la NSDL a démontré que le modèle centré sur les métadonnées, que tous connaissent, est insuffisant pour ce type de fonctionnalité. Nous avons conçu et implémenté un réseau d'information superposé au sein de Fedora, qui comporte toutes les fonctionnalités de l'entrepôt de métadonnées existant, mais qui modélise des relations, des services et de multiples types d'information à l'intérieur d'une application service web. Ce riche dépôt d'information fournira les bases de la prochaine étape de notre travail : implémenter une suite non limitée de services à l'utilisateur, à même de réaliser le "laboratoire pour le partage et la construction du savoir" imaginé dans le rapport initial du projet NSDL [59].

Remerciements : *Cet article reprend le travail de plusieurs personnes, en plus des auteurs. Le groupe Fedora, tout spécialement Chris Wilper et Eddie Shin, mérite un hommage pour le travail difficile qu'a nécessité l'implémentation de ces notions dans le logiciel libre Fedora. Les membres du groupe NSDL, en particulier Tim Cornwell, Elly Cramer et Naomi Dushay, ont joué un rôle majeur dans la formulation du modèle de données NSDL et son implémentation dans l'EDN. Le groupe NSDL dans son ensemble adresse ses plus vifs remerciements à Lee Zia, qui défend le projet auprès de la NSF depuis des années. Les réalisations décrites ici ont bénéficié de plusieurs subventions. Le travail concernant l'EDN de la NSDL a bénéficié des subventions n° 0227648, 0227656 et 0227888 de la National Science Foundation. Le travail concernant les réseaux d'information superposés a bénéficié de la subvention n° 0430906 de la National Science Foundation. Le travail sur Fedora est financé par la Andrew W. Mellon Foundation. Toutes les opinions, conclusions et recommandations contenues dans cet article sont celles de leurs auteurs et ne reflètent pas nécessairement les points de vue de la National Science Foundation ou de la Andrew W. Mellon Foundation. Un grand merci à Lucy Lagoze qui a montré à Carl Lagoze combien il est difficile pour un étudiant d'utiliser des moteurs de recherche et qui a livré quelques enseignements sur l'importance du contexte et des modèles d'usage.*

Merci, également, à Mike Keller et Vicky Reich de nous avoir autorisé à adopter et adapter un titre qu'ils ont utilisé dans un article antérieur [23].

Note du traducteur : [1] NDT : en anglais, Information network overlay (INO)

Bibliographie

[1] "Functional Requirements for Bibliographic Records," International Federation of Library Associations and Institutions March 1998. <http://www.ifla.org/VII/s13/frbr/frbr.pdf>.

[2] New Pathways to the National Science Digital Library, 2004 <http://www.infosci.cornell.edu/news/NSDL_Pathways.pdf>.

[3] "Report of the Science, Mathematics, Engineering, and Technology Education Library Workshop," National Science Foundation, Washington, DC, Workshop Report July 21-23 1998. <http://www.dlib.org/smete/public/report.html>.

[4] J. Abbas, C. Norris, and E. Soloway, "Middle School Children's Use of the ARTEMIS Digital Library," presented at ACM/IEEE Joint Conference on Digital Libraries (JCDL '02), Portland, OR, 2002.

[5] D. G. Andersen, H. Balakrishnan, and M. F. Kaashoek, "Resilient Overlay Networks," presented at 18th ACM SOSP, Banff, Canada, 2001.

[6] W. Y. Arms, N. Dushay, D. W. Fulker, and C. Lagoze, "A Case Study in Metadata Harvesting : the NSDL," Library Hi Tech, 21 (2), 2003.

[7] W. Y. Arms, D. Hillmann, C. Lagoze, D. Krafft, R. Marisa, J. Saylor, C. Terrizzi, and H. Van de Sompel, "A Spectrum of Interoperability : The Site for Science Prototype for the NSDL," D-Lib Magazine, 8 (1), 2002. <doi:10.1045/january2002-arms>.

[8] D. E. Atkins, Report of the Santa Fe Planning Workshop on Distributed Knowledge Work Environments : Digital Libraries, 1997 <http://www.si.umich.edu/SantaFe/report.html>.

[9] K. Auletta, "Under the Wire," New Yorker, January 17, 1994.

[10] C. Baru, R. Moore, A. Rajasekar, and M. Wan, "The SDSC Storage Resource Broker," presented at CASCON'98, Toronto, 1998.

[11] D. Bearman, G. Rust, S. Weibel, E. Miller, and J. Trant, "A Common Model to Support Interoperable Metadata. Progress report on reconciling metadata requirements from the Dublin Core and INDECS/DOI Communities," D-Lib Magazine, 5 (January), 1999. <doi:10.1045/january99-bearman>.

[12] J. Bekaert, P. Hochstenbach, and H. Van de Sompel, "Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National

Laboratory Digital Library," D-Lib Magazine, 9 (11), 2003. <doi:10.1045/november2003-bekaert>.

[13] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, (50), May, 2001.

[14] C. L. Borgman, "Digital libraries and the continuum of scholarly communication," Journal of Documentation, 56 (4), pp. 412-430, 2000.

[15] C. L. Borgman, "The invisible library : Paradox of the global information infrastructure," Library Trends, 51 (4), pp. 652, 2003.

[16] C. L. Borgman, "What are digital libraries ? Competing visions," Information Processing & Management, 1999 (35), pp. 227-243, 1999.

[17] D. Brickley and R. V. Guha, "RDF Vocabulary Description Language 1.0 : RDF Schema," W3C, Recommendation February 10 2004. <http://www.w3.org/TR/rdf-schema>.

[18] A. Crespo and H. Garcia-Molina, "Semantic overlay networks for p2p systems," Stanford University, Palo Alto 2003.

[19] R. Daniel Jr. and C. Lagoze, "Extending the Warwick Framework : From Metadata Containers to Active Digital Objects," D-Lib Magazine (November), 1997. <doi:10.1045/november97-daniel>.

[20] E. Fox, R. M. Akscyn, R. K. Furuta, and J. J. Leggett, "Digital libraries," Communications of the ACM, 38 (4), pp. 22-28, 1995.

[21] F. Fukuyama, The end of history and the last man. New York, Toronto : Free Press, 1992.

[22] R. Kahn and R. Wilensky, "A Framework for Distributed Digital Object Services," Corporation for National Research Initiatives, Reston, Working Paper cnri.dlib/tn95-01, 1995. <http://www.cnri.reston.va.us/k-w.html>.

[23] M. A. Keller, V. Reich, and A. C. Herkovic, "What is a library anymore, anyway ?," First Monday, 8, May 5, 2003.

[24] C. Lagoze, "The Warwick Framework : A Container Architecture for Diverse Sets of Metadata," D-Lib Magazine, 2 (7/8), 1996. <doi:10.1045/july96-weibel >.

[25] C. Lagoze, W. Arms, S. Gan, D. Hillmann, C. Ingram, D. Krafft, R. Marisa, J. Phipps, J. Saylor, C. Terrizzi, W. Hoehn, D. Millman, J. Allan, S. Guzman-Lara, and T. Kalt, "Core Services in the Architecture of the National Digital Library for Science Education (NSDL)," presented at Joint Conference on Digital Libraries, Portland, Oregon, 2002.

[26] C. Lagoze and J. R. Davis, "Dienst - An Architecture for Distributed Document Libraries," Communications of the ACM, 38 (4), pp. 47, 1995.

[27] C. Lagoze, S. Payette, E. Shin, and C. Wilper, Fedora : An Architecture for Complex Objects and their Relationships, 2005 <http://arxiv.org/abs/cs.DL/0501012>.

[28] C. Lagoze and A. Singhal, "Information Discovery : Needles and Haystacks," IEEE Internet Computing, 2005 (May/June), 2005.

[29] C. Lagoze, H. Van de Sompel, M. Nelson, and S. Warner, The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0, 2002 <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

[30] D. Levy, "Cataloging in the Digital Order," presented at The Second Annual Conference on the Theory and Practice of Digital Libraries, 1995.

[31] D. Levy, "Digital Libraries and the Problem of Purpose," Bulletin of the American Society for Information Science, 26 (6), 2000.

[32] Library of Congress, METS : An Overview & Tutorial, 2004 <http://www.loc.gov/standards/mets/METSOverview.v2.html>.

[33] C. Lynch, "The Battle to Define the Future of the Book in the Digital World," First Monday, 6 (6), June 4, 2001.

[34] C. A. Lynch and H. Garcia-Molina, "Interoperability, Scaling, and the Digital Libraries Research Agenda," IITA Digital Libraries Workshop May 18-19 1995. <http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html>.

[35] C. A. Lynch and M. A. Keller, googlization, digital repositories, distance education, and privacy, 2005 <http://www.learningtimes.net/acrlarchive.html>.

[36] B. Marshall, Y. Zhang, H. Chen, A. Lally, R. Shen, E. A. Fox, and L. Cassel, "Convergence of Knowledge Management and E-Learning : the GetSmart Experience," presented at ACM/IEEE Joint Conference on Digital Libraries (JCDL '03), Houston, TX, 2003.

[37] K. Martin, "Learning in Context," Issues of Teaching and Learning, 4 (8), September, 1998.

[38] G. McCalla, "The Ecological Approach to the Design of E-Learning Environments : Purpose-based Capture and Use of the Information about Learners," Journal of Interactive Media in Education, 7 (Special Issue on the Educational Semantic Web), 2004.

[39] F. McMartin and Y. Terada, "Digital Library Services for Authors of Learning Materials," presented at ACM/IEEE Joint Conference on Digital Libraries (JCDL '02), Portland, OR, 2002.

[40] National information Standards Organization (U.S.), The OpenURL Framework for Context-Sensitive Services, 2003<http://www.niso.org/standards/resources/Z39_88_2004.pdf>.

[41] W. Nejdl, B. Wolf, and C. Qu, "EDUTELLA : A P2P Networking Infrastructure Based on RDF," presented at WWW2002, Honolulu, 2002.

[42] A. M. Ouksel and A. Sheth, "Semantic Interoperability in Global Information Systems," SIGMOD Record, 28 (1), 1999.

[43] P. Parrish, "The Trouble with Learning Objects," Educational Technology Research and Development, 52 (1), pp. 49-67, 2004.

[44] S. Payette and C. Lagoze, "Flexible and Extensible Digital Object and Repository Architecture (FEDORA)," presented at Second European Conference on Research and Advanced Technology for Digital Libraries, Heraklion, Crete, 1998.

[45] President's Information Technology Advisory Committee : Panel on Digital Libraries, "Digital Libraries : Universal Access to Human Knowledge," PITAC February 2001. <http://www.itrd.gov/pubs/pitac/pitac-dl-9feb01.pdf>.

[46] M. Recker, Instructional Architect, 2004 <http://ia.usu.edu/>.

[47] M. Recker, J. Dorward, and L. M. Nelson, "Discovery and Use of Online Learning Resources : Case Study Findings," Educational Technology and Society, 7 (2), pp. 93-104, 2004.

[48] M. Recker and A. Walker, "Collaboratively filtering learning objects," in Designing Instruction with Learning Objects, D. A. Wiley, Ed., 2000.

[49] T. C. Reeves, The Impact of Media and Technology in Schools : A Research Report prepared for The Bertelsmann Foundation, 1998. <http://it.coe.uga.edu/ treeves/edit6900/BertelsmannReeves98.pdf>.

[50] V. Reich, "LOCKSS : A Permanent Web Publishing and Access System," D-Lib Magazine, 7 (6), 2001. <doi:10.1045/june2001-reich>.

[51] G. Salton, Dynamic information and library processing. Englewood Cliffs, N.J. : Prentice-Hall, 1975.

[52] A. Seaborne, RDQL - A Query Language for RDF, 2004. <http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/>.

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

[54] J. Surowiecki, The wisdom of crowds : why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations, 1st ed. New York : Doubleday :, 2004.

[55] E. Svenonius, The intellectual foundation of information organization. Cambridge, Mass. : MIT Press, 2000.

[56] Tucana Technologies, iTQL Commands, 2004 <http://kowari.org/oldsite/271.htm>.

[57] Tucana Technologies, Kowari metastore, 2004 <http://www.kowari.org/>.

[58] J. Ward, "A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative," presented at Joint Conference on Digital Libraries, Houston, 2003.

[59] F. Wattenberg, "A National Digital Libraries for Science, Mathematics, Engineering, and Technology Education," D-Lib Magazine, 1998 (October), 1998. <doi:10.1045/october98-wattenberg>.

[60] D. A. Wiley, "Connecting learning objects to instructional design theory : A definition, a metaphor, and a taxonomy," in The Instructional Use of Learning Objects : Online Version, D. A. Wiley, Ed., 2000.

[61] L. L. Zia, "The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program," D-Lib Magazine, 8 (11), 2002. <doi:10.1045/november2002-zia>.

αMETIST

Partie 1

Partie 2 :
Capitalisation / Mutualisation

Partie 3

Partie 4

D'un thésaurus vers une ontologie de domaine pour l'exploration d'un corpus

Claude Chrisment (1)
chrisment@irit.fr

Nathalie Hernandez(1)
hernandez@irit.fr

Françoise Genova (2)
genova@cluster.u-strasbg.fr

Josiane Mothe (1,3)
mothe@irit.fr

(1) IRIT, 118 route de Narbonne, 31040 Toulouse CEDEX, France
(2) CDS, 11 rue de l'Université, 67000 Strasbourg, France
(3) IUFM, 56 av. de l'URSS, 31078 Toulouse CEDEX, France

Mots-clés : astronomie, fouille texte, thésaurus, ontologie, linguistique information, linguistique de corpus, méthode, relation sémantique

Keywords : astronomy, text mining, thesaurus, ontology, corpus linguistics, semantic relation

Résumé : Dans cet article, nous proposons une méthodologie et des techniques associées pour permettre de transformer un thésaurus en une ontologie de domaine. L'originalité de notre approche repose sur le fait que l'ontologie est construite à partir de deux sources de connaissances non formalisées : celle issue du thésaurus et celle issue de documents du domaine. Nous avons appliqué notre démarche au cas de la transformation du thésaurus IAU de l'astronomie et les évaluations par des astronomes ont montré l'intérêt de nos propositions. L'utilisation de ressources de connaissances formalisées est un atout pour les systèmes qui doivent explorer des collections de documents de gros volumes.

Introduction

De nombreux thésaurus ont été créés dans différents domaines dans l'objectif de proposer un vocabulaire contrôlé pour l'indexation de ressources documentaires et pour l'aide à la formulation d'une

requête par un documentaliste. Ils ont nécessité de lourds efforts pour leur conception manuelle. L'existence de normes (ISO 2788 et ANSI Z39) permet d'uniformiser leur contenu en termes de liens sémantiques entre unités lexicales (synonymie, liens hiérarchiques et d'association). Cependant, leur format n'est pas normalisé : fichiers ASCII, HTML, bases de données co-existent. Pour faire face à ce problème les normes en cours d'élaboration, dans le cadre du W3C comme SKOS Core[1], visent à faire migrer les thésaurus vers des ressources disponibles sur le Web sémantique en se basant sur le langage OWL. La disponibilité de telles ressources sous format normalisé est un enjeu important dans le domaine de la veille et de l'accès à l'information [6].

La normalisation présente principalement trois avantages. Le premier est que l'uniformisation de leur représentation à partir de langages dédiés au web sémantique (tel que RDF et OWL) permettra à ces ressources d'être distribuées sur le web. De plus, ces ressources pourront être uniformément manipulées à partir d'outils dédiés aux ontologies pour leur visualisation, l'annotation, etc… Enfin, les processus de veille et d'accès à l'information pourront s'appuyer sur ces ressources élémentaires simples, sans avoir à faire face à l'hétérogénéité des formats.

D'un point de vue de la représentation des connaissances, les thésaurus ont un faible degré de formalisation. Ce sont des collections de termes qui sont organisées suivant une ou plusieurs hiérarchies avec des relations entre termes. Les thésaurus n'ont pas de niveau d'abstraction conceptuelle [13]. La distinction entre un concept et sa lexicalisation n'est pas clairement établie. Les relations de synonymies sont établies entre les termes mais les concepts ne sont pas identifiés. Ceci s'explique par l'utilisation initiale des thésaurus, qui n'ont pas pour objectif de refléter comment le monde peut être compris en termes de sens mais en termes de terminologie et de catégories servant à l'indexation manuelle de documents d'un domaine. Pour réduire la complexité de leur élaboration, les concepteurs de thésaurus n'ont pas intégré ce niveau d'abstraction.

[1] http ://www.w3.org/TR/swbp-skos-core-guide/

De plus, la couverture sémantique des thésaurus est limitée. En effet, les relations entre termes sont vagues et ambiguës. Les liens sémantiques qu'ils contiennent reflètent parfois l'utilisation prévue du thésaurus plutôt que les liens sémantiques réels entre termes. Ces relations peuvent ainsi englober les relations « est une instance de » ou « est une partie de » [4]. La relation associative « est lié à » est souvent difficile à exploiter car elle connecte des termes en sous-entendant différents types de relations sémantiques [14]. Par exemple, dans le thésaurus BIT[2] relatif au monde du travail, le terme « famille » est lié aux termes « femme » et « congé familial », la relation sémantique entre ces deux paires de termes est intuitivement différente. Par les choix faits lors de leur conception, les thésaurus manquent de formalisation et de cohérence par rapport aux ontologies légères.

Les ontologies légères ou formelles ne posent pas ce type de problème. Elles sont supposées respecter la relation de subsomption dans l'organisation hiérarchique des concepts. D'autre part, les liens d'associations entre concepts sont sémantiquement mieux décrits. Cependant, leur élaboration est coûteuse ; elle nécessite de nombreuses interventions manuelles. En effet, les techniques de construction d'ontologies de la littérature basent généralement l'élaboration de l'ontologie sur aucune connaissance préalable du domaine. Notre approche vise au contraire à réutiliser les thésaurus de domaine qui ont nécessité de lourds efforts de conception pour l'élaboration de nouvelles ressources d'un niveau formel plus élevé. La conception d'ontologies à partir de thésaurus présente l'avantage de reposer sur l'ensemble des termes qu'il contient et qui ont été identifiés par des experts comme étant représentatifs du domaine. Cependant, elle doit prendre en compte les différences fondamentales entre thésaurus et ontologie. La principale difficulté consiste à capturer la sémantique implicitement présente dans les thésaurus habituellement utilisés par des documentalistes.

En prenant en compte ces principales différences, nous proposons une méthode pour transformer un thésaurus en ontologie légère de domaine pour l'indexation de corpus en plusieurs étapes. Cette méthode vise à s'appliquer à n'importe quel thésaurus de domaine

[2]http://www.ilo.org/public/libdoc/ILO-Thésaurus/french/tr1740.htm

conçu sous les normes ISO 2788 et ANSI Z39. Ces thésaurus sont monolingues et ne sont pas organisés suivant des facettes.

Nous illustrons nos propositions à partir du thésaurus de l'astronomie IAU[3]; les validations s'appuient également sur ce thésaurus.

La section 2 présente la méthode que nous proposons. Cette section explicite les problématiques auxquelles la méthode doit répondre, les différentes étapes qu'elle met en place, ainsi que le schéma conceptuel de l'ontologie choisi. Les sections suivantes décrivent les étapes de la transformation de l'ontologie. La section 4 présente les mécanismes utilisés pour créer le niveau d'abstraction conceptuel à partir du thésaurus. La section 5 explique comment la structure de l'ontologie est construite (liens entre concepts).

1 Présentation de la méthode

La méthode que nous proposons vise à permettre l'élaboration d'une ontologie légère de domaine pour l'exploration de corpus, à partir d'un thésaurus. Afin de capturer la sémantique implicitement présente dans le thésaurus et de mettre à jour la connaissance représentée à partir de la connaissance actuelle d'un domaine, la méthode se base sur l'analyse de documents textuels. Les problématiques auxquelles doit répondre la méthode sont situées dans le cadre général de la construction d'ontologies à partir de textes et reposent sur différentes étapes.

1.1 Cadre général

La méthode que nous proposons s'appuie sur des documents textuels. La méthodologie TERMINAE[1] décrit les différentes étapes dans la construction d'une ontologie à partir de textes. Nous nous basons sur ces étapes pour spécifier la méthode permettant la transformation d'un thésaurus en ontologie. Afin d'identifier les éléments clés dans la transformation d'un thésaurus, nous reprenons

[3] http://www.site.uottawa.ca:4321/astronomy/index.html

les étapes de la méthodologie et les choix que nous faisons pour chacune d'entre elles.

1) La première étape de la méthodologie TERMINAE vise à spécifier les besoins auxquels doit répondre l'ontologie. Dans le cas de la transformation d'un thésaurus en ontologie légère de domaine pour l'exploration de corpus, les besoins que nous identifions sont les suivants :

- la spécification des termes du domaine et de leurs variantes lexicales afin de les détecter dans les granules documentaires,
- le regroupement de ces termes en concepts afin de déterminer les objets et notions référencés dans les documents,
- la structuration des concepts à partir de relations taxonomiques et associatives afin de permettre une indexation sémantique de qualité,
- la formalisation de l'ontologie dans un langage interprétable par le système afin qu'il soit capable de la manipuler.

Une ontologie légère créée pour l'exploration de corpus doit donc intégrer ces différents éléments.

