Valoriser la production grise Jacques Cuvillier

IUT Nantes, LABCIS/ICOMTEC Poitiers <Document Libre>

Introduction

Des articles, des cours, des exercices, des corrigés d'annales, des présentations, des fiches-résumés, des tutoriels, des documentaires... des centaines de milliers de documents issus des activités enseignantes ou associatives sont offerts au public et ne demandent qu'à être utilisés pour diffuser du savoir sur de nombreux sujets artistiques, sociaux, environnementaux, ainsi que sur nombre de matières d'enseignement.... C'est la production grise appelée aussi "*littérature grise*".

Une part significative de ces documents est en ligne, et ce depuis le début de l'utilisation publique d'Internet. Jusque là réservé à une sorte d'initiés, la Toile est vite apparue aux yeux de beaucoup comme le vecteur de communication qui allait permettre le diffuser le savoir à grande échelle. L'usage des pages html et les liens hypertextes était de plus relativement facile d'utilisation et d'apprentissage rapide. Avec l'enthousiasme et la générosité des auteurs , de nombreux sites ont fleuri un peu partout.

Une décennie s'est écoulée depuis l'essor d'Internet...

Pourtant de nombreux progrès restent à faire pour que le meilleur parti puisse être tiré des extraordinaires possibilités du réseau mondial. Dans certains secteurs, comme la recherche, des avancées rapides ont été observées, dans d'autres, ils ont été beaucoup plus lents. C'est le cas en particulier de la production grise, dont les ressources représentent un volume impressionnant à l'échelle de la Toile, mais dont l'exploitation mutualisée reste encore balbutiante. Pour comprendre peut-être faut-il se pencher sur les motivations essentielles qu'il y a à publier en ligne.

Pourquoi diffuser ?

On pourrait seulement répondre : « parce que c'est bien de diffuser du savoir », « parce que cela rendrait service à la société ». Mais il est sans doute plus facile de réunir les moyens pour démarrer un projet en invoquant des raisons moins générales, mais plutôt un intérêt bien circonscrit à l'organisation qui est à l'origine.

De fait, les tendances ont été très diverses, allant d'un extrême à l'autre. Tandis que les uns s'empressaient de se bricoler un « site » pour y placer spontanément leur production et la mettre à la disposition des collègues, des étudiants, de tout le monde, avec sans doute la satisfaction s'être investi pour quelque chose qui en vaille la peine, d'autres raisonnaient en part de marché et en montant de subventions. Les uns laissaient toutes leurs pages librement accessibles, les autres plaçaient login et mots de passe.

Entre les deux, il y a le cas général : l'action motivée par son utilité sociale, mais qui permet au passage de montrer son activité, en volume et en qualité, le site étant vu – entre autres choses – comme un outil de communication.

Les orientations ont cependant évolué diversement selon les secteurs et selon les matières. Beaucoup de documents en informatique, en électronique, beaucoup moins en gestion... allez savoir pourquoi...

Le rôle de pionniers des chercheurs[1]

C'est sans doute la recherche de niveau international qui a ouvert les voies les plus pragmatiques. Pour les organisations de chercheurs, la question « pourquoi publier » amenait des réponses claires :

- utiliser la toile pour publier, attendu que son audience devait dépasser de loin les actes de n'importe quel colloque ;
- utiliser la Toile pour faciliter la recherche des articles nécessaires à leurs travaux. Ce n'était pas en soi une innovation, vu que l'usage de bases de données accessibles par des moyens télématiques était pratiquée depuis de nombreuses années.

Utiliser la Toile pour s'affranchir autant que faire se peut des tarifs scandaleux pratiqués au niveau des revues scientifiques, et qui avaient fini par aboutir à cette situation absurde où certains laboratoires n'avaient même pas les moyens de s'abonner à des revues dans lesquelles ils étaient eux-mêmes sensés publier.