2) La deuxième étape repose sur le choix du corpus de référence à partir duquel l'ontologie est construite. Ce choix est un paramètre déterminant de l'élaboration de l'ontologie[3]. Le corpus doit décrire les éléments de connaissance qui seront intégrés dans l'ontologie. Dans le cas de la transformation d'un thésaurus, le corpus doit répondre à deux conditions. Il doit tout d'abord permettre de capturer la connaissance implicite qui n'est pas formalisée dans le thésaurus. Ensuite, le corpus doit aider à la mise à jour de la connaissance à partir de documents récents du domaine. L'ontologie étant créée pour des activités d'exploration de corpus, le corpus considéré doit aider à préciser le contexte associé à des documents du domaine d'intérêt considéré. Dans notre approche, le corpus est extrait de corpus existants et des experts doivent valider qu'il couvre l'ensemble du domaine sur une période représentative. Des résumés d'articles publiés dans des revues du domaine permettent de décrire ce type d'information. Les articles complets pourraient être utilisés

mais l'avantage des résumés est que les informations qu'ils détiennent sont synthétisées.

3) La troisième étape est celle de l'étude linguistique du corpus. Cette étape vise à extraire des documents les termes représentatifs du domaine et leurs relations (lexicales et syntaxiques) en utilisant des outils dédiés. A la fin de cette étape, on obtient un ensemble de termes, de relations entre ces termes et des regroupements. Dans le cadre de la transformation d'un thésaurus, cette étape intègre la connaissance représentée dans le thésaurus. Les termes présents dans le thésaurus sont représentatifs du domaine. Ils peuvent être regroupés à partir des relations du thésaurus. L'étude linguistique du corpus de référence est également nécessaire pour extraire les termes du domaine non présents dans le thésaurus et les relations entre termes qui n'y sont pas explicitées. Afin d'effectuer cette analyse, nous utilisons l'analyseur syntaxique SYNTEX[2]. Cet analyseur a l'avantage de se baser sur un apprentissage endogène pour effectuer des analyses sur des corpus de différents domaines. Il permet d'extraire les syntagmes des documents ainsi que leur contexte d'apparition (mots qu'ils régissent et par qui ils sont régis). Une méthode additionnelle doit cependant être élaborée pour définir les mécanismes permettant de sélectionner les termes et leurs relations, à partir de la connaissance extraite du thésaurus et des informations extraites du corpus. La méthode que nous proposons vise à répondre à cette problématique.

4) La quatrième étape correspond à la normalisation des résultats obtenus à l'étape précédente. A partir des termes et des relations lexicales, des concepts et des relations sémantiques sont définis. Au niveau de cette étape, le thésaurus peut être utilisé pour aider à la spécification des concepts.

5) La dernière étape est celle de la formalisation, le réseau sémantique défini à l'étape précédente est traduit dans un langage formel. La formalisation de l'ontologie créée à partir d'un thésaurus peut être réalisée à partir du langage OWL[9]. Ce langage, au cœur du web sémantique, a l'avantage d'être constitué de trois sous-langages d'un niveau de formalisation incrémentale. L'utilisation d'OWL-Lite permet une première formalisation de l'ontologie qui pourra évoluer. Ce langage permet de plus, de représenter l'ensemble des éléments

spécifiés par les besoins auxquels doit répondre une ontologie légère dans le cadre de l'exploration de corpus.

Pour la transformation d'un thésaurus, la méthode que nous proposons vise donc à mettre en œuvre les étapes 3, 4 et 5 spécifiées dans la méthodologie TERMINAE. Elle se base sur un mécanisme décomposé en différentes étapes décrites dans la section suivante.

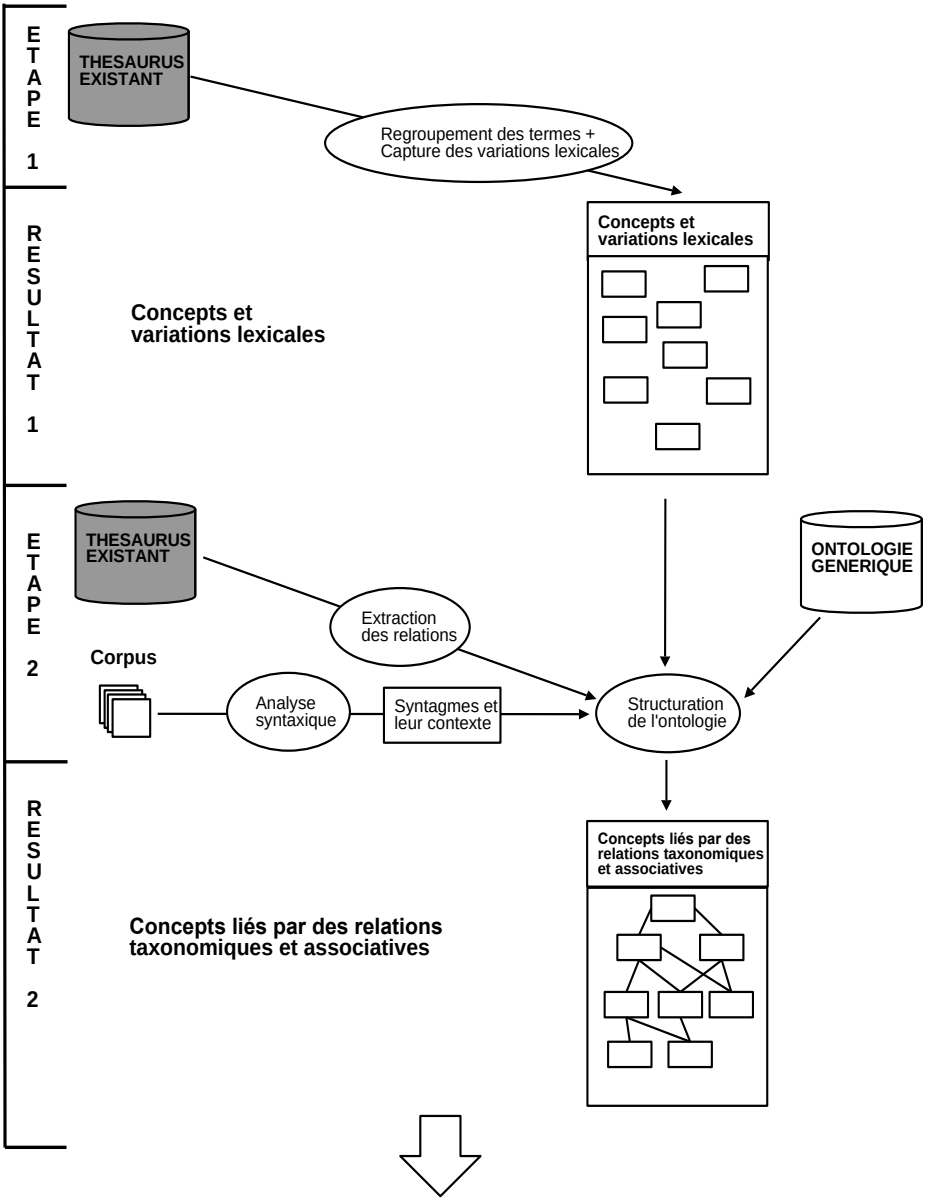
1.2 Etapes de la méthode

La méthode proposée repose sur trois étapes. Ces étapes sont décrites dans la figure 1.

La première étape vise à extraire du thésaurus un ensemble de concepts ainsi que leurs variations lexicales. Peu de méthodes dans la littérature mettent en œuvre cette étape. La majorité d'entre elles considère en effet qu'un concept est référencé à partir d'un seul terme [7][10][11][12].
Par l'utilisation d'un thésaurus, notre méthode vise à proposer un mécanisme automatique de regroupement des labels d'un même concept. Cette étape est décrite dans la section 3.

La deuxième étape permet de structurer les concepts de l'ontologie à partir de la détection de relations taxonomiques et associatives dans le thésaurus et dans le corpus. Cette étape soulève différentes problématiques de la construction d'ontologies. L'une d'elles relève de la difficulté à organiser les concepts par des relations taxonomiques. Dans notre cas, les relations hiérarchiques entre termes du thésaurus peuvent être utilisées pour aider à la détection de ces relations.

Cependant, un des inconvénients des thésaurus est que le niveau hiérarchique le plus général est souvent composé de nombreux termes. Afin d'organiser les concepts à partir d'un niveau d'abstraction comportant un nombre limité de concepts, nous proposons l'utilisation d'une ontologie générique.


ETAPE 1
THESAURUS EXISTANT
Regroupement des termes + Capture des variations lexicales
RESULTAT 1
Concepts et variations lexicales
Concepts et variations lexicales
ETAPE 2
THESAURUS EXISTANT
ONTOLOGIE GENERIQUE
Extraction des relations
Corpus
Analyse syntaxique
Syntagmes et leur contexte
Structuration de l'ontologie
RESULTAT 2
Concepts liés par des relations taxonomiques et associatives
Concepts liés par des relations taxonomiques et associatives

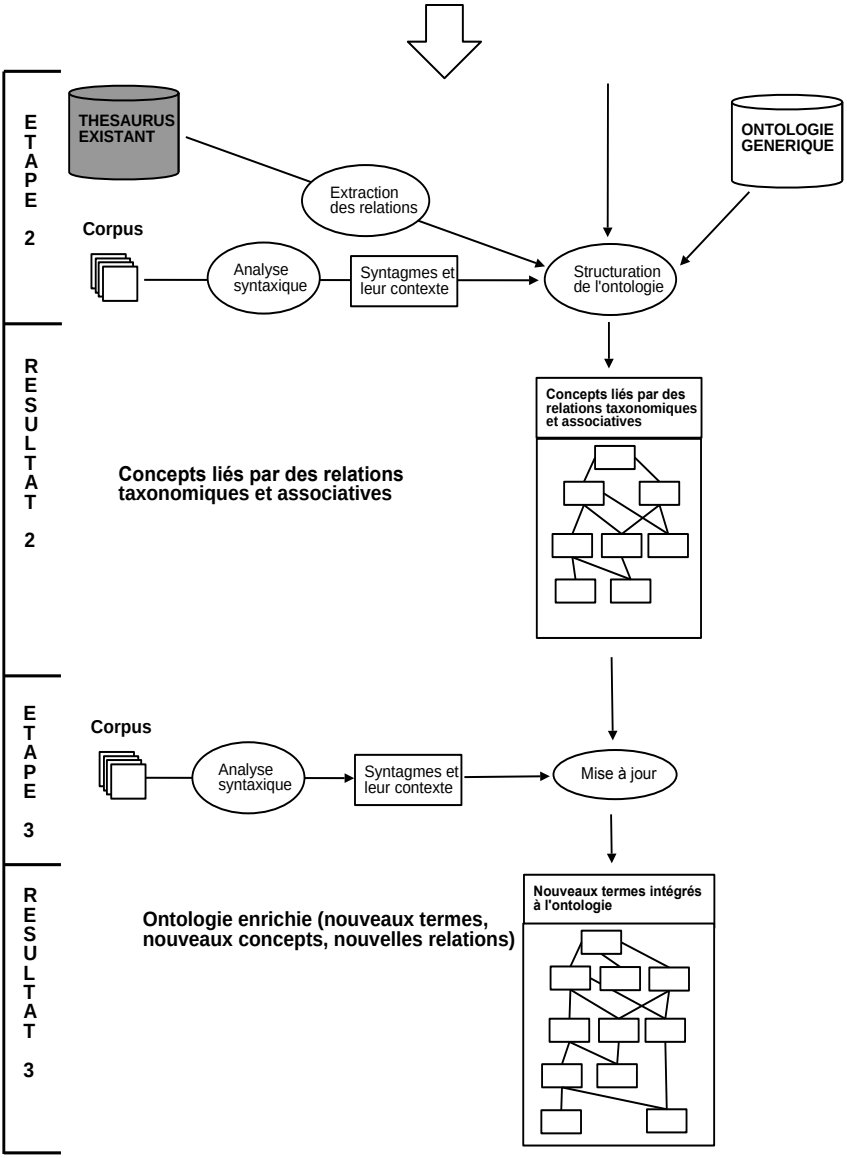


Figure 1 : Etapes de la méthode

Cette ontologie est utilisée pour définir semi-automatiquement les types abstraits du domaine et structurer l'ontologie. Le mécanisme développé est décrit dans la section 4. Une autre problématique que fait intervenir cette étape est la détection de relations associatives entre concepts et la désignation de ces relations sémantiques. Peu de méthodes désignent correctement les labels des relations. Nous proposons un mécanisme visant à proposer de façon semi-automatique ces relations ainsi que leur label. Le mécanisme repose sur l'analyse syntaxique du corpus de référence qui permet d'extraire les syntagmes constituant le lexique du corpus ainsi que le contexte dans lequel ils apparaissent (noms et verbes qu'ils régissent et par qui ils sont régis). Il est décrit dans la section 5.

Contrairement à ce que préconise la méthodologie TERMINAE, la formalisation de l'ontologie est réalisée à la fin de ces différentes étapes après la validation des éléments proposés par un expert du domaine. Ce choix est justifié par le besoin, du concepteur de l'ontologie et de l'expert du domaine, de visualiser les éléments de connaissance jusque là représentés. Le schéma conceptuel utilisé et la formalisation qui lui est associée sont présentés dans la section suivante.

1.3 Schéma conceptuel

Le schéma conceptuel définit la structure de l'ontologie qui est élaborée. Cette structure doit faciliter la transformation d'un thésaurus traditionnel en une ontologie et permettre la représentation des éléments spécifiés par les besoins auxquels doit répondre l'ontologie. Il se veut simple pour permettre son adaptation à tout thésaurus respectant les normes ISO 2788 et ANSI Z39. Comme nous l'avons justifié plus haut, l'implantation de ce schéma repose sur des éléments spécifiés dans le langage OWL-Lite.

Le haut niveau conceptuel est présenté dans la figure 2 et est décrit dans les différents sous-paragraphes suivants.

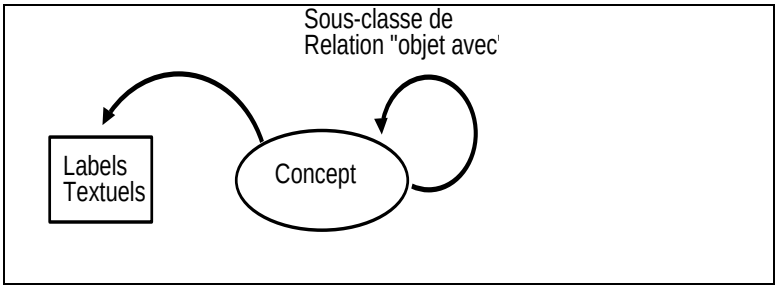


Figure 2 : Haut niveau du schéma conceptuel de l'ontologie

1.3.1 Concept et label textuel

Un concept est représenté à partir d'une classe OWL <owl:Classrdf:about="identifiant_unique">.

Elle est reconnue à partir d'un identifiant unique. Les labels d'une classe sont représentés par la propriété label <rdfs:label>. Dans le schéma conceptuel proposé dans [8], les labels sont de deux types : les labels représentant les termes principaux et ceux représentant les variations lexicales de ces termes. Cette approche peut être intéressante dans le cas où l'application et l'utilisateur doivent avoir une vision différente du contenu de l'ontologie. Dans la mesure où cette différenciation n'a pas lieu d'être dans la recherche de l'adéquation entre une ontologie et un corpus, ni dans le processus de RI, nous avons choisi de ne pas différencier les labels par rapport à leur rôle dans la désignation du concept. Les différentes variations lexicales des termes désignant le concept sont ainsi représentées par cette même propriété.

1.3.2 Relation entre concepts

Les concepts sont ensuite organisés à partir de relations taxonomiques représentées par la propriété <rdfs:subClassOf>.

Les concepts peuvent aussi être reliés entre eux à partir de relations non taxonomiques. Ce type de relations est représenté par l'intermédiaire de la propriété <owl:ObjectProperty> qui permet de lier deux concepts en spécifiant le concept de départ de la relation (rdfs:domain) et le concept d'arrivée (rdfs:range).

Des propriétés peuvent être ajoutées à la relation, telles que :

- la transitivité (<rdf:typerdf:resource="&owl;TransitiveProperty"/>),
- la symétrie (<rdf:typerdf:resource="&owl;SymmetricProperty"/>),
- la fonctionnalité (<rdf:typerdf:resource="&owl;FunctionalProperty"/>),
- l'inverse d'une autre relation <owl:inverseOfrdf:resource="#nom_propriété_inverse"/>.

1.3.3 Schéma conceptuel d'un thésaurus

L'ensemble des éléments du schéma conceptuel précédemment décrit ne sont pas présents dans un thésaurus. Un thésaurus est un ensemble de termes organisés suivant un nombre restreint de relations [5]. Les relations présentes dans un thésaurus répondant aux normes ANSI Z39 et ISO 2788 sont rappelées dans la figure 3.

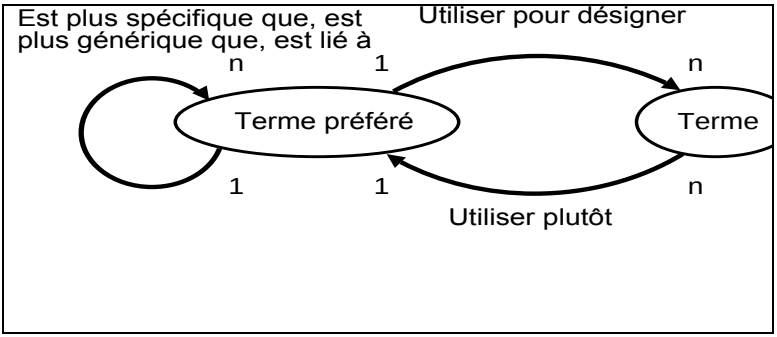


Figure 3 : Rappel des relations entre termes dans un thésaurus

Dans notre méthode, nous considérons que les thésaurus réutilisés pour construire une ontologie sont de ce type.

Nous faisons trois hypothèses sur la réutilisation des relations entre termes :

- les termes préférés sont les termes principaux du domaine et sont des indices pour constituer les termes désignant les concepts du domaine,
- les relations entre termes et termes préférés sont des relations de synonymies entre termes, elles permettent de regrouper les termes comme étant label d'un même concept,
- les relations entre termes préférés sont des indices pour définir des relations entre concepts.

A partir de ces hypothèses, nous définissons des méthodes permettant d'extraire les éléments du schéma conceptuel de l'ontologie d'un thésaurus et de documents textuels.

2 Conceptualisation du lexique du thésaurus

Cette étape vise à extraire du lexique du thésaurus une conceptualisation afin de formaliser un premier ensemble de concepts de l'ontologie.

2.1 Regroupement des termes en concepts

2.1.1 Regroupement basé sur les relations explicites UP et UPD

Afin d'extraire les concepts issus du lexique du thésaurus, les termes dits « préférés » ainsi que les relations du type « Utiliser plutôt » (UP) et « Utiliser pour désigner » (UPD) sont analysées. Nous interprétons ces relations comme des relations de synonymies entre termes.

Des groupements de termes sont réalisés à partir de chacun des termes préférés et de l'ensemble des termes auxquels ils sont liés par les relations UP et UPD.

Si t3 UP t1

alors *t1 et t3 sont regroupés*

t1 ***devient*** *terme préféré*

Si t1 UPD t2

alors *t1 et t2 sont regroupés*

t1 ***devient*** *terme préféré*

(R1)

2.1.2 Regroupement basé sur la fermeture transitive des relations UP et UPD

Les groupements précédents sont ensuite agrégés à partir de la fermeture transitive des relations UP et UPD. Dans le cas où un terme préféré à l'origine d'un premier groupement apparaît dans un autre groupement, tous les termes liés au terme préféré et le terme préféré lui-même sont ajoutés aux groupements auxquels il est lié par une des relations.

La fermeture transitive consiste à regrouper les termes à partir de la règle R2.

Si t1 UPD t2 **et** t2 UPD t3

alors t1 UPD t3 *(transitivité)*

t1, t2 et t3 sont regroupés

t1 ***devient*** *terme préféré principal*

Si t1 UP t2 et t2 UP t3

alors t1 UP t3

t1, t2 et t3 sont regroupés

t3 **devient** terme préféré principal

(R2)

La figure 4 schématise plusieurs exemples de groupements. Pour faciliter la lisibilité, les termes préférés sont en gras majuscules. Les termes regroupés par R1 sont soulignés en pointillé. Les termes regroupés par la règle R2 sont soulignés en trait plein.

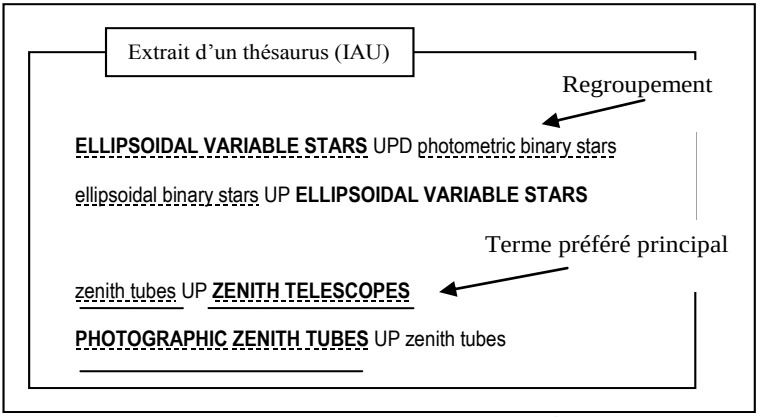


Figure 4 : Exemples de groupements des termes du thésaurus

Les groupements de termes ainsi réalisés constituent l'ensemble des labels des futurs concepts de l'ontologie.

2.1.3 Identifiant du concept

L'identifiant d'un concept est déterminé par le terme préféré à l'origine du groupement. Le choix de ce terme comme identifiant permet de garder un lien entre la future ontologie et le thésaurus. Les identifiants des concepts correspondent ainsi à des entrées du thésaurus. Un terme peut être polysémique (label de plusieurs concepts) dans le cas où il était lié dans le thésaurus à deux termes préférés distincts.

Si « *t1, t2,…et, tn regroupés* » **et** « *t1 terme préféré principal* »

alors *création du concept c d'identifiant t1 et de labels t1, t2,… et tn*

(R3)

2.2 Capture des variations lexicales

La forme lexicale sous laquelle se trouvent les termes du thésaurus est un sujet délicat et largement détaillé dans l'ensemble des normes dédiées au thésaurus [ISO 2788, AFNOR NF Z47-100, ANSI Z39]. Ceci s'explique par l'ambiguïté posée par le rôle des termes dans un thésaurus. Les termes peuvent soit représenter des catégories d'objets similaires, soit désigner le sens des objets. Dans le cas où le terme représente une catégorie, le pluriel du terme est préféré et, dans le cas où le terme définit le sens du terme, le singulier est choisi. Les normes ISO 2788 et ANSI Z39 proposent, pour différencier ces cas de figure, la distinction des termes à partir de leur type : les termes désignant des objets dénombrables et les termes désignant des objets indénombrables.

Lorsque peut être posée la question « combien d'objets représentés par le terme existent ? », le terme est intégré dans le thésaurus au pluriel, dans le cas contraire il l'est au singulier. Ces règles sont scrupuleusement respectées dans la plupart des thésaurus, comme dans le thésaurus de l'astronomie IAU. Il est cependant possible de trouver des variantes de l'application de ces règles.

Par exemple les termes sont au singulier sauf si l'usage impose le pluriel dans les thésaurus suivants :

- BIT

  http://www.ilo.org/public/french/support/lib/indexati/unit1/unit1.htm

  (Terminologie du travail, de l'emploi et de la formation)

- British Museum

  http://www.mda.org.uk/bmobj/Objintro.htm

- Alcohol and Other DrugThésaurus

  http://etoh.niaaa.nih.gov/AODVol1/titlepage.htm

Dans les ontologies, les termes sont utilisés pour référencer des concepts et décrire le sens associé aux objets qu'ils représentent. Il est donc important que les labels de l'ontologie ne représentent pas

des catégories mais des unités de sens. Les termes doivent donc être au singulier.

Des techniques de lemmatisation ou le recours à un expert peuvent être utilisées. Alternativement, une ressource lexicale telle que WordNet peut être utilisée. La figure 5 illustre les concepts identifiés dans la figure 4 pour lesquels les labels sont mis au singulier grâce à WordNet.

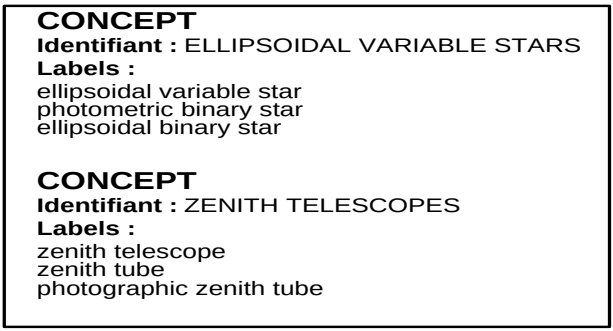


Figure 5 : Exemples de concepts labellisés par des termes au singulier

3 Construction de la structure de l'ontologie

La structure de l'ontologie définit les relations entre concepts établis suite aux étapes présentées dans la section précédente. La structure comprend des relations taxonomiques de type « est un » et des relations associatives qui sont obtenues par les méthodes décrites ici.

Certains liens hiérarchiques entre concepts sont directement issus des liens explicites présents dans le thésaurus. Des niveaux hiérarchiques supérieurs y sont ajoutés à partir de l'analyse des têtes et expansions des labels des concepts et de la création de types abstraits. La figure 6 schématise ces différents mécanismes.

3.1 Etapes de construction

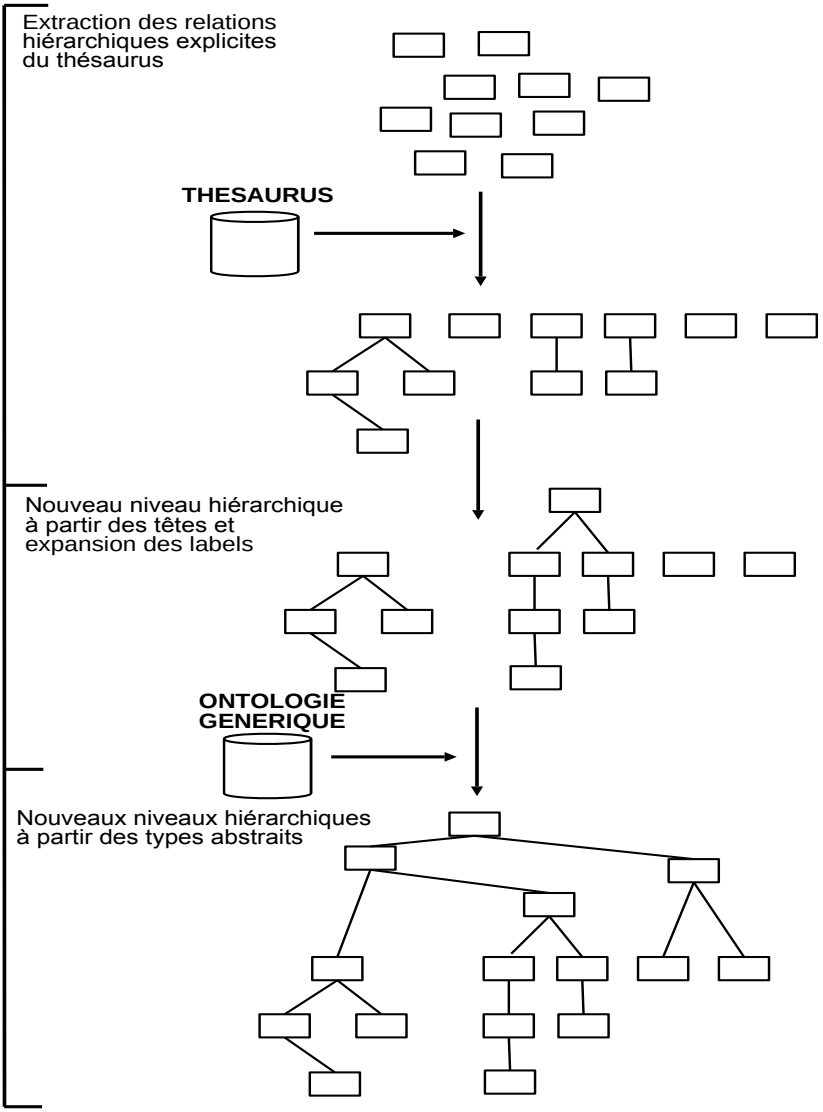


Figure 6 : Mécanisme de construction de la hiérarchie de concepts

3.2 Extractions des relations hiérarchiques explicitées dans le thésaurus

Les concepts sont d'abord organisés hiérarchiquement à partir de la relation « sous-classe de » du schéma conceptuel de l'ontologie. Afin d'extraire ce type de relation du thésaurus, les relations « est plus spécifique que » et « plus générique que » du thésaurus sont prises en compte. L'ensemble de ces relations définies pour les termes, devenus maintenant labels d'un concept, est retenu comme relations candidates pour représenter des relations « sous classes » entre le concept et le concept auquel se rapporte le terme lié dans le thésaurus. Les relations candidates doivent ensuite être analysées avec précaution car elles peuvent englober des relations de type « partie de » ou « instance de ». Nos travaux ne proposent pas de méthode automatique pour réaliser cette désambiguïsation.

Il faut noter que beaucoup de thésaurus de domaine prennent la peine de considérer les relations « est plus spécifique que » et « plus générique que » de façon stricte. Cela est également le cas pour le thésaurus de l'astronomie IAU qui sert de validation à notre approche.

Si « *t1 est plus spécifique que t2* »

et *«t1 label du concept c1 »*

et *« t2 label du concept c2 »*

alors *c1 est une sous-classe de c2*

(R4)

3.3 Suppression de la redondance dans les relations hiérarchiques

Les thésaurus n'étant pas formalisés, des redondances dans la structure hiérarchique de l'ontologie construite avec les règles de R1 à R4 peuvent exister. La relation de généricité est une relation transitive et permet le type d'inférence suivant : si A « est une sous classe de » B et B « est une sous classe de » C, alors A « est une sous classe de » C, A, B, C étant des concepts.

La figure 7 en présente un exemple ; les flèches entre les rectangles représentant les concepts symbolisant la relation « est une sous classe de ». Par la propriété de transitivité de la relation « est une sous classe de », la relation « est une sous classe de » entre palenetary nebula et nebula est donc inutile.

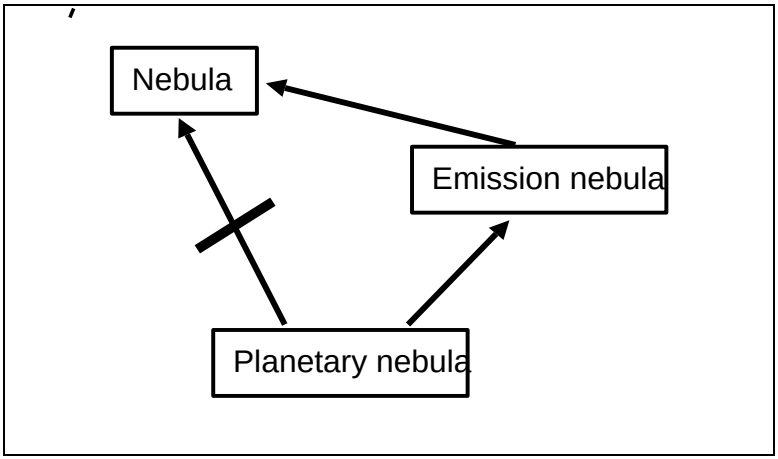


Figure 7 : Exemple de redondance dans la hiérarchie de l'ontologie

Afin de supprimer les relations redondantes, la pertinence de chacune des relations « est une sous classe de » est vérifiée.

La suppression de la redondance est formalisée par la règle R5.

Pour tout concept $c \in C$

Si $\forall$ ($c_i \in C$ et $c \neq c_i$,) , $\exists$ chem1, chem2

tel que chem1=chemin(c,c_i)

et chem2=chemin(c,c_i),

et chem1≠chem2

alors « *suppression de l'arc à l'origine du chemin le plus court »*

(R5)

3.4 Nouveaux niveaux hiérarchiques

Une des lacunes des thésaurus est que leur plus haut niveau hiérarchique contient généralement un très grand nombre de termes [13]. Ces termes sont ceux pour lesquels aucune relation « est plus spécifique » n'a été définie. Ceci s'explique par le fait que les thésaurus ne définissent pas de catégories génériques permettant de répertorier l'ensemble des termes du domaine. Cette même lacune est constatée dans les ontologies obtenues par la transformation d'un thésaurus. Ceci pose problème lorsqu'un utilisateur ou une application choisit d'explorer l'ontologie par une navigation de haut en bas. Le grand nombre de concepts du premier niveau rend le départ de sa navigation délicate. Par exemple, le niveau hiérarchique le plus générique de l'ontologie extraite du thésaurus IAU à cette étape de la transformation contient 1132 concepts.

Nous proposons donc l'ajout de niveaux hiérarchiques plus génériques qui facilitent la navigation dans l'ontologie. D'autre part, nous proposons la définition de concepts génériques (ou types abstraits) permettant de caractériser les concepts.

Un concept générique ou abstrait fait référence à une notion abstraite et n'admet pas d'instance. Il est soit un véritable concept du domaine, soit un concept ajouté pour structurer la représentation. Dans [13], les concepts génériques sont définis à partir d'un schéma de catégorisation de haut niveau existant dans le domaine. Les concepts du plus haut niveau de l'ontologie sont liés manuellement aux concepts de ce schéma. Ce procédé ne peut pas être appliqué à tous les domaines, car de tels schémas n'existent pas toujours. De plus, il demande un travail manuel à l'expert qui doit affecter les milliers de classes de l'ontologie à l'une des centaines de classes du schéma. Nous proposons donc une autre approche plus automatisée.

3.5 Premier niveau de généralisation : tête et expansion des syntagmes

Pour créer un premier niveau d'abstraction, les concepts sont regroupés à partir de la tête des termes de leur label. Cette approche est suivie dans OntoLearn[15] pour créer la hiérarchie de concepts. Les concepts ayant des labels comportant la même tête sont définis comme étant des sous classes du concept labellisé par la tête (règle

R6 et figure 8). Si ce concept n'existe pas dans l'ontologie, il est créé et appartient au nouveau niveau 0 de l'ontologie (règle R7 et figure 9). Ce mécanisme permet de créer un nouveau premier niveau de la hiérarchie contenant un nombre plus réduit de concepts.

Si tete(F^{-1} (c_1))= tete(F^{-1} (c_2))

et tete(F^{-1}(c_1))$\in$ L_{Onto}

alors c_1 « *est une sous-classe de* » F(tete(F^{-1} (c_1))),

c_2 « *est une sous-classe* de » F(tete(F^{-1} (c_1)))

(R6)

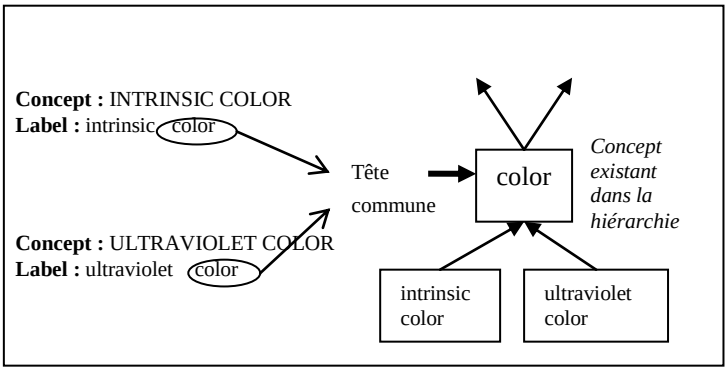


Figure 8 : Nouveau niveau hiérarchique obtenu par la tête des labels appartenant à l'ontologie

Si tete($F^{-1}(c_1)$)= tete($F^{-1}(c_2)$)

et tete($F^{-1}(c_1)$) $\notin L_{Onto}$

alors création d'un nouveau concept c_3, $c_3 \in C_{Onto}$

label de c_3 = tete($F^{-1}(c_1)$),

c_1 « *est une sous classe de* » c_3,

c_2 « *est une sous classe de* » c_3

(R7)

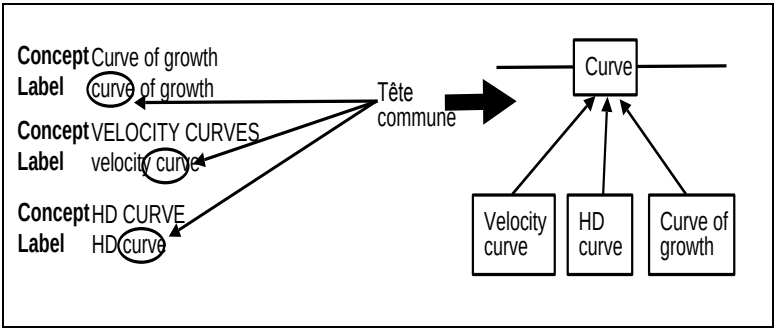


Figure 9 : Nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie

3.6 Deuxième niveau de généralisation : types abstraits

La définition des types abstraits vise à identifier les concepts génériques dont dépendent les concepts du niveau 0 de généralisation précédent. Cette définition comporte deux étapes. Dans un premier temps, il s'agit de définir les types abstraits du domaine, puis de les associer aux concepts. La règle R8 synthétise les étapes qui sont décrites ci-dessous.

Si c ⇔ sw **et** sw∈{synsets_WordNet}

Alors c « *est sous classe* de » ta -- ta pour type abstrait

ta « *est le plus spécifique hyperonyme de* » sw

(R8)

– Définition des types abstraits

Afin de définir ces types de façon automatisée, une ontologie de haut niveau comme par exemple WordNet ou DOLCE, est utilisée. Tout d'abord, les concepts de niveau 0 doivent être mis en correspondance avec les concepts de l'ontologie. Les types abstraits sont alors définis à partir des concepts les plus génériques associés aux concepts détectés. Nous décrivons dans cette section l'utilisation de WordNet pour expliciter cette étape.

Concernant la mise en correspondance des concepts de niveau 0 avec les synsets de WordNet, les labels des concepts de l'ontologie en cours de construction sont comparés aux entrées de WordNet. Chaque synset ainsi détecté est candidat pour représenter le concept dans WordNet. Dans le but de limiter les synsets extraits aux synsets se rapportant effectivement aux concepts de l'ontologie, un mécanisme de désambiguïsation est mis en place. Il prend en compte quatre éléments :

– le glossaire fourni par WordNet pour décrire en langage naturel le sens du synset,

– les synsets descendant du synset en question par la relation hyperonymie dans Wordnet,

– les synsets ancêtres du synset en question dans WordNet par la relation hyponymie dans WordNet,

– les labels des concepts descendant du concept dans l'ontologie par la relation « est sous classe de ».

Lorsque plusieurs synsets correspondent à un label d'un concept de niveau 0, le synset choisi est obtenu par trois méthodes de désambiguïsation qui sont mises en oeuvre séquentiellement.

Les termes très généraux décrivant le domaine traité par l'ontologie, sont tout d'abord spécifiés avec des experts du domaine. Ils sont ensuite recherchés dans le glossaire associé par WordNet à chacun des synsets candidats. Par exemple, le terme recherché dans le glossaire pourrait être « astronomie ». Si un de ces termes est retrouvé, le synset candidat est automatiquement choisi. Sinon, la méthode (2) est appliquée.

Les synsets fils du synset sont comparés aux concepts fils du concept dans l'ontologie. Si au moins un des labels se rapportant aux concepts fils est retrouvé dans les synsets fils, alors le synset est choisi. Sinon, la méthode (3) est appliquée.
Les synsets ancêtres du synset candidat sont analysés par la proposition (1). Un synset candidat est choisi dans le cas où la proposition est vérifiée et, dans le cas contraire, le concept n'est pas associé à un synset de WordNet car aucun synset n'a pu être désambiguïsé.
Concernant l'identification des types, les synsets les plus génériques (i.e. les plus lontains ancêtres) des synsets désambiguïsés sont proposés pour représenter les concepts génériques de l'ontologie. Ils sont ensuite validés par un expert et intégrés à l'ontologie en tant que nouveaux concepts.

– Association des concepts aux types abstraits

Pour les concepts de niveau 0 de l'ontologie ayant été liés à un synset désambiguïsé, un lien est établi entre le concept et le type abstrait correspondant. Le lien est représenté dans l'ontologie en définissant le concept comme sous-classe du type abstrait.

Dans le cas où la désambiguïsation n'a pu avoir lieu ou que les labels du concept n'étaient pas dans WordNet, l'association concept/type abstrait est réalisée manuellement.

La figure 10 présente des exemples de types abstraits extraits pour notre cas d'application.

Property : a basic or essential attribute shared by all members of a class

Phenomenon : any state or process known through the senses rather than by intuition or reasoning

Event : something that happens at a given time

Science : a particular branch of statistic knowledge

Instrumentation : an artifact (or system of artifacts) that is instrumental in accomplishing some end

Substance : that which has mass and occupies space

Relation : an abstraction belonging to or characteristic of two entities or parts together

Location : a point or extent in space

Angle : the space between two lines or planes that intersect ; the inclination of one line to another

Plane : an unbounded two-dimensional shape

Region : the extended spatial location of something ;

Object : a tangible and visible entity

Natural object : an object occurring naturally ; not made by man

Artifact : a man-made object taken as a whole

Figure 10 : Nouveau niveau hiérarchique obtenu par la tête des labels n'appartenant pas à l'ontologie

4 Détection des relations associatives

La deuxième étape dans la formalisation de la structure de l'ontologie vise à définir des relations associatives entre concepts de l'ontologie. Ces relations sont tout d'abord extraites des relations du thésaurus. De nouvelles relations entre les concepts sont ensuite extraites à partir de l'analyse du corpus de référence. Nous présentons dans cette section ces différents éléments.

4.1 Spécification de relations entre types abstraits

La spécification des relations sémantiques entre types abstraits de l'ontologie est fondée sur la proposition de relations associées à chaque type par une analyse syntaxique automatique du corpus de référence. Ces propositions servent de base à la définition manuelle de relations entre paires de type abstrait et sont synthétisées dans la règle R9.

Soient ta_1 et ta_2 deux types abstraits

avec $ta_1, ta_2 \in C_{Onto}$

Soient r et r' deux relations sémantiques

avec $r,r' \in R_{Onto}$

$\sigma_{ROnto}: R_{Onto} \rightarrow C \times C$

$r(ta_1,ta_2)$

$G^{-1}(r)$ spécifiés dans le domaine

Si $r'(c_1,c_2)$

et $c_2, c_2 \in C_{Onto}$

et c_1 « *est sous classe de* » ta_1 , c_2 « *est sous classe de* » ta_2

et $G^{-1}(r')$=« *est lié à* »

alors $G^{-1}(r') \in G^{-1}(r)$

(R9)

4.1.1 Proposition de relations

A partir de l'analyse syntaxique réalisée sur le corpus de référence, le contexte des labels de chacun des concepts est extrait. Nous entendons par contexte, les syntagmes dont les labels sont tête ou expansion, les compléments d'objet et les sujets de verbes dans lesquels les labels apparaissent. Ces contextes sont ensuite regroupés à partir des types abstraits auxquels se rapportent les concepts.