1 Voir 15 années de publications scientifiques : http://www.tours.inra.fr/prc/internet/documentation/communication_scientifique/comsci.htm

A partir de ces objectifs, une approche rationnelle s'est mise en place[2], afin de traiter les questions cruciales :

- traiter correctement les questions juridiques, notamment avec un usage adapté du copyright et des cessions de droits d'auteurs ;
- converger vers des standards de fichiers permettant de rendre les contenus utilisables sur tout système ;
- développer les outils permettant l'auto-édition des articles par les chercheurs ;
- se mettre d'accord sur un usage cohérent des métadonnées qui permettent de décrire des ressources ;
- convenir de procédures permettant d'explorer le contenu des lieux de stockage des ressources[3].

En ce qui concerne le premier point – à lui seul il nécessiterait un long article – contentons-nous de dire que **la mise en ligne de tout document doit être accompagnée d'une cession de droits.** La simple mention souvent rencontrée « tous droits réservés » est totalement inadéquate pour des documents dont la finalité est par exemple d'être reproduits et distribués à des étudiants. Il existe une certaine variété de licences adaptées à des contextes particuliers[4]. Je citerai les licences GNU/FDL, Creative Commons, Licence d'Art Libre (pour les oeuvres graphiques en particulier) et aussi la Charte de Document Libre – déjà parue dans la revue du Gesi – qui stipule le minimum de cessions de droits de la part de l'auteur (libre à lui de les élargir) et qui propose un mode de travail collaboratif.

Pour ce qui est du second point, le standard ouvert de prédilection est bien entendu le XML. C'est la manière idéale de transmettre un contenu « interopérable ». Autre avantage : le contenu et la forme du document sont véhiculés par des fichiers distincts. Les cessions de droit peuvent s'appliquer au contenu seul, indépendamment de la mise en page, ce qui est une bonne chose, car les feuilles de style qui déterminent la forme du document sont souvent l'objet d'un copyright et d'une protection à part.

Les outils d'auto-édition ont été créés. En fait, il en existe un certain nombre. Le plus populaire est sans doute SPIP[5], mais il y a aussi LODEL[6]. Mais l'outil de prédilection des chercheurs de la mouvance Open Archives est certainement Eprints. Développé principalement à l'université de Southampton, ce logiciel libre écrit en perl permet l'«auto-archivage» de documents ainsi que la gestion des métadonnées. Il doit son existence à la communauté de développeurs Eprints et sur la constance de Christopher Gutteridge.

Vous avez dit "métadonnées" ?

Les deux derniers points, concernent sans doute un aspect aussi important qu'ignoré du public. Les métadonnées, c'est quoi ? Prenons par exemple le cas d'une photographie. Peut-on dire qu'elle se suffise à elle-même ? Si dans certains cas elle se passe de commentaires – comme par exemple le portrait de la reine d'Angleterre ou celui du premier ministre – la plupart du temps, elle n'est exploitable qu'assortie d'une légende. Que représente-t-elle ? à quelle date a-t-elle été prise ? et si on s'intéresse de près un cliché pris par un appareil numérique et qu'on examine les données enregistrées, on aura aussi un certain nombre de renseignements portant sur les conditions de prise de vue, de définition de couleurs, etc...

Le regret que l'on peut avoir vis à vis des moteurs de recherche les plus en vogue, c'est qu'ils ne traitent que le document lui-même. Tout au plus indiquent-t-ils le format du fichier (HTML... PDF...) l'adresse du site d'où la page est tirée, et de courts extraits du contenu qui renferme les mots-clés que vous avez entrés. En quelque sorte, ce sont déjà des métadonnées. Mais c'est bien pauvre. Pour en savoir plus, vous devrez charger le document entier et l'examiner. Si le résultat de la recherche porte – comme c'est fréquent – sur des centaines, voire des milliers de pages, vous concevez les limitations du principe.

Pour un document bien décrit, les métadonnées sont à peu près ce que vous trouveriez sur la fiche d'une bibliothèque : le titre, le sujet, le nom de l'auteur, l'éditeur, la date de parution, éventuellement des mots-clés, ainsi que des éléments portant sur la manière de classer le document selon un système établi : la classe Décimale Universelle, ou Dewey....

Pour avoir une utilité au niveau d'un système aussi vaste que celui de la Toile, il faut cependant s'accorder sur une manière cohérente de présenter les métadonnées. Ce fut l'objet de l'initiative Dublin Core[7].

Cette initiative a trouvé un aboutissement conforme à sa vocation : elle est devenue en février 2003 la norme ISO 15 836.