Les termes apparaissant fréquemment dans les contextes regroupés sont retenus pour caractériser le type abstrait et servir de proposition

aux labels des relations associatives que ses concepts fils peuvent avoir. Prenons, pour illustrer cette idée, le cas des contextes des concepts dépendant du type abstrait « instrumentation » dans l'ontologie de l'astronomie. Les termes apparaissant le plus fréquemment sont les verbes anglais « observe » et « mesure ». Ces termes indiquent que les instruments astronomiques sont utilisés pour observer ou mesurer les autres concepts du domaine.

4.1.2 Définition de relations entre types

	Property	**Instrumentation**
Property	Influences Is influenced by Determined by Dertermines Exclude Has part Is part	Makes Observes
Phenomenon	Is a property of induces	Observes Measures
Event	Is a property of induces	Observes Measures
Science	Is studied by	Is Used to studied
Natural Object	Is a property of	Is observed by
Instrumentation	Is maked by Is observed by	Is ou has part exclude

Tableau 1 : Extrait de la matrice des relations entre types abstraits

La définition des relations sémantiques est réalisée entre chaque paire de type abstrait. Une matrice à double entrée est ensuite

réalisée. Cette matrice contient en ligne et en colonne l'ensemble des différents types abstraits identifiés manuellement sur la base des propositions précédentes. Chaque case de la matrice contient les relations possibles. Un extrait de la matrice proposée pour le domaine de l'astronomie est présenté dans le tableau 1.

Il est important de noter que la diagonale de la matrice témoigne de relations particulières. Elles relient en effet des concepts de même type. Une proposition particulière est donc ajoutée pour ce type de relation, la proposition est la relation « partie de ». Les concepts étant de même type, ils peuvent avoir été liés parce que l'un d'eux spécifie une partie de l'autre. Sur la base des propositions précédemment faites, un expert du domaine identifie les relations qui peuvent lier les concepts génériques deux à deux et reporte les labels qu'il choisit dans les cases de la matrice.

4.1.3 Association des relations vagues du thésaurus et des relations entre type

Les relations vagues du thésaurus « est lié à » sont d'abord retranscrites dans l'ontologie. Ainsi, deux termes liés dans le thésaurus donneront lieu à une association entre les concepts dont ils sont labels dans l'ontologie. Cette association est ensuite spécifiée grâce aux relations identifiées dans la matrice entre les types abstraits associés à ces concepts. Par exemple, la relation identifiée entre les types abstraits « instrumentation » et « natural object » étant la relation « observes », la relation « est lié à » du thésaurus entre « coronagraph » et « solar corona » (concepts issus de ces deux types) est modifiée en la relation «coronagraph» « observes » « solar corona ». Si plusieurs relations sémantiques sont identifiées, le choix est laissé à l'expert du domaine.

Le mécanisme mis en place peut s'apparenter à celui proposé dans [Sorgel 2004]. Les relations entre concepts sont en effet établies à partir de l'analyse des relations du thésaurus et de la définition de patrons permettant de retrouver les relations sémantiques spécifiées dans l'ensemble du corpus. Plutôt que d'avoir à spécifier individuellement les relations vagues dans le thésaurus entre termes, l'expert doit seulement valider ou invalider les propositions qui lui sont faites sur la base de l'analyse du corpus et des relations entre les

types abstraits. Ainsi, l'analyse que nous mettons en place facilite le travail de l'expert.

4.2 Détection de nouvelles relations associatives

Contrairement aux approches de la littérature visant uniquement à transformer un thésaurus en ontologie à partir de la connaissance représentée dans celui-ci, nous proposons d'établir de nouvelles relations associatives entre les concepts à partir de l'analyse de documents textuels du domaine (règle R10).

Sur la base de la matrice précédemment établie, de nouvelles relations sont décelées entre les concepts de l'ontologie. Pour cela, le contexte des différents labels des concepts dans le corpus est analysé. Deux approches sont utilisées pour considérer le contexte.

La première prend en compte les termes qui occurrent fréquemment autour des labels de concepts de l'ontologie.

La seconde se base sur l'analyse distributionnelle réalisée par le module UPERY de SYNTEX [2]. Ce type d'analyse consiste à rapprocher des syntagmes en fonction de la ressemblance de leur contexte. Les syntagmes déduits de l'analyse syntaxique sont rapprochés s'ils sont formés autour de la même relation et des mêmes têtes et queues. Par exemple, en considérant les syntagmes « star », « galaxy », « star mass » et « galaxy mass », les syntagmes « star » et « galaxy » sont rapprochés par le contexte « mass ». UPERY permet de rapprocher des syntagmes à partir d'un poids de proximité. Ce poids prend en compte la productivité d'un terme et la productivité d'un concept. A partir d'un seuil fixé empiriquement sur ce poids, le module détecte des relations entre syntagmes mais ne désigne pas la relation sémantique qui les relie. Nous proposons donc d'utiliser les résultats de ce module pour la détection de nouvelles relations associatives qui sont typées par l'intermédiaire de la matrice.

Lorsqu'un label apparaît dans le contexte d'un concept ou dans les termes qui lui sont associés par l'analyse distributionnelle et qu'aucune relation ne lie les deux concepts dans l'ontologie, une relation est proposée entre les deux concepts. Cette relation prend en

compte le type des deux concepts et est établie à partir de la matrice élaborée à l'étape précédente.

Par exemple, dans le contexte du label « luminosity » référençant le concept de même nom, le label « galaxy » correspondant au concept « galaxy » est retrouvé. Ces concepts, étant de type « property » et « natural object », la relation « has a » est proposée entre « galaxy » et « luminosity » (cf tableau 1). Aucune relation n'ayant été précédemment établie entre ces deux concepts, la nouvelle relation est ajoutée à l'ontologie.

Soient ta_1 et ta_2 deux types abstraits

avec $ta_1, ta_2 \in C_{Onto}$

Soient r et r' deux relations sémantiques

(R10)

Conclusion

Le procédé de transformation d'un thésaurus en ontologie légère que nous proposons repose sur trois étapes principales : l'extraction d'information du corpus, l'identification des concepts issus du thésaurus, la construction de structure de l'ontologie (hiérarchie de concepts et relations associatives entre concepts).

Les procédés sont simples à mettre en œuvre et permettent d'extraire une ontologie légère. Ils nécessitent une validation par un expert du domaine, mais le travail qui lui est demandé est allégé par la proposition d'éléments à chacune des étapes. Le travail demandé à l'expert est moins important que celui demandé par les approches proposées dans [13] [16] car son travail consiste uniquement à valider les propositions. Contrairement aux approches présentées dans la littérature, le procédé mis en place vise non seulement à transformer le thésaurus mais aussi à intégrer de nouvelles connaissances dans l'ontologie (ajout de relations entre concepts).

Une contribution importante de notre travail est la proposition permettant de déceler puis de labelliser les relations associatives

entre concepts. Elle repose sur la notion de type abstrait qui correspond à des concepts de haut niveau d'abstraction.

La définition de relations sémantiques, validée par des experts est rapide, compte tenu du nombre limité de types abstraits. Ces relations permettent d'inférer des relations au niveau des concepts de plus bas niveau, en les associant à l'analyse syntaxique du corpus.

Cette méthodologie est bien adaptée lorsque le thésaurus initial est construit en respectant la sémantique de la relation « est un ». En revanche et comme nous l'avons souligné précédemment, lorsque ce n'est pas le cas, une étape supplémentaire doit être ajoutée afin de distinguer les différentes relations telles que « est une partie de » ou « est une instance de ».

Les premières évaluations de nos propositions, dans le cadre de l'astronomie, ont débuté. Les experts du domaine sont satisfaits des résultats. La pertinence des concepts créés et définis à partir de plusieurs labels a été validée par les astronomes. La validation a montré que la totalité des concepts créés étaient pertinents et que pour 85 % d'entre eux l'ensemble des labels était correct. Pour 15 %, les labels ne sont pas corrects car ils se rapportent à des sous-concepts des concepts pour lesquels ils sont définis. Ces labels ont donc été supprimés et ont mené à la création de nouveaux concepts définis comme sous-concepts des concepts auxquels ils étaient rattachés à l'origine.

L'organisation hiérarchique des concepts est réalisée par les règles R4 et R5. Les relations sont définies à partir de la relation « est plus spécifique » « est plus générique » du thésaurus. Ces relations ont mené à la définition de 2882 relations « sous classe de » dans l'ontologie. Parmi celles-ci, 193 relations redondantes ont été trouvées. Elles ont donc été supprimées de la hiérarchie de concepts. Les relations étant définies dans les spécifications du thésaurus pour ne comprendre que des relations du type « est plus spécifique » « est plus générique », seules 5% de ces relations ont été analysées. La totalité d'entre elles a été validée.

Les perspectives de ce travail sont multiples. Concernant le contenu de l'ontologie, une première perspective concerne sa mise à jour. Les thésaurus reflètent des connaissances dans un domaine à un instant

donné. Il est important que cette connaissance puisse être mise à jour. Nous travaillons donc sur la mise à jour d'une ontologie de domaine légère, en s'appuyant sur la connaissance qui peut être extraite des corpus (nouveaux concepts, nouvelles relations). Une autre perspective de ce travail concerne l'utilisation concrète d'une ontologie dans le domaine de la veille. En réalité, nous avons proposé un système d'exploration de corpus (OntoExplo) basé sur des ontologies de domaine [6]. Ce travail s'appuyait sur des ontologies préexistantes. Dans cet article au contraire, nous nous sommes attachées à montrer comment une telle ressource pouvait être construite.

Remerciements **:** *Les travaux présentés dans ce papier ont été réalisés dans le cadre des projets WS-Talk "WS-Talk: Web services communicating in the language of their user community" supporté par le Sixth Framework Programme of the European Community (2002-2006), COOP-CT-2004 006026 et le projet Masse de Données en Astronomie supporté par le ministère délégué à la Recherche et aux Nouvelles Technologies. Nous tenons à remercier particulièrement les astronomes du CDS qui ont évalué nos propositions.*

Bibliographie

[1] N. Aussenac-Gilles, B. Biébow, S. Szulman, Modélisation du domaine par une méthode fondée sur l'analyse de corpus, dans les actes de la conférence IC'2000, Journées Francophones d'Ingénieurie des connaissances, pp 93-103, 2000.

[2] D. Bourigault, Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), Nancy, 2002, pp. 75-84

[3]. A. Condamines, Sémantique et Corpus, Hermès science publications, ISBN 2-7462-1055-X, 2005.

[4] D. H. Fischer, From Thesauri towards Ontologies?, in: el Hadi, Maniez & Pollitt (Eds.): Structures and Relations in Knowledge Organization, dans 5th Int. ISKO Conference, Lille, France, 1998. Würzburg: Ergon, pp. 18-30, 1998.

[5] D.J. Foskett, Thésaurus, In Encyclopedia of Library and Information Science, A. Kent, H. Lancour (Eds), p.416-463, 1980.

[6] N. Hernandez, J. Mothe, Ontologies pour l'aide à l'exploration d'une collection de documents, Veille Stratégique Scientifique & Technologique Systèmes d'information élaborée, Bibliométrie, Toulouse, Novembre 2004.

[7] A. Maedche and S. Staab. Mining ontologies from text. In Proceedings of EKAW-2000, Springer Lecture Notes in Artificial Intelligence (LNAI-1937), Juan-Les-Pins, France, 2000. Springer, 2000.

[8] A. Miles, D. Brichley, SKOS Core GuideW3C Working Draft 10 May 2005, http://www.w3.org/TR/swbp-skos-core-guide/

[9] D. L McGuinness, F. Van Harmelen, OWL Web Ontology Language Overview, W3C Recommendation http://www.w3.org/TR/owl-features/, 10 février 2004.

[10] E. Morin, C. Jacquemin, Projecting Corpus-Based Semantic Links on a Thésaurus, Dans 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Maryland, USA, Juin 1999.

[11] Sang Ok Koo, Soo Yeon Lim, Sang Jo Lee Building an Ontology Based on Hub Words for Information Retrieval IEEE/WIC International Conference on Web Intelligence (WI'03) ,October 13 - 17, 2003 Halifax, Canada.

[12] M. Sanderson, W.B Croft., Deriving concept hierarchies from text, in Proceedings of the 22nd annual conference ACM SIGIR, pp 206-213, 1999.

[13] D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer and S. Katz, Reengineering Thesauri for New Applications: the AGROVOC Example, Journal of Digital Information, Volume 4 Issue 4, Article N° 257, 2004

[14] D. Tudhope, H. Alani, C. Jones, Augmenting Thésaurus Relationships: Possibilities for Retrieval, Journal of Digital Information, Volume 1 Issue 8, Article No. 41, 2001.

[15] P. Velardi, P. Fabriani, M. Missikoff: Using text processing techniques to automatically enrich a domain ontology, FOIS, pp 270-284, 2001:

[16] B. Wielinga, G. Schreiber, J. Wielemaker, and J. A. C. Sandberg. From thésaurus to ontology. Internation Conference on Knowledge Capture, Victoria, Canada, Octobre 2001.

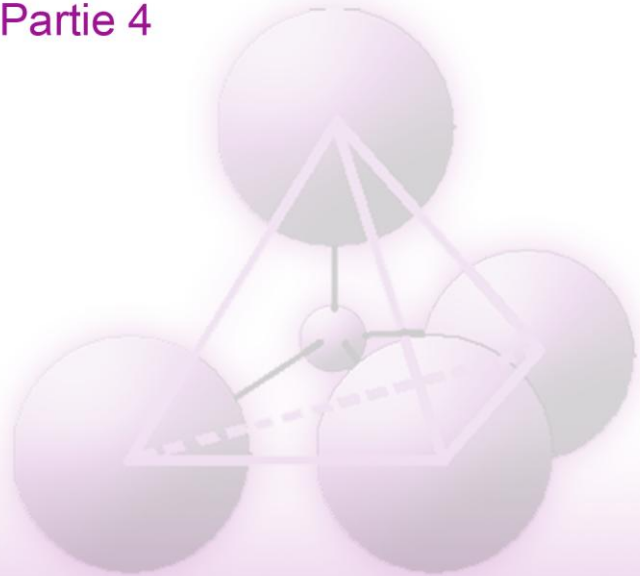
αMETIST

Partie 1

Partie 2

Partie 3 :
Coups de flash

Partie 4



Développement de la veille à l'INRS : approches et retours d'expériences

Françoise Grandjean (1)
francoise.grandjean@inrs.fr

Guillaume Moureaux (2)
guillaume.moureaux@inrs.fr

Michel Servais (3)
michel.servais@inrs.fr

(1) INRS, avenue de Bourgogne, 54500 Vandoeuvre 03 83 50 21 56
(2) INRS, avenue de Bourgogne, 54500 Vandoeuvre 03 83 50 20 00
(3) INRS, avenue de Bourgogne, 54500 Vandoeuvre 03 83 50 21 34

Mots-clés : institut recherche, sécurité travail, veille, retour expérience, analyse information, système information, logiciel, prototype, description système, INRS, France, Dilib

Keywords : research institute, work safety, wakefulness, experience feedback, information system, software, system description

Résumé : Depuis 1995, le centre de documentation du site de l'INRS (Institut National de Recherche et de Sécurité) situé à Vandoeuvre, a expérimenté différentes approches de l'infométrie et de la veille dans le but de soutenir les activités de recherches de son site. Cet article présente le cheminement, les réflexions et les difficultés qui ont marqué ce parcours.

Ce parcours a permis d'acquérir de l'expérience et de susciter un intérêt pour la veille. Des projets pilotes ont été menés sur des sujets d'études tels que le stress, la génomique du mésothéliome, les risques biologiques émergeants ou les particules ultrafines. Ils ont été réalisés en collaboration avec le LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications) et l'INIST (INstitut de l'Information Scientifique et Technique) et se sont essentiellement appuyés sur la plate-forme DILIB (Digital Library) développée par ces organismes. Parallèlement, une sensibilisation à la veille a été conduite grâce à un séminaire de réflexion interne, des

stages de formation et la création d'une rubrique de communication interne dédiée à ces sujets sur l'intranet de l'INRS.

Aujourd'hui, la nécessité d'organiser et de structurer une activité de veille au sein de l'INRS s'impose en raison de la quantité accrue d'information disponible, du développement des NTIC (Nouvelles Technologies de l'Information et de la Communication) mais aussi des choix stratégiques dans l'orientation des études et des actions menées en matière de prévention des risques professionnels.

Notre démarche a permis de mettre en évidence la nécessité d'homogénéiser l'accès à l'information et de structurer les informations réunies selon un modèle commun pour une exploitation collective efficace. Mais les aspects humains et relationnels se révèlent problématiques car il s'agit de convaincre et de dépasser les réticences que soulève la mise en place de nouveaux processus.

Introduction

Cet article vise à montrer comment les méthodologies de veille ont fait leur chemin progressivement au sein de l'INRS (Institut National de Recherche et de Sécurité). Il s'agit pour nous de partager notre approche, nos expériences, nos difficultés et nos observations.

L'objectif de cet article est aussi de montrer comment nous avons acquis précocement une connaissance profonde de la discipline et avons préparé la culture d'entreprise afin de permettre l'introduction de la veille à l'INRS non pas comme une révolution mais plutôt en douceur et comme une évidente nécessité.

Ce retour d'expériences mettra en évidence une démarche d'abord de type expérimental avec le développement d'outils d'exploration et d'analyse de l'information faisant appel aux techniques de l'infométrie. Il montrera comment nous avons parallèlement travaillé à une nécessaire sensibilisation des personnels à tous les niveaux de responsabilité de l'institut, du chercheur au directeur en passant par les chefs de projet.

Il montrera également comment nous avons réalisé des prototypes d'outils de collecte, d'analyse et de diffusion et d'analyse de

l'information aussi bien dans le cadre d'un système informatif opérationnel que d'un système informatif de management et de décision. Et nous évoquerons les travaux qui sont aujourd'hui en perspectives ainsi que les difficultés et obstacles que nous avons aujourd'hui à franchir.

5 Importance de l'information dans les missions de l'INRS

Que ce soit à l'échelle des individus ou des sociétés, aujourd'hui, chacun doit être compétitif. De l'invention à la production de masse tout doit aller plus vite dans la course à l'innovation. Cependant, il s'agit de ne pas oublier de préserver les hommes et les femmes acteurs de cette compétition.

C'est le rôle du système français de prévention des risques professionnels. Ce rôle consiste à protéger les individus face aux risques qui apparaissent notamment sous l'effet de cette pression économique. Mieux encore, les acteurs de ce système se doivent, autant que possible, de précéder cette course en proposant des solutions de prévention.

Au sein de ce dispositif l'INRS est dépositaire des connaissances scientifiques et techniques dont le système de prévention a besoin pour mettre en œuvre cette prévention. Le rôle de cet institut se décline en trois missions majeures :

Anticiper : Du risque toxique au bien-être physique et psychologique, l'INRS conduit des programmes d'études et recherches pour améliorer la santé et la sécurité de l'homme au travail. Le bilan de ces actions lui permet également de déterminer les besoins futurs en prévention. Tous les cinq ans, un programme définit son cadre général d'action.

Sensibiliser : L'institut conçoit de nombreux produits d'information : 4 revues, 300 brochures, 150 affiches, 70 vidéos, des cédéroms, un site internet. Ils sont diffusés auprès d'un large public, composé de chargés de sécurité, médecins du travail, ingénieurs, opérateurs,

formateurs... Certaines actions ponctuelles font l'objet de campagnes de prévention auprès du grand public.

Accompagner : L'INRS propose une aide technique aux entreprises : 40 000 demandeurs y font appel chaque année pour résoudre un problème de prévention. L'institut transmet son savoir-faire et ses compétences par 70 offres de formation ou d'aides pédagogiques adaptées aux besoins des animateurs de la prévention en entreprise. Ses experts participent à de nombreux groupes de travail, nationaux, européens ou internationaux, pour la rédaction de textes à caractère réglementaire ou normatif.

L'INRS est réparti sur trois sites, à Paris, à Vandoeuvre et à Neuves-Maisons, chacun disposant d'un centre de documentation adapté à ses activités. Sur le site de l'INRS situé à Vandoeuvre, la documentation traditionnelle est depuis toujours étroitement liée aux travaux d'études et de recherches qui constituent l'essentiel de l'activité scientifique. Elle apporte sa contribution dès l'origine de ces projets et les accompagne tout au long de leur réalisation.

Les trois services de documentation de l'INRS mettent à disposition sur le réseau interne l'ensemble des sources documentaires dont ils disposent. En effet, les 3 centres de documentation de l'INRS bien que de nature et de missions différentes, se sont regroupés pour offrir un accès commun à l'information scientifique et technique sur le site intranet de l'institut (Interligne) (voir illustration 1).

Illustration 1 : Interligne : Le site intranet de l'INRS.

Interligne permet d'accéder à des services de bases de données internes mais aussi externes telles que BiblioSciences de l'INIST, (Pascal, Francis, Inspec, Current Contents…), Kompass, Perinorm. Interligne ouvre également l'accès vers un serveur de cédérom offrant entre autres, l'accès aux bases diffusées par le centre canadien d'hygiène et de sécurité (CCHST).

L'utilisateur d'Interligne a aussi accès aux revues auxquelles les centres INRS de Paris, Vandoeuvre et Neuves Maisons sont abonnés. Elles sont présentées accompagnées de l'état de leur collection, du nom de la personne à contacter pour y accéder en version papier et d'un lien direct sur le site de l'éditeur pour l'accès aux sommaires voire au texte intégral lorsque la revue existe également au format électronique.

En effet depuis 2 ans, l'accès au texte intégral des articles de périodiques s'est beaucoup développé. Après une période expérimentale, où l'accès à la version électronique d'une revue était compris dans l'abonnement à la version papier, une politique d'accès payant s'est mise en place progressivement. Il a donc fallu trouver une solution pour faire face à cette tendance.

C'est dans ce but, que fin 2002, l'INRS a adhéré au consortium COUPERIN qui regroupe la plupart des universités et des organismes de recherche publics français. Ainsi un accord a pu être signé avec l'éditeur Elsevier pour accéder aux 1700 revues électroniques en texte intégral du service ScienceDirect

6 Des approches pour se familiariser avec l'infométrie et la veille

Parallèlement à cette documentation en évolution, depuis 1995, des expériences ont été menées sur le développement d'applications de traitement et d'analyse de l'information grâce à des collaborations informelles établies avec le LORIA et l'INIST. En effet, Jacques Ducloy et ses collaborateurs au LORIA puis à l'INIST, ont développé la plate-forme d'investigation documentaire, DILIB, qui est une bibliothèque de puissantes fonctions de traitement de l'information structurée au format XML. Elle a été implantée sur le serveur du Centre de Services Informatiques de l'INRS et a été utilisée pour la réalisation de nos différents outils de traitement de l'information.

L'objectif initial était de développer un outil muni d'une interface hypertexte permettant un accès intuitif à des fonds documentaires. L'outil a été utilisé pour exploiter les fonds documentaires généraux des centres INRS de Paris et de Lorraine, ainsi que des bases documentaires personnelles de chercheurs de l'institut. Il permettait de visualiser les fréquences et les associations de mots du titre, mots du résumé, descripteurs et auteurs mais aussi de naviguer dans les notices bibliographiques. Au-delà de l'objectif initial, ces interfaces de consultations étaient en fait de véritables serveurs infométriques munis également de représentations graphiques de l'information permettant d'explorer des fonds documentaires avec une optique d'analyse.

Au début des années 2000, cette expérience a été complétée lorsque ces outils ont été mis en œuvre à la demande de chercheurs pour explorer non plus des bases internes mais des fonds résultant de l'interrogation de bases de données externes comme MEDLINE,

NIOSH, PsycINFFO sur des thématiques intéressant l'institut. Ces travaux en collaboration avec le LORIA, ont notamment abouti à la réalisation des applications WebStress qui avait pour objectif d'explorer le vaste fonds des publications concernant les problèmes de stress au travail et Transcriptome/Bibliome (voir illustration 2) en collaboration avec l'INIST qui exploitait des documents traitant de l'expression génétique du mésothéliome, tumeur liée à l'exposition à l'amiante.

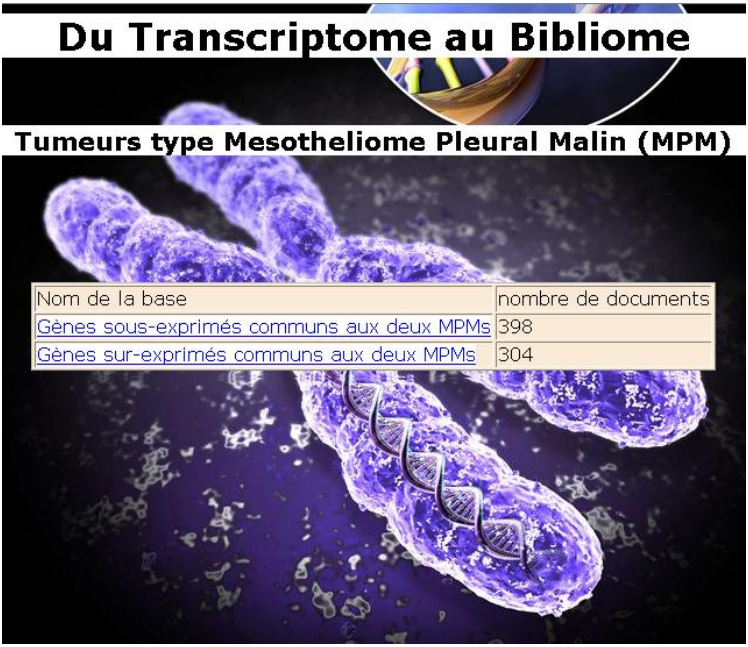


Illustration 2 : Serveur documentaire sur le mesotheliome.

Cependant, après ces premières expériences et avant d'aller plus loin dans le domaine de l'analyse de l'information, l'idée est apparue que l'INRS devait mener une réflexion sur ce qui précède cette analyse : les problèmes de collecte de l'information et donc la veille. De plus, les participants aux premières expériences ayant été enthousiasmés, l'idée était aussi d'organiser une opération visant à sensibiliser l'ensemble du personnel de l'INRS à la veille.

Septembre 2003 a donc vu l'organisation d'un séminaire interne concernant la veille. Une journée de sensibilisation permet alors de faire avancer la réflexion interne du public INRS en situant la veille dans un cadre de prospective stratégique. Tables rondes, retours d'expérience d'organismes externes et expériences internes se déroulent au long de cette journée. Les différentes déclinaisons de la veille sont proposées : veille pour une production de connaissances, veille pour des décisions de société, aspects politico-stratégiques et médiatiques, structures et outils internes.

Parallèlement, la documentation développe alors, sur son site intranet, une nouvelle rubrique consacrée à la veille grâce au concours d'une étudiante (Christelle Martin), du DESS Information Scientifique et Technique et Intelligence Economique (ISTIE) de Nancy. Sa mission était de réaliser une étude comparative des outils existants et de les appliquer à titre d'exemple à la thématique des risques biologiques émergents et plus particulièrement aux risques biologiques dans les métiers du bois. Le site en question répertorie et décrit sources d'informations, méthodes et outils pour la veille, le but étant de mieux appréhender les spécificités de chaque outil et d'aider tout un chacun à les mettre en œuvre pour ses propres besoins. En 2003 et 2004 cette rubrique intranet est accompagnée de stages de formation interne sur le thème de la veille animés par l'INIST.

Plus récemment, nous avons pu étudier et tester un nouveau produit d'analyse de fonds documentaires, présenté par l'équipe Orpailleur du LORIA (Emmanuel Nauer). Cet outil nommé IntoBib est issu de la technologie DILIB mais complété par les technologies PHP et SQL. Le but est de fournir, au chercheur ou au spécialiste de l'information scientifique et technique, un environnement dans lequel il puisse exploiter les données issues de sa veille, de façon dynamique cette fois, contrairement aux serveurs d'investigation classiques dont les explorations sont prévues dès la construction du serveur.

```
<ref>
  <TITR>Joining the trek with Keith up the Serpentine
Road--the lattice from another perspective.</TITR>
  <AUTE>
    <e>Wolosewick, J J</e>
  </AUTE>
  <SOUR>Biol-Cell. 2002 Dec; 94(9): 557-9</SOUR>
  <JOUR>Biology of the cell under the auspices of the
European Cell Biology Organization</JOUR>
  <ISSN>0248-4900</ISSN>
  <YEAR>2002</YEAR>
  <LANG>
    <e>English</e>
  </LANG>
  <PAYS>France</PAYS>
  <DEEN>
    <e>Cytoplasm chemistry</e>
    <e>Cytoskeleton chemistry</e>
    <e>Organelles chemistry</e>
    <e>Cytoplasm ultrastructure</e>
    <e>Cytoskeletal Proteins analysis</e>
    <e>Cytoskeletal Proteins ultrastructure</e>
    <e>Cytoskeleton ultrastructure</e>
    <e>Microscopy, Electron methods</e>
    <e>Organelles ultrastructure</e>
  </DEEN>
  <TYPE>
    <e>Editorial</e>
  </TYPE>
</ref>
```

Illustration 3 : Données au format XML pour l'application amiante.

Des fonctionnalités de fouille (dénombrements, classifications, extractions de règles, etc.) peuvent être déclenchées à la demande pour analyser plus précisément certains sous-ensembles de données. Le principe technique est que les actions de l'analyste sur l'interface hypertexte sont traduites en requêtes SQL et les résultats de traitement traduits en graphiques et en chiffres. Les temps de traitement sont très courts grâce à la conception du serveur qui pré calcule les résultats lors de sa génération à l'aide de fonctions PHP.

A titre d'essai, nous avons expérimenté l'outil avec un corpus de références bibliographiques sur l'amiante issu de MEDLINE. Là encore le savoir-faire acquis au sein de l'INRS a permis la transformation de ces données au format XML (voir illustration 3) pour l'importation directe dans cette application. Les résultats qui se sont dégagés se sont avérés intéressants dans la mesure où ils indiquaient des tendances en présentant des pics d'intérêt et de publications pour le sujet, liés aux dates des décisions politiques ou aux échos médiatiques.

7 Des prototypes de systèmes informatifs

En 2004 une étude d'instruction de projet sur les risques professionnels liés aux particules ultrafines est initiée par un laboratoire de l'INRS. Le chargé de projet (Olivier Witschger) dont la mission est de faire le point sur ces risques et sur l'intérêt de lancer une étude sur ce sujet, est intéressé par la mise en place d'un processus de veille pour soutenir l'étude. Sensibilisé par le précédent séminaire de veille il avait constaté que la documentation traditionnelle et les outils mis à sa disposition ne suffisaient pas pour répondre à son besoin. Nous décidons alors d'avoir recours aux services d'un veilleur, étudiant du DESS ISTIE (Guillaume Moureaux) dans le cadre de son stage d'étude en collaboration avec les membres de la cellule veille de l'INIST (Catherine Czysz, François Parmentier, Philippe Houdry et Solveig Vidal) qui lui transmettent leur savoir faire et nous donnent accès à certains de leurs moyens d'investigation.

La première tâche consiste alors à réunir un fonds de références bibliographiques sur le sujet des particules ultrafines. Pour cela les bases documentaires mises à disposition par l'INIST au travers du service BiblioSciences sont employées. Ce service présente l'intérêt de permettre l'interrogation de 11 bases de données externes avec une seule interface d'interrogation et de permettre le téléchargement des résultats sous un format unique. Un fonds documentaire sera donc effectivement constitué après plusieurs réunions avec les chercheurs impliqués dans le projet pour définir le besoin et préciser le vocabulaire requis pour l'interrogation des sources.

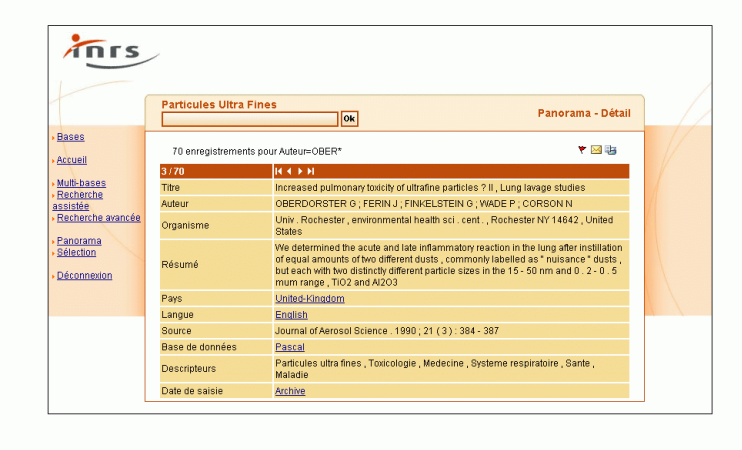


Illustration 4 : La base de donnée documentaire particules ultrafines par l'interface CinDoc Web

Mais à cette étape on n'a encore qu'un résultat brut puisqu'il reste encore à traiter le fonds collecté afin de le diffuser et de l'exploiter. Une rapide analyse de l'existant permet de faire le point sur les outils logiciels disponibles à l'INRS et utilisables à cet effet. Le choix est fait de créer une base de notices bibliographiques à l'aide du logiciel documentaire CINDOC récemment acquis par l'INRS. Cette base sera accessible grâce à l'interface Web de CINDOC (voir illustration 4) depuis les trois centres de l'INRS constituant ainsi un moyen de diffusion idéal. De plus, le logiciel est muni d'index interrogeables qui permettront une exploitation des données recueillies par les personnes qui devront analyser ce fonds.

Mais avant cela il s'agissait de préparer les données afin de réaliser ce produit. Les données sont dans un format propriétaire incompatible avec une importation immédiate dans une base CINDOC. Il convient donc de transformer les données dans un format approprié. Une phase de formatage et de traitements divers, effectués sous Unix à l'aide des outils de la plate-forme DILIB est donc engagée. La plate-forme DILIB est d'abord mise à jour et les outils de développement installés avec le concours du Centre de Services Informatiques de l'INRS (Michel Servais).

Le développement des programmes de formatage commence alors. Il s'agit de passer par un format XML qui servira de format transitoire. Le format propriétaire de BiblioSciences est donc transformé en format XML qui sera lui même ensuite transformé en format Ajout Piloté. L'Ajout Piloté est en effet le format d'importation des données de CINDOC qui permettra de constituer la base documentaire.

Dans le même temps, d'autres traitements intermédiaires sont ajoutés à la chaîne de formatage. En effet, à l'étape du format XML il devient possible d'appliquer des traitements supplémentaires comme le dédoublonnage et l'indexation semi-automatique des références bibliographiques.

L'objectif de ces outils complémentaires est d'effectuer un retour vers des problématiques d'analyse en se proposant d'explorer des fonds issus des bases de données externes.

Le dédoublonnage est en effet un pré requis pour pouvoir réaliser ensuite une analyse infométrique du fonds. Quant à l'indexation semi-automatique elle permet de marquer les références bibliographiques avec les termes que l'on veut analyser. Le principe étant de détecter des termes ou expressions spécifiques présents dans les notices bibliographiques et de les représenter par un terme générique qui sera ensuite analysé en termes de fréquence.

Ces travaux permettent de produire des historiques des publications sur tel ou tel sujet sous forme de représentations graphiques. Grâce à cette application on peut montrer par exemple que le nombre de publications concernant un type de risque ne progresse plus alors que l'industrie correspondante est florissante. On peut ainsi se poser la

question de l'intérêt de la proposition d'autres études dans ce domaine. Cette application constitue donc un système informatif décisionnel à destination de la direction scientifique de l'institut. Afin de tester ce système d'analyse, de tels travaux ont été réalisés à titre d'exemple sur des thématiques connues de longue date ou bien nouvelles comme les nanoparticules, les éthers de glycol, les troubles musculo-squelettiques, le stress, les fibres céramiques, le traitement au niveau de l'indexation permettant de différencier les aspects prévention, épidémiologiques et toxicologiques.

L'objectif ici est évidemment de susciter l'intérêt et de sensibiliser les directions en montrant l'apport d'une veille stratégique et prospective. Cet outil pourrait en effet contribuer à la prise de décision concernant les orientations scientifiques de l'institut dans le cadre du Plan à moyen terme quinquennal 2008-2012.

8 Perspectives

Où en sommes-nous aujourd'hui ? Actuellement, nous commençons à développer, en collaboration avec Michel Servais, un système automatisé d'alerte, basé sur les profils proposés par ScienceDirect de Elsevier. Il s'agit de construire des équations de recherche pour chacun des sujets des 7 projets transversaux en cours au sein de l'INRS. Les alertes envoyées par ScienceDirect seront redirigées vers le serveur de messagerie de l'INRS, basculées dans une base de données de type SQL et après validations sélectives des résultats, diffusées dans une rubrique « Veille PTI » d'Interligne (voir illustration 5).


Robot de Surveillance Documentaire (SERVEUR : 192.168.2.21)
139 documents sélectionné(s) au 19/04/2006

SCIENCE DIRECT Alerte(s): stress+auteurs+psychosocial

Page: 2 / 20 de la liste des 139 document(s) [début | suite | précédent | fin]

2006-03-28 **Work related post-traumatic stress as described by Jordanian emergency nurses**
Accident and Emergency Nursing, In Press, Corrected Proof, Available online 27 March 2006
Anders Jonsson and Jehad Halabi

2006-03-25 **Observational Stress Factors and Musculoskeletal Disorders in Urban Transit Operators**
Journal of Occupational Health Psychology, Volume 11, Issue 1, January 2006, Pages 38-51
Birgit A. Greiner and Niklas Krause

2006-03-25 **Relationships Among Organizational Family Support, Job Autonomy, Perceived Control, and Employee Well-Being**
Journal of Occupational Health Psychology, Volume 11, Issue 1, January 2006, Pages 100-118
Cynthia A. Thompson and David J. Prottas

2006-03-23 **Rapports au travail, contrôle et santé dans les centres de gestion de la relation–client**
Psychologie du Travail et des Organisations, In Press, Corrected Proof, Available online 22 March 2006
M. Lourel

2006-03-23 **Confirmatory factor analysis of posttraumatic stress symptoms in emergency personnel: An examination of seven alternative models**
Personality and Individual Differences, In Press, Corrected Proof, Available online 22 March 2006
Leanne Andrews, Stephen Joseph, Mark Shevlin and Nick Troop

2006-03-21 **Musculoskeletal complaints and psychosocial risk factors among physicians in mainland China**
International Journal of Industrial Ergonomics, In Press, Corrected Proof, Available online 20 March 2006
Derek R. Smith, Ning Wei, Yi-Jie Zhang and Rui-Sheng Wang

Illustration 5 : Les alertes ScienceDirect accessibles sur le site intranet

Que ressort-il du parcours effectué ? Au niveau de la Documentation, l'intérêt est évident, dans la mesure où c'est le virage qu'elle doit prendre rapidement pour évoluer vers d'autres missions.

Le rôle traditionnel de recherche et de fourniture d'informations bibliographiques a été remis en cause par l'explosion de sources documentaires électroniques à disposition des chercheurs via l'intranet. La Documentation s'est aussi attachée à les aider à les utiliser et les exploiter (formation des utilisateurs sur les sources, aide au démarrage d'une recherche bibliographique, constitution de bases de données, mise en place d'alertes..). Elle se doit de jouer désormais un rôle moteur dans l'analyse et l'exploitation des données bibliographiques dans le cadre de la veille.

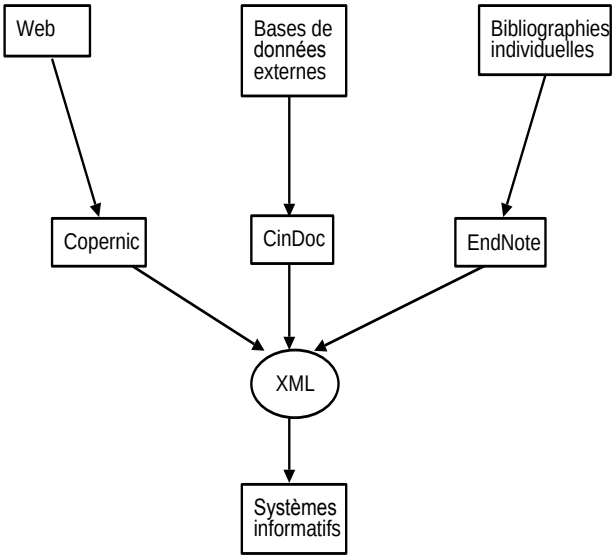


Illustration 6 : Modélisation de prototypes de systèmes informatifs avec les outils de l'INRS

Tandis que les technologies de l'information font leur progrès nous tentons autant que possible d'apporter des innovations dans notre manière de pratiquer la gestion de l'information à l'INRS. Nous suivons pour cela une démarche expérimentale qui nous permet de nous former aux méthodes et aux technologies qui apparaissent. Cette acquisition d'expérience nous permet d'aller chaque fois plus loin et de rendre l'INRS progressivement autonome dans sa gestion de l'information sur la base des expériences et savoir-faire acquis dans le cadre de collaborations.

Par exemple, nous avons commencé à acquérir une certaine pratique des technologies XML. Cette maîtrise nous permet aujourd'hui et de plus en plus, d'échanger des données entres différentes applications. (voir illustration 6) Dans ce domaine nous avons notamment développé des passerelles logicielles faisant appel à la technologie XML. Elles permettent dans un cas de regrouper des bibliographies

constituées sous EndNote en une base documentaire CINDOC. Cela constitue une sorte d'outil de veille collective rassemblant les recherches bibliographiques des membres d'un groupe de travail donné. Nous avons également constitué des outils destinés à monter un système de veille orienté web.

Ces expériences nous ont permis d'apprendre à réaliser différents prototypes à toutes les étapes du processus de veille et de tester sur des groupes d'utilisateurs les solutions les plus appropriées à la culture d'entreprise de l'INRS. De même, nous aurions pu faire le choix d'investir dans des solutions commerciales clés en main. Mais nous avons préféré faire d'abord nos expériences afin de pouvoir le moment venu être capables d'une véritable maîtrise des outils et de l'exploitation de ces outils à leur plein potentiel.

Dans le même temps, nous nous sommes employés à faire progresser l'idée qu'une gestion moderne de l'information à l'INRS qui pourrait profiter non seulement à chacun, mais encore plus à l'ensemble et qu'il devient nécessaire d'avoir une politique globale dans ce domaine. De plus, chacun commence à réaliser par lui même qu'il ne peut pas maîtriser à lui seul la masse d'information qui se déverse sur le web, dans l'intranet ou bien envahit sa boite de messagerie et son bureau.

En fait de société de l'information, nous serions plutôt actuellement dans une société de la surinformation ou chacun a la responsabilité ou parfois le besoin vital de ne rien ignorer mais n'a pas le temps matériel de traiter la quantité d'information qui lui incombe. C'est ainsi que certains chercheurs de l'INRS cherchent à monter des groupes de travail non pas pour mettre en commun l'information mais pour en partager l'analyse dans des processus d'intelligence collective. Car le problème n'est plus de trouver l'information, mais de savoir laquelle est l'information adéquate et pertinente.

Si pendant un temps, à l'INRS, la tendance a consisté à donner à l'utilisateur tous les moyens de s'informer par lui-même, aujourd'hui, cette situation tend à s'inverser. Des chercheurs commencent en effet à revenir chercher de l'aide auprès des spécialistes pour trouver l'information ou plutôt pour la trier puis pour l'analyser lorsqu'elle est trop abondante. Cependant l'utilisateur conserve le désir d'être autonome et a parfois du mal à admettre que

des médiateurs sont nécessaires, à l'interface, entre lui et l'information.