Cette norme est une chance qu'il convient de saisir. Tout site sérieux qui se destine à la diffusion de documents devrait en faire usage au travers d'un outil d'indexation utilisant des métadonnées de ce type, si possible sous forme de fichiers XML.

---

2 voir : http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm#Openarchiveinitiative

3 Dans le langage consacré de la mouvance "Open Archives" (OAI), le lieu de stockage de ressource est un "entrepôt" et la méthode de colecte des métadonnées est le OAI PMH : Open Archives Protocol for Metadata Harvesting.

4 Voir la page "licences " sur le site <Document Libre> http://www.documentlibre.org/licences.html. En fait, les documents indexés par <Document Libre> sont placés sous la licence choisie par l'auteur. Voir aussi : http://www.vecam.org/article.php3?id_article=278

5 htp://www.spip.net

6 http://www.lodel.org

7 http://www.dublincore.org

Le principe est simple : on utilise des balises, un peu comme dans le code html. On compte dans la forme de base (c'est à dire sans éléments de raffinement) quinze éléments pour décrire un document[8]. Ceci permet de préciser par exemple son titre, son auteur, le sujet qu'il traite, la période ou le lieu de référence, le public auquel de document est destiné, l'éditeur ...

L'usage des métadonnées permet d'opérer :

- une sélection thématique portant sur le sujet traité et le lieu ou la période concernée, les mots-clés – eux-même portés par un élément « subject »
- une sélection par les éléments spécifiques du document : son type (cours, mémoire...) son auteur, la date de sa publication, son format de fichier...
- une sélection portant sur le public de destination.

Toute combinaison de ces éléments de sélection est évidemment possible. Il devient alors facile de sélectionner un document du genre : un *cours* sur la *production d'énergie* au *19° siècle* pour les étudiants en *IUT/BTS.* La recherche utilise ici quatre éléments du Dublin Core : type, subject, coverage et audience dont les valeurs respectives sont dans cet exemple les termes en italique.

Il est dès lors beaucoup plus facile et plus rapide d'explorer le contenu d'une bibliothèque numérique en manipulant de petits fichiers XML de moins d'1Ko que de télécharger « pour voir » toute une série de documents entiers. Produire sur son site de telles fiches est donc un moyen très puissant d'exposer son contenu.

Les métadonnées ont par ailleurs un autre avantage lorsqu'elles sont aux normes : elles permettent de mutualiser les contenus, de manière à ce que d'un point de la Toile, on puisse avoir accès au contenu d'une multitude de sites. C'est un peu le rôle que se sont donnés les moteurs de recherche. Mais ceux-ci n'utilisent généralement pas les métadonnées, et sont donc, on le voit bien, encore bien trop peu sélectifs.

La mutualisation a-t-elle quelque chose de gênant ?

Gênant pour qui ? certainement pas pour l'utilisateur qui trouve dans un système fédéré une ouverture directe sur un très grand choix de documents.

La question se pose plus particulièrement pour les teneurs de site, en fonction des objectifs qu'ils se sont fixés. En réalité, le cas d'une action motivée uniquement par le souci d'offrir du savoir est sans doute assez rare. Il se double la plupart du temps d'un souci de communication légitime qui a été déjà évoqué. L'auteur qui agit de lui-même a besoin d'être reconnu, l'établissement expose peu ou prou son image et entretient un lien privilégié avec son personnel, ses étudiants, l'association maintient l'intérêt et la motivation de ses membres. Bref, il y a forcément une crainte à entendre parler d'un système qui viendrait en quelque sorte constituer un sur-ensemble à même de court-circuiter les éléments de communication mis en place. C'est pourtant ce qui a tendance à se produire lorsque des moteurs de recherche offrent des « *liens profonds* » qui donnent accès directement à des pages présentes sur le site sans qu'il soit nécessaire de passer par la porte d'entrée – je veux dire la page d'accueil du site. Un système bien construit doit donc respecter la personnalité d'un site et ne pas permettre un accès à ses ressources qui ferait passer le visiteur à côté des éléments de communication mis en place.