Pour ce qui concerne l'avenir de l'INRS, 2005 voit se définir des projets à l'échelle de chaque département scientifique mais aussi des projets transversaux rassemblant les moyens de plusieurs départements qui ont des besoins croissants de collecte, de tri, de validation, d'analyse et de diffusion de l'information. Pour 2005-2008, l'INRS a été retenu comme maître d'œuvre du projet européen d'observatoire des risques professionnels dont le principe même consiste à surveiller la littérature scientifique et toutes les autres sources d'information.

Cette offre nouvelle en matière d'outils et de méthodes, va pouvoir aider à faire des choix stratégiques au moment où se met en place la préparation du prochain Plan à Moyen Terme et sans doute permettre à chacun de mieux repérer et exploiter l'information utile pour les missions de l'INRS.

Bibliographie

[10] Rihn b., Mohr s., Grandjean f., Nemurat c.,From transcriptomics to bibliomics.,Medical sciences monitor, vol.9, no.8, 2003, pp.89-95.

[11] Grandjean f., Mur d., Puzin m., Ciccotelli j., Falcy m.,Séminaire interne "La veille", Institut National de Recherche et de Sécurité, INRS, Vandoeuvre, 18 septembre 2003.,Vandoeuvre, Institut National de Recherche et de Sécurité, INRS, 2003, 47p.

[12] Jolibois s., Mouze-Amady m., Chouaniere d., Grandjean f., Ducloy j.,WEBSTRESS : a web-interface to explore a multidatabase bibliographic corpus on occupational stress,Work and stress, vol. 14, no 4, octobre 2000 pp.283-296.

[13] Jolibois s., Chouaniere d., Ducloy j., Grandjean f., Mouze-Amady m.,Un exemple d'utilisation de l'ULMS, base multilingue de connaissances biomédicales.,Documentaliste, vol.37, no.2, juin 2000, PP.94-103.

[14] Jolibois s., Nauer e., Chouaniere d., Mouze-Amady m., Ducloy j., Grandjean f.,Standardisation of a multidatabase bibliographic corpus.,Consensus Workshop on "stress at work" organise par l'AMI (UK),

Copenhague, 21-22 juin 1999,Vandoeuvre, Institut National de Recherche et de Sécurité, INRS, Service Epidémiologie en Entreprise, EE, 1999

[15] Ducloy j., Nauer e., Jolibois s., Grandjean f., Chouaniere d., Mouze-Amady m.,DILIB, how to use an XML technology to build Intranet or Internet services oriented towards scientific survey,Consensus Workshop on "stress at work" organise par l'AMI (UK), Copenhague, 21-22 juin 1999,Vandoeuvre, Institut National de Recherche et de Sécurité, INRS, Service Epidémiologie en Entreprise, EE, 1999

[16] Jolibois s., Chouaniere d., Ducloy j., Grandjean f., Mouze-Amady m.,La gestion informatisée de corpus bibliographiques. Adaptation des normes et formats documentaires.,Bulletin des bibliothèques de France, vol. 45, no. 1, 2000, pp. 998-108.

[17] Jolibois s., Chouaniere d., Mouze-Amady m., Grandjean f., Servais m.,Standardisation of a multibase bibliographic corpus.,Journal of the American Society for Information Service,Vandoeuvre, Institut National de Recherche et de Sécurité, INRS, Service Epidémiologie en Entreprise, EE, 1999

[18] Chouaniere d., Jolibois s. ; Mouze-Amady m., Grandjean f., Francois m.,Une base documentaire sur le stress professionnel.,Travail et sécurité, n° 579, décembre 1998, pp. 8-11, ill.

INCISO :

Elaboration automatique d'un index de citations des revues espagnoles en sciences sociales.

José M. Barrueco (1)
Jose.Barrueco@uv.es

Julia Osca-Lluch (2)
m.julia.osca@uv.es

Thomas Krichel (3)
krichel@openlib.org

Pedro Blesa (4)
pblesa@dsic.upv.es

Elena Velasco (2)
elenavelascoarroyo@yahoo.es

Leonardo Salom (2)
leosamu@eui.upv.es

(1) Biblioteca de Ciencias Sociales, Universidad de Valencia, 46022 Valencia
(2) Instituto de Historia de la Ciencia y Documentación López Piñero (Universidad de Valencia-CSIC) 46010 Valencia
(3)Palmer School, 720 Northern Boulevard, Brookville 11548-1300, USA
(4) Dept. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 46022 Valencia

Cet article est une traduction de Clotilde Roussel (INIST-CNRS) et Magali Rasolomanana (INIST-CNRS). Il a été révisé par Catherine GUNET (INIST-CNRS) et mis en ligne par l'équipe ARTIST.

Il est paru sous la référence originale :

INCISO: ***Automatic Elaboration of a Citation Index in Social Science Spanish Journals***

Mots-clés : bibliométrie, recherche scientifique, littérature scientifique, périodique électronique, sciences sociales, évaluation, projet, index citation, Espagne, INCISO (Indice de Ciencas Sociales)**,** logiciel, description système, architecture système

Keywords : bibliometrics, scientific research, scientific literature, electronic periodical, social sciences, evaluation, project, citation index, spain, software, system description, system architecture.

Résumé : Les index de citations sont des outils clés dans le système de communication scientifique pour deux raisons. Tout d'abord, ce sont une excellente source d'informations pour interroger la littérature scientifique puisqu'ils permettent de naviguer au moyen de liens entre les documents représentés par des références bibliographiques. Ensuite, ils permettent d'évaluer la production scientifique. Le comptage des citations est un processus courant pour évaluer la qualité d'un article scientifique. En Espagne, une telle évaluation n'est possible qu'en utilisant des outils élaborés par l'ISI mais dont la couverture des revues publiées hors des pays anglo-saxons est limitée. L'évaluation de la production scientifique espagnole se limite donc aux travaux publiés dans des revues internationales. Il n'existe aucun outil pour évaluer la recherche (notamment en Sciences humaines et sociales) publiée dans les revues locales. Dans le cadre du projet de recherche INCISO, nous étudierons la possibilité de créer automatiquement un index de citations. L'objectif du projet est de développer un logiciel permettant de générer automatiquement des index de citations et de créer un échantillon d'index de citations pour les Sciences sociales.

Cette recherche est financée grâce à la subvention HUM2004-05532 du ministère espagnol des Sciences et de l'Education.

Introduction

Les revues scientifiques sont le principal moyen pour la communauté scientifique de communiquer les résultats de ses recherches. Le facteur d'impact des revues scientifiques est devenu un outil clé pour évaluer non seulement la diffusion et la visibilité de ces revues, mais aussi l'importance et la qualité de la recherche scientifique. Pour

calculer le facteur d'impact des revues dans une discipline donnée, il est nécessaire de constituer des bases de données bibliographiques dans lesquelles tous les travaux publiés dans les revues les plus importantes du domaine seront répertoriés. De plus, il faudra que ces bases de données contiennent des informations sur les références bibliographiques de chaque article afin de pouvoir établir des liens entre l'article citant et l'article cité. Enfin, le système doit pouvoir compter le nombre de fois où un article est cité. De telles bases de données s'appellent des index de citations. Actuellement, l'Institute for Scientific Information (ISI) publie trois index couvrant toutes les disciplines (Science Citation Index, Social Science Citation Index et Arts and Humanities Citation Index). Les données de ces index sont utilisées pour évaluer la recherche dans les universités du monde entier.

Les coûts élevés et la grande complexité technique qu'entraine la création d'index de citations ont freiné jusqu'ici, le développement de nouvelles bases de données qui pourraient être utilisées en complément des produits de l'ISI. Dans le cas des pays non-anglophones, un tel complément serait utile car l'ISI ne traite que des revues internationales en grande majorité de langue anglaise.

En 1983, Garfield a signalé que le facteur d'impact servait avant tout à la littérature anglo-saxonne et que donc toute évaluation fondée sur ce facteur d'impact n'était valable qu'au sein de cette communauté anglo-saxonne. Différents auteurs ont analysé les données du SCI et les ont comparées à la production scientifique des pays non anglo-saxons ; ils ont constaté que la discrimination envers ces pays était évidente. On peut observer ce problème dans les sciences dures et les technologies, mais il est encore plus marqué dans les sciences humaines et sociales, car dans ces domaines, les chercheurs publient souvent dans des revues nationales ou régionales car ces dernières sont davantage en rapport avec la portée locale de leur recherche. La recherche publiée dans les revues locales n'est donc pas couverte par l'ISI et ne peut être évaluée.

Le présent projet n'est pas le premier à vouloir développer des index de citations en Espagne ; il y a eu plusieurs tentatives depuis les années 1990 et notamment « l'Index de citations et les index bibliométriques des revues espagnoles en médecine interne et ses

spécialités » (Terrada et coll., 1991), « l'Index de citations de la documentation en espagnol » (Moya et coll., 1998), « l'Index de citations des revues espagnoles en sciences humaines » (Sanz et coll., 1998), « l'Index de citations des revues espagnoles en psychologie » (Tortosa et coll., 2002), « l'Index de citations des sciences économiques et des affaires » (Hernández et coll., 2003) et plus récemment « l'Index des sciences sociales » (Jiménez-Contreras et coll., 2004). Tous sont focalisés dans le cadre géographique spécifique de l'Espagne sur un secteur scientifique spécifique, avec une couverture temporelle délimitée.

La plupart des projets que nous avons cités ont malheureusement disparu par manque de financement mais tous partageaient les mêmes caractéristiques :

- ils étaient basés sur un enregistrement manuel des références et des citations,
- ils se concentraient sur une discipline concrète,
- ils utilisaient un échantillon réduit de revues (4-5 dans certains cas) et saisissaient aussi peu d'information que possible afin de réduire la charge de travail des opérateurs de saisie.

Notre conclusion est que la constitution d'index de citations généraux selon des moyens traditionnels demande des ressources beaucoup trop coûteuses pour généraliser de tels index au niveau national. Autrefois, seul l'ISI disposait des ressources nécessaires pour élaborer des index de revues papier. Toutefois, de nouvelles voies se sont ouvertes avec la généralisation de l'internet comme nouveau moyen de communication, avec la prolifération des revues électroniques au niveau national comme au niveau international et avec la possibilité de créer des index par des moyens automatiques. Si les articles étaient disponibles dans des formats numériques, un système informatique pourrait alors en extraire les références automatiquement. Avec un tel système les coûts seraient nettement réduits et de nouveaux index couvrant de nouveaux types de documents (par exemple la littérature grise) pourraient voir le jour.

Pour essayer de développer plus avant cette idée et en nous basant sur les travaux des auteurs décrits plus loin, nous avons décidé d'étudier la possibilité de développer un système informatique

capable de créer automatiquement des index de citations pour les publications espagnoles. Notre projet a obtenu une subvention de recherche de trois ans de la part du ministère espagnol de la science et de la technologie. Cette subvention a débuté en juillet 2005 et nous avons appelé le projet INCISO (Indice de Ciencias Sociales ou Index des Sciences sociales). INCISO avait pour but général de réduire les coûts du processus en remplaçant l'homme par un système informatique capable de constituer automatiquement un index de revues électroniques. Le projet avait deux objectifs principaux :

> 1) concevoir un système informatique pour élaborer un index de citations de manière automatisée. Le système pourra s'appliquer à des disciplines multiples mais il sera testé sur une sélection de revues espagnoles en sciences sociales.
>
> 2) élaborer et diffuser un index de citations pour les sciences sociales basé sur une sélection de revues espagnoles. Cet index sera mis à disposition de toute la communauté scientifique et sera librement accessible sur le site web du projet : http://inciso.openlib.org/.

La suite de cet article est organisée comme suit. La deuxième partie décrit certains autres projets de recherche de niveau international qui travaillent sur l'extraction automatique et les liens entre les références dans le but de constituer des index de citations. Dans la troisième partie, nous abordons la méthodologie et les étapes de notre projet. L'architecture d'INCISO est discutée dans la quatrième partie. La cinquième partie décrit l'état d'avancement du projet et conclut l'article.

1 Autres travaux sur le sujet

La généralisation des formats électroniques d'édition et de distribution des articles scientifiques a permis de développer de nouvelles fonctionnalités de recherche documentaire comme par exemple les recherches en texte intégral, les liens entre les références bibliographiques ou les index de citations autonomes. Notre projet

s'inscrit dans ce dernier aspect. Roth (2005) décrit plusieurs autres projets de recherche relatifs au développement des index de citations qui pourraient éventuellement concurrencer Science Citation Index (SCI). De la liste de Roth nous retenons deux groupes de projets : les projets commerciaux et les projets académiques.

Les projets commerciaux sont généralement menés par les sociétés d'édition afin de développer et d'améliorer leurs offres de services bibliographiques. D'un point de vue technique, ils utilisent les informations issues de documents et de références déjà disponibles dans les bases de données des éditeurs. De telles informations ont été générées au cours du processus éditorial et elles sont en général au format SGML ou XML. La grande qualité des données permet de développer des services à forte valeur ajoutée. Les défis techniques que ces projets doivent relever sont l'établissement de liens entre les références et le texte intégral sur différentes plates-formes et la gestion des droits d'accès aux documents. A titre d'exemple :

- Chemical Abstract propose une interrogation des références citées remontant à 1997. Chaque notice est liée à d'autres notices qui le citent correctement grâce à deux fonctionnalités : « Get related » (obtenir les références apparentées) et « Get citing references » (obtenir les références citantes), cette dernière fonctionnalité permettant aux utilisateurs de connaître le nombre de fois où un article a été cité. Voir http://www.cas.org/casdb.html,
- Scopus est généralement considéré comme un éventuel concurrent du SCI puisqu'il fournit des résultats de recherche qui incluent des résumés, des références citées et des liens vers les références citées (Roth 2005). Scopus est une initiative d'Elsevier, le géant de l'édition STM qui a récemment ajouté 13 millions de notices de brevets. Voir http://www.scopus.com.,
- CrossRef. est un service collaboratif d'établissement de liens entre les références qui fonctionne un peu comme un standard numérique. Il ne contient aucun contenu en texte intégral, mais plutôt des liens vers celui-ci grâce aux identifiants d'objets numériques (Digital Object Identifier-DOI) associés aux métadonnées des articles fournies par les éditeurs participants à CrossRef. Le résultat final est un

système de liens souple grâce auquel un chercheur peut cliquer sur une citation dans une revue et accéder à l'article cité. CrossRef a démarré en 2000 quand un groupement des principaux éditeurs scientifiques a fondé la Publishers International Linking Association, Inc (PILA) qui gère Crossref. Voir http://www.crossref.org.

Les projets académiques travaillent sur toute sorte de types de documents disponibles sur l'internet. Ils analysent les documents en texte intégral afin d'en extraire les références et les citations qui seront reliées aux documents d'origine s'ils sont disponibles sous forme électronique. Dans ce cas les données sont extraites automatiquement par des programmes informatiques. Ces projets ne sont pas tous d'un niveau de qualité similaire mais d'une manière générale leur qualité est inférieure à celle des projets commerciaux. Le principal enjeu de ces projets consiste à améliorer le processus technique d'extraction des métadonnées les plus pertinentes.

- Citebase permet d'interroger des documents dont les références ont été analysées et d'obtenir des résultats classés par facteur d'impact. Citebase est le fruit de l'Open Citation Project, projet bénéficiant du soutien du Joint Information Systems Committee du Royaume-Uni et de la National Science Foundation des Etats-Unis.
- CiteSeer est un système logiciel à double fonction, l'extraction de citations et leur implémentation dans le dispositif informatique, permettant ainsi de produire une base de données contenant plus de 200 000 documents indexés, avec plus de deux millions de références bibliographiques. Il a été développé dans les laboratoires de recherche de NEC par Steve Lawrence, Kurt Bollacker et C. Lee Giles. Voir http://citeseer.ist.psu.edu.
- CitEc est un index de citations dans le domaine des sciences économiques référençant les documents électroniques disponibles dans la bibliothèque numérique de RePEc. CitEc utilise une version modifiée du logiciel de CiteSeer pour établir des liens entre les références de documents disponibles en libre accès (principalement des documents de travail). Pour chaque notice dans RePEc, CitEc propose les fonctionnalités suivantes : « cited by » (cité par) si le

document a été cité par d'autres documents également disponibles dans RePEc et « get references » (obtenir les références) quand les références de l'article citant sont effectivement reliées aux documents cités. Voir http://netec.ier.hit-u.ac.jp/CitEc.

- Google Scholar est une base de données de littérature scientifique contenant des articles évalués par les pairs, des thèses, des livres, des prépublications, etc. provenant d'éditeurs scientifiques, de sociétés savantes et d'archives de publications électroniques. Google Scholar analyse et extrait automatiquement les citations et les présente sous forme de résultats séparés même si les documents auxquels ces citations font référence ne sont pas disponibles en ligne. Voir http://scholar.google.com.

2 Méthodologie et déroulement du projet

Les auteurs de cet article ont une grande expérience dans le développement d'index de citations autonomes puisqu'ils ont développé le service CitEc décrit ci-dessus. Avec ce nouveau projet, il va s'agir de transposer l'expérience acquise dans une discipline spécifique à des publications dans une langue autre que l'anglais issues de plusieurs disciplines, avec comme facteur commun le même pays de publication. Une grande partie du logiciel développé pour CitEc sera utilisé et testé dans ce nouvel environnement.

La méthodologie que nous allons suivre pour extraire et relier les données concernant les références, comprend sept étapes :

1. Il nous faut sélectionner les sources de données. Le système sera testé sur un échantillon de revues espagnoles en sciences sociales. Dans un premier temps, cet échantillon est réduit à dix revues représentant toutes les disciplines. La sélection a été faite selon les critères suivants : les revues doivent disposer d'une version électronique avec au moins quatre numéros publiés et un système d'évaluation par les pairs pour s'assurer de la qualité du contenu. Etant donné

que l'index va être créé automatiquement, il est essentiel que les revues disposent d'une version électronique. Le nombre de revues électroniques en Espagne est encore faible. Néanmoins il se développe rapidement comme on peut le voir dans le répertoire des revues électroniques espagnoles en sciences humaines et sociales (disponible à : http://citas.uv.es/DifusionRevistas/Revistaselectronicas/index.html). Y figurent les nouvelles revues créées exclusivement sous forme électronique et d'autres revues qui s'orientent vers le format électronique tout en maintenant une version papier. La sélection, basée sur la disponibilité ou non du format électronique, implique que d'importantes revues seront exclues de l'échantillonnage parce qu'elles n'existent que sous forme papier. Nous sommes conscients que les meilleures revues espagnoles sont exclues et que les résultats doivent être interprétés en conséquence et ne doivent pas être utilisés pour évaluer la recherche. La situation devrait s'améliorer à l'avenir au fur et à mesure que de plus en plus de revues passent à l'électronique.

2. Il nous faut obtenir les informations bibliographiques concernant les articles publiés dans les revues sélectionnées. A l'avenir, il serait souhaitable de collaborer avec les fournisseurs d'informations (les éditeurs) afin de définir des moyens automatisés pour alimenter le système. Cela signifie qu'il faut mettre en place des procédures permettant d'alerter INCISO quand de nouveaux articles sont publiés. Pour ce faire, nous utiliserons de nouvelles technologies dans le domaine des bibliothèques électroniques telles que le protocole OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting), voir http://openarchives.org.
3. Les informations bibliographiques de chaque article ayant une adresse électronique renvoyant vers le document en texte intégral seront stockées dans une base de données mySQL. Ces documents seront considérés comme les documents citants. Un autre fichier de la base de données recueillera les métadonnées inhérentes aux documents publiés en Espagne dans le domaine des sciences sociales au cours des dix dernières années. Ces documents sont

potentiellement les documents cités. Ces métadonnées sont considérées comme faisant autorité (contrôlées) car elles proviennent de sources de qualité. Seules les citations faisant références à ces documents seront retenues et considérées comme de vraies citations. Toutes les autres citations seront écartées.

4. Pour chaque document citant, on télécharge le fichier contenant le document en texte intégral. A l'heure actuelle, INCISO ne gère que les fichiers en format pdf. Le fichier est converti au format ASCII afin que le texte puisse être facilement extrait et manipulé.
5. Une fois le fichier converti, on lance l'analyse syntaxique de l'ensemble du texte en vue d'identifier et de délimiter la section contenant les références bibliographiques. Si cette étape aboutit, il faut ensuite identifier chaque référence citée et la fractionner en fonction de ses différents éléments comme l'auteur, le titre, la revue, etc. C'est l'étape la plus importante, car l'efficacité du processus dépend essentiellement de la qualité et de la cohérence de ces résultats. Un problème majeur est que les références sont différentes selon les disciplines. La démarche adoptée par la plupart des projets décrits plus haut consiste à extraire de manière aussi précise que possible tous les éléments des références. Ces projets ont donc tenté d'analyser les notices de manière exhaustive. A notre avis, une telle analyse est compliquée et gourmande en ressources car la qualité des données source est très hétérogène. Notre approche est différente. Le système ne va identifier que les éléments de base de la référence et essaiera ensuite de localiser le document référencé dans la base de métadonnées contrôlées. S'il trouve le document, les bonnes métadonnées viendront compléter la référence.
6. Toutes les données extraites au cours des étapes précédentes sont stockées dans une base de données de références. Cette base de données servira à faire des études bibliométriques sur les résultats.
7. Le projet propose deux types de résultats. D'une part, l'index de citations qui sera utile pour évaluer la recherche en sciences sociales menée en Espagne et d'autre part, un

ensemble de documents techniques concernant le système qui sera d'un intérêt majeur pour la communauté des chercheurs dans le domaine des bibliothèques numériques. Tous les résultats seront publiés en libre accès sur le web.

INCISO va développer un système informatique pour réaliser de manière automatisée le processus décrit précédemment. La conception du système s'appuiera sur les principes fondamentaux suivants :

- la multidisciplinarité. Dans un premier temps, le système sera appliqué aux revues en sciences sociales mais reposera sur une architecture modulaire permettant d'adapter facilement de nouvelles fonctions au noyau de base afin de répondre aux besoins spécifiques des différentes disciplines.
- les logiciels libres. Le système sera complètement écrit en Perl. Les logiciels complémentaires nécessaires relèveront du logiciel libre comme par exemple ceux qui utilisent GNU ou des licences similaires. Le système fonctionnera sous DebianGNU/Linux sur une machine localisée à l'université polytechnique de Valence (Espagne) et utilisant mySQL comme système de gestion de base de données et Apache comme serveur web.
- l'autonomie et la continuité. Une des principales exigences à prendre en compte dans la conception du système est qu'il faudra que ce dernier puisse fonctionner avec le minimum de maintenance possible. Les systèmes actuels reposent sur le travail éditorial d'administrateurs ce qui nécessite des ressources financières pour les rémunérer. Si nous mettons en place un système automatisé au maximum et si nous parvenons à constituer une masse critique de documents, alors il se peut que certains éditeurs souhaitent contribuer au système en y versant leurs publications. Cela permettrait d'assurer un flux continu de documents et le système pourrait fonctionner par lui-même sous l'impulsion des éditeurs.
- l'ouverture. Les données produites seront accessibles à toute la communauté scientifique ainsi qu'à d'autres projets au niveau international. Le premier prolongement du projet pourrait concerner les revues publiées en Amérique latine.

Latindex (http://www.latindex.org) est un annuaire de revues électroniques recensées par le CINDOC qui pourrait servir à sélectionner les revues de qualité à inclure dans INCISO.

3 Architecture du système

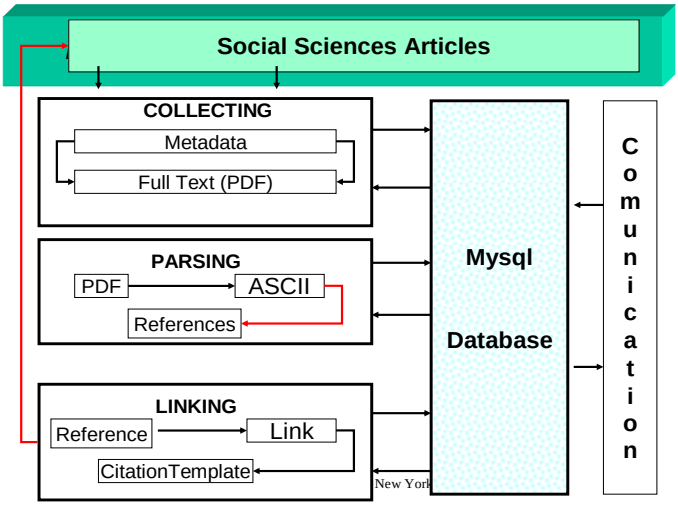


Figure 1 : l'architecture d'INCISO

Comme le montre la figure 1, l'architecture d'INCISO est fondée sur deux éléments principaux. Tout d'abord, nous travaillons sur un corpus d'articles publiés dans des revues espagnoles en sciences sociales. Nous avons constitué une banque de métadonnées contrôlées décrivant chacun des articles. Ces métadonnées sont stockées dans une base de données bibliographiques. Nous n'entrerons pas ici dans les détails de cette base de données car ce n'est pas le sujet du présent article. Ensuite, nous disposons d'une série de trois modules

correspondant aux trois étapes du processus d'établissement de liens entre les références et les documents (Barrueco, 2005) :

1. collecte des métadonnées et du texte intégral des documents,
2. analyse des documents afin de repérer là où se trouvent les références et extraire les éléments (auteurs, titre, etc.) de chacune de ces références,
3. création des liens entre les références et le document qu'elles représentent si celui-ci est disponible dans INCISO.

Il est important de noter que chaque module est tributaire de la production du module qui le précède. Ainsi, l'intégration de chaque document implique de valider les trois niveaux de traitement. Chaque document se voit attribué un statut correspondant à l'étape du processus où il se trouve. Le statut initial « nofulltex » (pas de texte intégral) et le dernier « linked » (lien établi) si tout se passe bien.

1. **Collecte**. La collecte implique trois étapes différentes : (1) collecte des métadonnées des documents, (2) téléchargement du texte intégral des documents et (3) conversion en un format prêt à être analysé par le système informatique. Les métadonnées des documents citant sont complétées avec l'URL du texte intégral des articles. Dans certains cas, les URLs fournies peuvent être erronées ou le serveur web peut être hors service lorsque le système tente d'accéder aux documents. Alors, un statut spécial est attribué aux articles et le processus s'arrête jusqu'à l'intervention manuelle du personnel éditorial qui vérifie et corrige le problème rencontré. Une fois que le fichier en texte intégral est sauvegardé sur le disque dur, nous commençons le processus de conversion. Dans un premier temps, nous vérifions si le fichier en texte intégral est compressé. Si c'est le cas, un algorithme de décompression est utilisé. Dans un second temps, nous vérifions le format du fichier. Actuellement, seuls les documents en pdf sont acceptés. Heureusement, le pdf est un format très répandu pour publier des articles scientifiques sur l'internet. La dernière étape consiste à convertir le document pdf en ASCII. Dans

ce but, nous utilisons pdftotext, le logiciel développé dans le cadre du visionneur Xpdf. Tous les fichiers pdf ne peuvent pas être convertis avec suffisamment de qualité pour permettre l'extraction. La qualité des fichiers pdf dépend principalement du logiciel utilisé pour créer les fichiers et également du bon codage des polices de caractères.

2. **L'analyse** syntaxique est l'étape la plus compliquée. Les auteurs utilisent divers formats pour leurs références et ces formats peuvent varier au sein d'un seul et même article. En outre, la manière dont les références sont indiquées dans les documents varie d'une discipline à l'autre. Compte tenu de l'importance de la phase d'analyse syntaxique, nous avons choisi de commencer avec un logiciel déjà testé plutôt que de développer un nouveau logiciel en recommençant à zéro. Notre choix s'est porté sur le logiciel développé pour le projet CitEc, qui a été décrit dans des articles comme celui de Lawrence (1999). Le logiciel de CitEc est capable d'identifier la partie du document contenant la liste des références. Ensuite, il peut séparer les différentes références de cette liste. Enfin, il procède à l'analyse de chaque référence pour en retrouver les différents éléments. Pour le moment, il n'identifie que l'année de publication, le titre et les auteurs. Cependant, ces quatre[NdT] éléments sont suffisants pour notre objectif. La qualité des références bibliographiques fournies dans les revues est variable. Par exemple, dans un même article, il est fréquent de trouver différentes formes du nom d'un même auteur, différentes formes du titre d'une même revue, etc. Nous utilisons les métadonnées contrôlées pour compléter les références et nous en améliorons la qualité avec des métadonnées provenant des éditeurs.
3. **Etablissement des liens**. Une fois que nous avons analysé les documents, l'étape suivante consiste à regarder si certaines des références extraites avec succès renvoient à des documents disponibles dans la base de données d'INCISO. Dans ce cas, il faut établir un lien entre les deux

[NdT] – La phrase précédente ne mentionne que trois éléments.

documents. Pour ce faire, nous comparons chaque référence analysée aux métadonnées contrôlées stockées dans la base de données bibliographiques d'INCISO. Actuellement, nous considérons qu'une référence représente un document d'INCISO quand :

a. l'analyse du titre de la référence et le titre dans notre collection de métadonnées sont assez proches ;
b. l'année de publication des deux items est identique ;
c. au moins un des auteurs des articles correspond aux auteurs de la notice de métadonnées.

Dans ce processus, nous prenons chaque référence, nous extrayons le titre analysé et nous le convertissons en une version normalisée appelée le titre clé. Ici tous les différents espaces et articles sont enlevés et toutes les majuscules sont converties en minuscules. Ensuite, nous sélectionnons dans notre base de données bibliographiques tous les documents qui contiennent dans leur titre tous les mots du titre clé de la référence. Tous les articles sélectionnés sont susceptibles de devenir le document cité. Dans un deuxième temps, nous calculons la distance de Levenshtein de chaque titre clé du document candidat avec le titre de clef de la référence. Si cette distance dépasse de 8 % la longueur du titre clef de la référence, le document est rejeté. Enfin, nous vérifions si l'année de publication des articles candidats et celle de la référence sont identiques. Si c'est le cas nous estimons que la référence correspond au document que nous avons. Les auteurs sont comparés uniquement quand le titre est court et qu'il ne permet pas de différencier les éléments. Les informations sur les citations sont stockées dans une table de la base de données mySQL. Cette base de données sera utilisée pour développer des indicateurs bibliométriques.

Conclusions

Dans cet article nous avons décrit une méthodologie pour développer automatiquement un index de citations. En mettant en œuvre cette méthodologie, le projet d'INCISO va essayer de

réduire les coûts élevés liés au développement d'index de citations par des moyens traditionnels. En cas de succès, cela ouvrira la voie aux pays non-anglophones pour développer leurs propres index qui pourraient être utilisés comme complément de l'ISI dans l'évaluation de la recherche.

Actuellement, nous venons juste de commencer à développer le logiciel. On s'attend à avoir les premiers résultats en 2006. Alors une période d'évaluation commencera afin de déterminer si les résultats sont suffisamment bons pour permettre à la fois la recherche d'informations et l'extraction des indicateurs bibliométriques.

Il existe d'autres projets au niveau international opérant dans le même domaine. L'innovation d'INCISO réside dans l'utilisation d'une base de données de métadonnées contrôlées qui normalise les références extraites des documents.

Bibliographie

[1] Barrueco, José Manuel, and Thomas Krichel (2005) "Building an autonomous citation index for grey Literature: RePEc, the economics working papers case" The Grey Journal, An International Journal on Grey Literature, vol. 1, no. 2, pp. 91–97

[2] Delgado López-Cozar, Emilio y otros (2005). INRECS: Índice de impacto de las revistas españolas de ciencias sociales. Biblio 3W, Revista Bibliográfica de Geografía y Ciencias Sociales, Vol. X, no. 574.

[3] Hernández Mogollón, Ricardo (2003). Citaedem.. Indice de citas de economía de la empresa. Memoria y resultados. Universidad de Extremadura.

[4] Lawrence, Steve, Kurt Bollacker, and C. Lee. Giles (1999) "Indexing and retrieval of scientific literature", proceedings of eighth International Conference on Information and Knowledge Management, CIKM99, pp. 139–146.

[5] López Piñero Jose María, Terrada María Luz. (1994). El consumo de información científica nacional y extranjera en las revistas

médicas españolas: un nuevo repertorio destinado a su estudio. Medicina Clínica, vol.. 102, pp. 104-112.

[6] Osca-Lluch Julia and Haba Julia (2005). Dissemination of Spanish Social Sciences and Humanities Journals. Journal of Information Science, vol. 31, no. 3, pp.229-236.

[7] Osca-Lluch, Julia. (2005). Some considerations on the use the impact factor of scientific journals as a tool to evaluate research in psychology. Scientometrics, vol. 65, no.2, pp.189-197.

[8] Roth, Dana L. (2005) "The emergence of competitors to the Science Citation Index and the Web of Science", Current Science, 2005, vol. 89, no. 9. pp. 1531–1536.

[9] Tortosa, Francisco, Civera, Cristina, Osca-Lluch, Julia, Barrueco, José Manuel, Quiñónes, Elena, Peñareanda, María, Martinez, Francisco, López, Juan José (2005). "Creación de un índice de citas de revistas españolas de psicología". I Jornadas Españolas de Indicadores para la Evaluación de la Ciencia, Madrid. Disponible en: http://www.cindoc.csic.es/info/fesabid/25.htm

αMETIST

Partie 1

Partie 2

Partie 3

Partie 4 :
ARTIST, un lieu d'expérimentation

A propos du numéro zéro d'AMETIST : rapport sur une expérience d'appropriation

Jacques Ducloy (1)
jacques.ducloy@inist.fr

Patricia Gautier(1)
patricia.gautier@inist.fr

Magali Rasolomanana(1)
magali.rasolo@inist.fr

Clotilde Roussel(1)
clotilde.roussel@inist.fr

Djamila Safa(1)
djamila.safa@inist.fr

Pierre Wirtz(1)
pierre.wirtz@inist.fr

(1) INIST / CNRS , France

Mots-clés : écriture numérique, appropriation des techniques éditoriales, revue numérique.

Keywords : electronic journal, electronic writing, information technology appropriation

Résumé : Ce rapport relate les principaux faits marquants liés à la production du numéro zéro d'AMETIST. Il donne des éléments sur les moyens à mettre en œuvre dans un centre de documentation de la recherche pour initialiser une activité éditoriale autour de revues scientifiques avec des contraintes techniques liées aux domaines relevant de l'ingénierie et de l'écriture numérique. Les problèmes liés à la dualité des supports papier / numérique sont abordés.

Introduction

Avec le projet de revue AMETIST, nous poursuivons un objectif ambitieux à moyen terme : arriver à créer un périodique de référence à la fois sur un plan scientifique et technique. En pratique et à court terme, nous démarrons avec une équipe de soutien logistique qui, bien que disposant d'un environnement privilégié à l'INIST, est encore inexpérimentée. Nous sommes donc dans des conditions proches d'une équipe scientifique de terrain voulant se lancer dans une « aventure d'appropriation des pratiques éditoriales en mode numérique ».

Nous vous proposons ici un compte rendu qui se veut transparent sur cette expérience d'appropriation. En particulier, nous ferons état de difficultés pour lesquelles les professionnels de l'édition auront probablement un regard critique là où ils bénéficient d'une longue expérience et de savoir-faire que nous ne contestons pas. Sur le forum ARTIST et dans des articles antérieurs [2][3] nous avons identifié la nécessité pour la communauté académique de reprendre à sa charge une partie du processus éditorial. En effet, en s'appuyant uniquement sur des critères relevant de l'économie de l'édition, les conditions de diffusion ne permettent pas toujours de rentabiliser les traitements éditoriaux nécessaires à la dissémination d'activités scientifiques en émergence.

En raisonnant avec une approche différente, celle de l'économie de l'innovation, les organismes de recherche sont confrontés à une alternative décisionnelle : l'externalisation par une politique de subvention vers l'édition commerciale ou l'appropriation en s'appuyant notamment sur leurs réseaux de spécialistes en IST. Dans sa réalisation, le « projet AMETIST » est une expérimentation de cette deuxième proposition. Cet article, et ceux qui suivront sur ce thème, veulent donner des éléments d'appréciation à la question suivante : pour un centre bibliothéconomique ou documentaire, quel est l'investissement nécessaire à une appropriation des pratiques et techniques éditoriales ?

Nous aborderons d'abord les aspects techniques liés au lancement de la revue AMETIST sous un angle général. Puis nous illustrerons ces réflexions à partir de deux exemples. A propos de l'article sur la

transformation d'un thésaurus en ontologies nous traiterons de l'écriture numérique des articles scientifiques à contenu technique. A propos de la traduction de l'article de Carl Lagoze, nous étudierons les pratiques collectives et la mise en ligne des traductions. Enfin, nous évoquerons l'appropriation des pratiques éditoriales par une entité documentaire.

Avertissement

Pour la rédaction de cet article nous avons été confrontés à des difficultés particulières liées à la dualité des supports et aux contraintes de notre échéancier. En effet, l'impression du fascicule papier demande un délai incompressible d'un mois avant sa diffusion et celui-ci sera alors définitivement figé. Pour l'édition numérique, les contraintes sont nettement plus souples. Les deux versions de cet article seront donc assez différentes. La version papier présente un résumé de nos observations au 15 juillet 2006. Le document numérique sera sensiblement plus détaillé et remis à jour jusqu'en novembre 2006.

1 Appropriation des techniques éditoriales

Avant d'entrer dans les détails techniques, voici un rappel des conditions initiales de notre expérience.

L'idée de monter une revue électronique à partir de l'expérience du forum ARTIST a germé il y a un an environ. Son contenu devait s'organiser autour d'une dynamique à trois composantes. En premier lieu, nous voulions, de façon classique, encourager la recherche autour de la thématique de l'appropriation et promouvoir les « bons articles sur le sujet ». Ensuite, nos premières expérimentations sur les forums ont montré l'intérêt et les difficultés des démarches collectives. La revue devrait donc servir à cadrer les discussions en leur donnant un objectif et une échéance. Enfin, il nous avait paru opportun de lancer un banc d'essai pour l'écriture numérique.

Sur un plan technique, nous nous devions de devenir « exemplaires » en termes de respect des contraintes de normalisation et d'indexation. Le comité éditorial a validé ces options et nous a ajouté

une contrainte complémentaire avec une dualité de supports de diffusion : papier et numérique.

En pratique, l'INIST a dégagé un poste de secrétaire de rédaction qui a été pourvu par une documentaliste qui doit maintenant s'approprier cette fonction. L'INIST apporte également un réseau interne de compétences informatiques, documentaires ou éditoriales mais qui n'ont jamais été impliquées dans une initiative identique. Nous sommes donc dans une situation assez représentative de l'appropriation des techniques éditoriales par un établissement de recherche qui veut se lancer dans ce type de projet, sans expérience significative préalable.

1.1 Initialisation du cycle éditorial

Nous désirons donc parvenir à un modèle éditorial où l'auteur conçoit un article, dans un contexte d'écriture numérique. Il doit en extraire une version lisible sur papier (en format A5). Il devrait également disposer d'un ensemble de recommandations comportant une charte graphique et un code typographique[1].

En réalité, nous partons d'une situation vierge. Nous avons décidé de sortir une maquette, le numéro zéro, pour le 15 septembre 2006 à partir d'une sélection d'articles faite lors des journées VSST de janvier 2006. Au moment de la rédaction de cet article le planning prévu est le suivant :

- 15 septembre 2006, Semaine du Document Numérique (SDN 2006 du 18 au 22/09/2006).[2]
    - Diffusion du numéro 0 en version papier et présentation d'un prototype de la version numérique.
    - Recherche d'articles pour le numéro 1 par le comité de rédaction (la SDN 2006 donne une occasion de contacter des auteurs potentiels).

[1] Le code typographique désigne un guide de bonnes pratiques qui définit précisément les règles typographiques.

[2] https://diuf.unifr.ch/event/sdn06/accueil.html

- Novembre 2006
    - Version finale du numéro 0 en mode numérique,
    - Première version des recommandations aux auteurs incluant la charte graphique, le code typographique et des éléments techniques (formats) et documentaires (vocabulaire).
- Mars 2007 : sortie du numéro 1, appel aux contributions pour le numéro 2.

Autrement dit, nous visons un cycle stabilisé pour septembre 2007.

Pour ce numéro 0, nous sommes donc confrontés à une démarche intermédiaire, inverse de celle qui est recherchée à moyen terme. En effet, nous avons sélectionné des articles conçus pour des actes diffusés dans un format A4. Nous avons proposé aux auteurs quelques adaptations et révisions pour produire en priorité la version papier (en format A5) et nous approfondissons la version électronique en fin de cycle.

1.2 Choix d'un format de document

La pérennité des documents est un des enjeux de l'écriture numérique. Les documents numériques sont codés dans des formats qui évoluent avec le temps. Par exemple, il est aujourd'hui difficile, voire impossible, de lire des images ou des documents texte qui ont été créés sur des ordinateurs au cours des décennies antérieures. C'est pourquoi nous devons mener une réflexion sur les moyens de présenter l'information dans un format qui soit lisible dans encore 10, 20 ans ou plus.

Nous avons ouvert sur ARTIST un forum de discussion sur les bonnes raisons de rédiger un document dans un format structuré et plus précisément XML. Dans la philosophie des projets tels que Cyberthèses[3], nous souhaiterions à terme disposer d'une souche normalisée de nos articles reposant sur un schéma ou une DTD tel que DocBook ou TEI, en utilisant des mécanismes automatiques

[3] http://www.cybertheses.org/

(feuilles de style XSLT par exemple) pour produire les différentes versions (PDF ou XHTML).

La mise en pratique n'est pas si simple et nous évoquerons dans la section 2 quelques difficultés rencontrées dès que l'on dépasse l'information purement textuelle pour intégrer des éléments techniques (figures, formalismes divers) propres aux articles scientifiques intégrant une dimension ingénierie.

Le choix de l'option « écriture numérique » d'articles pour une revue scientifique pose également des problèmes de structuration. Dans ce numéro, nous avons traité un article conçu initialement pour une version papier, et donc une lecture essentiellement linéaire. Pour le mettre en ligne avec des mécanismes de lecture non linéaires nous avons été amenés à le restructurer fondamentalement dans une logique d'enrichissement. Cela conduit à définir en fait deux structures pour le même document.

Dans l'avenir, il faudra probablement prévoir l'option inverse, avec un article conçu dans un contexte interactif d'où il faudra extraire une version imprimable, dans une logique d'appauvrissement. Nous allons donc rencontrer des problèmes intéressants du point de vue de la codification XML tels que : comment gérer facilement cette dualité de structure, et notamment pendant la phase de relecture ?

Avec le choix des options électroniques, la revue AMETIST devrait donc devenir un banc d'essai pour des travaux sur l'évolution des documents structurés.

1.3 Composition, gestion et mise en ligne des documents

Si nous prévoyons donc de passer à terme vers un ensemble XML, pour les premiers numéros nous devrons traiter un flot de documents majoritairement écrits en format Word.

Concernant la version papier, et pour ce numéro, nous avons utilisé Word pour assembler les articles et produire le fascicule « prêt-à-tirer ». Sans entrer dans les détails, cette solution ne nous paraît pas optimale et, pour la version papier du numéro 1, nous devrons « nous» approprier un ensemble de composition plus professionnel.

Concernant la version numérique, il existe une offre de CMS[4] pour créer et gérer un site éditorial. Nous avions déjà utilisé SPIP[5] pour le site ARTIST et nous expérimentons LODEL[6] pour mettre en ligne le numéro 0 d'AMETIST. Au moment où la version papier de cet article est rédigée nous n'avons pas encore assez de recul pour rendre compte de façon significative de cette expérience qui sera détaillée dans la version électronique[7].

2 Autour de l'écriture scientifique technique et numérique

Les exemples de cette section concernent l'article de Claude Chrisment et al. (IRIT) « D'un thésaurus vers une ontologie de domaine pour l'exploration d'un corpus ». Il avait été accepté pour le colloque VSST de Lille et repéré par le comité de rédaction d'AMETIST. Pour des raisons de calendrier et dans l'esprit de ce numéro 0, nous avions choisi de limiter les demandes de modifications du contenu de l'article dans sa version papier afin de concentrer nos efforts sur la version numérique.

La version papier ne devait donc être qu'une simple « amélioration de la mise en forme de l'article initial ». Les problèmes rencontrés ont cependant nécessité des discussions approfondies avec les auteurs à propos de la charte éditoriale des notions formelles. Nous avons également rencontré des difficultés avec la gestion des figures.

2.1 Homogénéisation des formalismes

La principale difficulté rencontrée pour l'amélioration technique de la version papier concernait l'écriture de règles formalisant les transformations. Le document initial contenait des règles telles que :

---

[4] *Content Management System* ou *Système de Gestion de Contenu*.

[5] http://www.spip.net

[6] http://www.lodel.org

[7] http://ametist.inist.fr

Si t3 UP t1 alors t1 et t3 sont regroupés, avec t1 terme préféré

Si t1 UPD t2 alors t1 et t2 sont regroupés, avec t1 terme préféré

(R1)

Nous avons retravaillé l'écriture en proposant de rendre plus lisible les « opérateurs formels ».

Si t3 UP t1 **(R1)**

alors t1 et t3 *sont regroupés*

t1 **devient** *terme préféré*

Si t1 UPD t2

alors t1 et t2 *sont regroupés*

t1 **devient** *terme préféré*

La complexité de certaines règles nous a demandé une compréhension totale du contenu de l'article et une discussion de fond avec les auteurs pour obtenir un résultat conforme à leur volonté.

A ce sujet, nous avons également rencontré des difficultés avec l'affichage des caractères spéciaux, tels que les opérateurs ensemblistes, intégrés dans les règles. Nous avons pu ainsi approfondir le traitement des caractères Unicode dans les polices du logiciel Word…

Cet exemple illustre le besoin d'une expertise conjointe sur les techniques éditoriales et la connaissance du domaine scientifique traité.

2.2 Adaptation des figures aux différents médias

La gestion des figures nous a permis d'appréhender les limites des mécanismes de génération des versions multiples d'un article à partir d'une souche unique.

L'article initial contenait une figure complexe résumant les trois étapes de la méthodologie adoptée. Ce schéma avait été conçu pour un format A4 conformément aux spécifications du colloque VSST.

Dans la revue numérique, nous avons choisi de l'utiliser comme un moyen d'orientation vers les différentes parties de l'article, avec des mécanismes d'interaction divers. La figure a été de ce fait assez considérablement modifiée à l'aide des diapositives de la présentation orale. Le plan de l'article en version numérique va donc se trouver sensiblement modifié.

Dans la revue papier, pour le passage en format A5 sa réduction rendait les légendes illisibles. Nous avons du l'éclater sur deux pages, en utilisant d'ailleurs la structure dégagée dans la version numérique.

Pour une diffusion de l'article seul en version A4 ou pour une diffusion en PDF sur une archive ouverte par exemple, la figure initiale retrouve son intérêt.

Nous avons donc rencontré un exemple significatif où le même « concept graphique » doit donner lieu à plusieurs réalisations distinctes en fonction du média cible. Dans ce cas précis la figure n'était pas une simple illustration (que l'on peut zoomer sans problème particulier) mais un élément structurant de l'article. Cet exemple nous paraît assez révélateur des différences que l'on peut rencontrer entre les « revues SHS » et les articles relatifs à un domaine relevant de l'ingénierie. Pour AMETIST, il renforce l'idée d'un banc d'essai pour un axe fort de la recherche et développement autour de ce type de communication scientifique.

3 Autour des traductions : travail coopératif et mise en ligne spécialisée

Le chapitre précédent a illustré la nécessité de collaborations entre une équipe technique et les auteurs d'un article pour une amélioration rétrospective. Dans cette nouvelle partie, nous allons aborder l'angle du travail coopératif à plus large échelle, à propos de

la traduction de l'article de Carl Lagoze : « Qu'est-ce qu'une bibliothèque numérique, au juste ? ».

3.1 Historique des travaux et contributions

Le lancement de cette action a déjà mobilisé un ensemble d'acteurs dans un espace de temps très court (12 heures). En effet, le matin du 16 Novembre 2005, Jean-Michel Salaün nous a interpellés dans le cadre des échanges de la liste du RTP-DOC[8]. Le jour même, et simultanément, Frédéric Martin de la BnF[9] et l'équipe ARTIST signalions notre intention de traduire cet article. Nous avons aussitôt sollicité Bonnie Wilson éditrice en chef de D-Lib Magazine et ensuite contacté Carl Lagoze. Miracle de la communication par internet : à 17 heures 30, nous avions toutes les autorisations. Nous pouvions donc commencer à œuvrer…

En revanche les travaux proprement dits ont duré bien plus longtemps.

Les deux premières étapes « traduction par Frédéric Martin et révision par Catherine Gunet (de l'INIST) » ont demandé de nombreux allers et retours entre traducteur et réviseur, se sont déployées sur deux mois. En parallèle, une traduction arabe a été réalisée au sein de l'équipe ARTIST.

Nous avons ensuite organisé les discussions terminologiques[10] à partir des difficultés relevées par les traducteurs jusqu'en juin 2006.

La dernière phase est dédiée à l'intégration des corrections et à la mise en ligne. Le lecteur aura à sa disposition deux réalisations assez différentes. La version papier sera une traduction classique reprenant, paragraphe par paragraphe, le découpage initial. Pour la version en ligne, nous prévoyons une navigation entre les versions anglaises et françaises pour celui qui désire approfondir un passage dans sa formulation originale.

[8] http://rtp-doc.enssib.fr/

[9] Bibliothèque nationale de France

[10] http://artist.inist.fr/rubrique.php3?id_rubrique=113

Pour les travaux collectifs, nous avons exploré une voie qui, après réflexion, nous paraît prématurée. Cette expérience chronophage a retardé la mise en ligne. Nous voulions favoriser des discussions sur les traductions proprement dites. Pour cela nous avions prévu un découpage du texte en petits tableaux trilingues pour faciliter les échanges. Cette hypothèse de travail n'a pas été concluante. Il faudra donc envisager une formule un peu différente pour ce travail de traduction « basique ». Les discussions terminologiques ont été plus fructueuses.

3.2 Les discussions terminologiques

Nous avons demandé au traducteur et au réviseur de nous faire part de leurs difficultés de traduction. Pour chaque terme repéré, nous avons rédigé une fiche terminologique qui a servi de point d'ancrage à un forum de discussion.

Nous avons relevé plusieurs difficultés : certaines inhérentes à des expressions connues donnant lieu à discussion sur des ambiguïtés d'interprétation et d'autres sur des termes courants qui, employés dans un sens métaphorique, ont suscité de vives discussions.

Le site en ligne ne reflète pas l'intégralité des débats, dans la mesure où nous avons eu quelques échanges sous différentes formes (téléphonique, réunion…) sur la plupart des discussions. Concernant l'article proprement dit, le choix définitif a été laissé au traducteur et au réviseur.

Certaines expressions équivoques nous ont confrontés à la réalité du métier de traducteur/réviseur. Nous avons pu appréhender la difficulté de choisir la traduction idoine. Dans un souci de pertinence, un forum a été mis en place sur ARTIST. Les discussions ont été initialisées par des fiches terminologiques faisant référence à des référentiels terminologiques divers (termSciences[11], ATILF[12], Grand dictionnaire terminologique de l'OQLF[13], Glossaire du site

[11] http://termsciences.inist.fr

[12] http://atilf.atilf.fr/tlf.htm

[13] http://www.granddictionnaire.com/

« Libre accès à l'information scientifique et technique »[14]). Chaque internaute pouvait participer à la discussion et proposer une traduction pour une expression.

En voici trois exemples caractéristiques :

- **Digital library** : si la traduction de « library » par bibliothèque fait quasiment l'unanimité, l'ambiguïté est liée au terme digital que l'on peut traduire par numérique ou électronique voire virtuelle. Les trois variantes de cette expression étant quasiment synonymes, les débats ce sont vite stabilisés. Le traducteur a choisi « bibliothèque numérique » dans le titre.

- **Institutional repository** : la discussion a été fortement alimentée par une forte contribution de Guylaine Baudry argumentant sur la proposition de « dépôt institutionnel » comme traduction de « institutional repository ». Elle a attiré l'attention sur un glissement de sens en français autour du terme « archivage ».

  > *En fait, "archive/archivage" en français veut dire exactement l'inverse du sens en anglais en informatique de "to archive/archives".*

  D'autres contributions ont plutôt soutenu l'expression « archives institutionnelles ». Le traducteur a en définitive conservé sa version initiale «entrepôt institutionnel ».

- **Stuff** : ce terme est utilisé tout au long de l'article pour désigner le composant de base d'une bibliothèque numérique qui correspond à un vaste champ métaphorique.

  Si la discussion a été riche, elle n'a cependant pas permis de converger vers un terme générique faisant l'unanimité. Le traducteur a finalement opté pour l'expression « matériau numérique » dont le spectre métaphorique est plus restreint mais qui rend l'article parfaitement lisible et compréhensible.

[14] http://www.inist.fr/openaccess/

Cette démarche nous a instruits sur la difficulté à la fois de trouver une traduction claire et adaptée et qui correspond à une valeur d'usage de la communauté.

3.3 La mise en ligne des traductions, aspects techniques

Dans la version en ligne des traductions, nous voulons présenter simultanément et de manière interactive les deux versions de l'article côte à côte et paragraphe par paragraphe. Plus précisément, le document se présente en français et le lecteur peut interagir pour accéder localement aux paragraphes en anglais. Pour faire apparaître et disparaître une ou l'autre version, les spécifications du langage XHTML ne suffisent pas. Il a donc fallu pour cela rajouter des mécanismes en JavaScript. Ces scripts réagissent au clic sur les drapeaux français et anglais pour faire apparaître la version concernée.

Nous avons rencontré une difficulté pour intégrer un mode dégradé à disposition des lecteurs qui ne disposeraient pas d'un navigateur récent (ou qui auraient désactivé JavaScript). En effet, JavaScript est un langage plus ou moins normalisé et surtout totalement dépendant du navigateur qui l'exécute. Pour présenter les paragraphes côte à côte, nous avons du utiliser des tableaux HTML. Sans JavaScript, les deux versions du document sont ouvertes côte à côte sans que l'utilisateur puisse masquer l'une ou l'autre partie. Avec le JavaScript, seule la version française est affichée et des boutons permettent de naviguer d'une version à l'autre.

Avec l'usage de JavaScript, nous avons donc du prendre quelques libertés par rapport aux critères de normalisation liés à la pérennité. Mais, dans le cas des traductions nous mettrons en place des solutions plus propres à moyen terme. En effet, une traduction est une mise en relation de deux versions d'un même texte. Chaque version peut donner lieu à un document normalisé (et pérenne), seul le mécanisme générique de mise en ligne est spécifique d'un logiciel d'affichage.

4 Mutations liées à l'appropriation des pratiques éditoriales

Dans les sections précédentes, nous avons présenté un échantillon des problèmes techniques rencontrés dans le traitement éditorial d'articles scientifiques comportant des spécificités propres à l'ingénierie ou à la technologie. Dans cette section, nous voulons compléter cette réflexion en abordant les difficultés liées aux changements de pratiques ou de point de vue des personnels issus du monde de l'IST et qui auraient à prendre en charge une activité éditoriale. En effet, comme de nombreuses entités documentaires ou bibliothéconomiques, l'INIST étudie l'opportunité d'une telle hypothèse qu'il faut donc évaluer. Nous venons d'ailleurs de mettre en évidence le besoin simultané d'experts scientifiques et de spécialistes en sciences de l'information qui sont justement un de nos points forts.

Comment une entité documentaire peut-elle s'approprier donc les pratiques éditoriales ? Pour un approfondissement du sujet nous conseillons la lecture de l'ouvrage de Thierry Chanier [1]. Nous proposons simplement ici de revisiter les trois verbes qui nous sont parfois utilisés pour résumer notre activité : « collecter, traiter et diffuser l'information scientifique et technique ».

4.1 Collecter

Pour certains professionnels ce premier verbe cristallise parfois une perte progressive de marge de manœuvre.

En effet, la chaîne documentaire traditionnelle organise cette collecte à partir des informations fournies par les éditeurs. Cette action demandait au bibliothécaire une réflexion décisionnelle stratégique et complexe lorsque les éditeurs étaient très nombreux et qu'il fallait sélectionner les articles à analyser. Elle tend à se réduire à sa plus simple expression avec l'alimentation directe à partir des métadonnées des fournisseurs. Dans les archives institutionnelles avec dépôt par les auteurs, l'action des professionnels est souvent limitée à la mise en place technique du système et à l'accompagnement des déposants. La collecte se réduit donc à des

fonctions de gestion éventuellement complétées par des pratiques de veille et par l'animation d'un comité d'experts.

La prise en compte de l'édition électronique ouvre de nouvelles perspectives sur la collecte mais implique un changement radical de point de vue.

En effet, la création d'une revue électronique demande en fait la mise en place d'un réseau de collecte et de sélection totalement différent de celui des contacts traditionnels des centres de documentation. Pour AMETIST, nous nous situons à l'étape initiale de cette action et nous en mesurons les difficultés : il faut par exemple s'appuyer sur des relais ayant une fonction d'animation de la recherche et non seulement sur nos collègues documentalistes de laboratoire. Pour y répondre, le forum ARTIST, qui maintient une animation permanente de la communauté d'auteurs potentiels, nous a été utile mais il n'a pas été suffisant pour initialiser le processus. Pour les premiers numéros, nous partons par exemple de sélection d'articles de colloques avec lesquels nous avons des relations privilégiées.

Le tissu de relations concerné par l'action de collecter est donc totalement différent des réseaux documentaires traditionnels, et les modes de relation doivent également être repensés.

4.2 Traiter

Le traitement proprement dit pose naturellement de nombreux problèmes d'appropriation de pratiques. Dans les sections précédentes nous avons repéré des problèmes techniques qui révèlent des besoins de plan de formation. Nous nous limitons ici aux problèmes d'organisation.

Le traitement d'un article demande en fait la collaboration de deux grands secteurs de compétences : techniques documentaires ou éditoriales d'une part, scientifiques et rédactionnelles de l'autre. Chaque secteur peut être divisé en activités plus spécialisées, par exemple : catalogage, consolidation éditoriale et transformation de schéma pour la partie technique ; indexation et relecture pour la partie rédactionnelle.

Dans les chaînes documentaires ou bibliothéconomiques classiques, on cherche à uniformiser des actions sur un grand nombre d'objets informationnels. L'informatique joue un rôle simplificateur et unificateur. Dans les traitements éditoriaux, on cherche au contraire à mettre en valeur les contenus d'un petit nombre de documents. L'informatique va au contraire servir à exprimer les spécificités thématiques. Notre première expérience nous a montré une imbrication des activités techniques et la nécessité d'une compréhension du contenu des articles beaucoup plus forte que nous ne l'avions imaginée.

Autrement dit, il nous semble difficile de vouloir produire une revue scientifique, avec des contraintes techniques propres au domaine et dans un contexte numérique fort, sans réunir une équipe de spécialistes travaillant en étroite coopération. En effet, les auteurs et réviseurs sont en permanence confrontés aux limites actuelles de l'édition électronique et il faut faire des compromis en prenant en compte d'une part des contraintes relevant de l'ingénierie logicielle et, d'autre part, de la connaissance du domaine scientifique.

Les bibliothèques et centres de documentation ont souvent segmenté leur organisation avec une structure hiérarchique de services, dont les cellules de base regroupent des personnes qui exercent le même métier. Une telle organisation était jugée optimale avec les contraintes d'une chaîne de production traditionnelle. Elle ne l'est plus pour un traitement éditorial fortement lié au type de contenu comme pour AMETIST. Pour les petits centres le problème est un peu différent dans la mesure où les acteurs sont géographiquement rapprochés. En revanche, il faut veiller au bon niveau d'expertise de l'équipe de rédaction.

4.3 Diffuser

Pour ce dernier verbe, avec une approche un peu simplificatrice (qui fait abstraction de points importants comme les abonnements), la mutation peut se formuler simplement.

Un des rôles traditionnels d'une bibliothèque ou d'un centre de documentation est de diffuser vers une communauté bien définie de l'information sélectionnée dans une offre mondiale.

Dans la pratique éditoriale, ou de dissémination des résultats de la recherche, il s'agit de faire exactement le contraire : diffuser au niveau mondial les informations élaborées par une communauté scientifique bien définie.

Conclusion

Créer une revue telle qu'AMETIST, c'est prévoir et maîtriser plusieurs aspects : le contenu scientifique, les techniques éditoriales, la normalisation et l'offre logicielle. L'équipe ARTIST a dû retravailler chaque article pour l'adapter au média choisi et s'approprier les techniques et pratiques afférentes.

Comme cela a été évoqué dans l'introduction, l'objectif de l'expérience d'appropriation des technologies dans la production de la revue AMETIST est multiple. La méthode expérimentée devrait servir de base aux prochaines parutions de la revue et constituer un point de départ pour la formalisation d'un processus d'édition. La capitalisation de ce savoir faire, qui sera amélioré progressivement en fonction des retours des auteurs, de ceux des lecteurs et des constats a postériori de l'équipe, est une étape incontournable du phénomène d'appropriation, objectif d'ARTIST.

A l'issue de cette première phase, nous avons déjà un premier retour d'expérience sur la gestion de la dualité de supports papier d'une part et électronique de l'autre.

La revue papier nécessite de choisir un format et un travail de mise en page important pour fournir un fichier propre à l'imprimeur et ainsi obtenir une qualité d'impression optimale. De plus, la revue papier doit respecter des délais incompressibles d'édition. Une édition papier est par définition figée : il faut donc un effort de relecture plus important.

La revue électronique se distingue fondamentalement de la revue papier car la lecture n'est plus uniquement linéaire et permet d'introduire de l'interactivité. Le travail en amont est différent, puisqu'il faut faire des choix technologiques qui ne se posent pas avec la version papier : choix d'un logiciel de mise en ligne, choix

d'un format de données, choix d'un mode de lecture et de diffusion…

Pour les auteurs, cette dualité de support va donc induire un coût de rédaction plus important. Nous commençons à entrevoir qu'elle peut aussi améliorer la qualité de rédaction simultanée des deux versions et, dans le même temps, de la réflexion scientifique sous-jacente.

Pour les institutions, au-delà des aspects techniques pour lesquels il est toujours possible de planifier des programmes de formation ou de reconversion, le problème le plus difficile à court terme est probablement celui de la prise en compte de l'évolution des pratiques dans les institutions de la recherche.

De beaux sujets pour des appels à communication.

Bibliographie

[1] T. Chanier. *Archives ouvertes et publication scientifique. Comment mettre en place l'accès libre aux résultats de la recherche ?* L'Harmattan, Paris.2004

[2] J. Ducloy. *Plaidoyer pour un réseau d'inventaires des résultats de la recherche*, Colloque VSST, Toulouse 2004 <http://archivesic.ccsd.cnrs.fr/sic_00001147>.

[3] J. Ducloy, L. Grasset. *Appropriation des réseaux d'inventaires scientifiques par les entités de recherche en émergence*, Colloque International sur l'Information numérique et les enjeux de la société de l'information, Tunis, Tunisie 2005 <http://archivesic.ccsd.cnrs.fr/sic_00088884>

.

Impression : SPEI 54420 Pulnoy

Partenaires scientifiques

CIDE

Colloque
International sur le
Document
Electronique

SDN

Semaine du
Document
Numérique

RTP-DOC

Réseau
Thématique
Pluridisciplinaire
Document et
Contenu

VSST

Veille
Stratégique
Scientifique et
Technologique

Soutien logistique

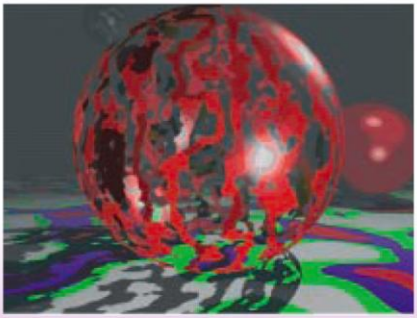
CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

inist Institut de l'Information Scientifique et Technique

V.S.S.T. '2007

VEILLE STRATEGIQUE SCIENTIFIQUE & TECHNOLOGIQUE

SYSTÈMES D'INFORMATION ELABORÉE,
BIBLIOMÉTRIE, LINGUISTIQUE,
INTELLIGENCE ÉCONOMIQUE



MARRAKECH

21 - 25 octobre 2007

associé à l'UPC & la SFBA

http://atlas.irit.fr