Un autre aspect du problème est de remarquer que les tendances des organismes de concertation sont d'imposer des standards en terme de formats et de procédures. Ce qui a bien fonctionné pour la recherche qui a adopté de manière concertée une attitude normative, reste encore improbable pour la production grise qui utilise en guise de standards, des formats issus des outils encore les plus largement utilisés : présentations PowerPoint, documents Word, quand ce n'est pas des formats spécialisés du genre Solidworks ou Orcad.

Enfin, on doit aussi admettre que pour bien fonctionner, une organisation étendue requiert fatalement certaines contraintes de la part de ses collaborateurs. Il faut que les auteurs se plient à un minimum de règles en terme de licences, et qu'ils portent une attention suffisante à l'indexation correcte de leurs travaux dans une classification qui ne doit son efficacité qu'à la rigueur avec laquelle elle est respectée.

Pour être réaliste, il faut aussi prévoir une certaine assistance aux auteurs et ne pas attendre de leur part une complète autonomie.

8 Pour plus de précisions, consulter par exemple ce site en français : http://savoirscdi.cndp.fr/culturepro/actualisation/metadonnees/el_dublin.htm

Les conditions d'un bon fonctionnement

- Les conditions nécessaires à une bonne valorisation de la production grise seraient donc celles-ci :
- admettre l'existence de documents aux formats multiples et les gérer comme tels,
- ne pas contourner les sites détenteurs de documents et respecter leurs caractères propres,
- apporter une solution souple et évolutive aux questions relatives aux licences, aux modes de classement, notamment en permettant aux auteurs un éventail de solutions pouvant les satisfaire,
- apporter aux auteurs l'assistance nécessaire pour les aider à établir et maintenir à jour l'indexation de leurs documents.
- ne pas ignorer les habitudes des utilisateurs et retenir que des publics différents ont des pratiques différentes. On sait que sur un même sujet, le militant d'un organisme de défense de l'environnement ne recherchera pas un article selon la même méthode ni avec es mêmes termes qu'un ingénieur du CNRS...

Un outil adapté à la production grise

L'outil d'indexation <Document Libre>[9], développé depuis quelques années, a pris en compte toutes ces exigences et est maintenant prêt à servir, en particulier depuis qu'un accord de collaboration a été signé avec l'université de Poitiers qui héberge à présent son nouveau serveur. Il s'agit d'un ensemble de scripts php associés à une base de données. Ce système se tient à disposition des sites qui utilisent ses services. Ses caractéristiques essentielles tiennent en quelques points :

- **Indépendance des collections**

Chaque site désireux de participer au réseau continue à gérer ses propres collections de documents. Le gestionnaire du site choisit le nom de cette collection et décide des documents qui en feront partie.

- **Respect des sites gestionnaires**

Il n'est pas possible d'accéder au dispositif de recherche autrement qu'en se connectant d'abord sur le site d'un gestionnaire de collection. Si la recherche ne débouche sur aucun résultat dans cette collection, un bouton d'élargissement permet cependant d'étendre cette recherche à toutes les collections du réseau.

- **Personnalisation des interfaces**

Dès qu'un gestionnaire de collection s'inscrit dans le réseau, la création de son interface de médiathèque est simple et rapide. Il peut choisir une interface standard et une feuille de style standard qu'il aménage à sa manière pour respecter par exemple sa charte graphique, mais il peut aussi utiliser le service en ligne[10] qui lui permet de choisir une interface paramétrable et de déterminer entièrement sa présentation en téléchargeant un modèle et en créant visuellement sa mise en page avec l'éditeur de dessin d'OpenOffice.org. Sitôt l'envoi (téléchargement montant) de son fichier de dessin, il reçoit automatiquement sa feuille de style css. Le gestionnaire du site peut ainsi choisir à sa guise parmi les nombreuses possibilités, le mode de recherche qu'il juge le plus approprié : de la recherche rapide par « expression régulière » à la recherche thématique multicritères... et placer dans sa page de recherche de 0 à 9 boites déroulantes et de 0 à 2 champs de texte.

- **Assistance à l'indexation**

L'auteur qui propose un document commence par le mettre en ligne sur un site auquel il a accès. S'il désire le placer dans une collection mutualisée, il remplit un formulaire en trois étapes destiné à recueillir les éléments servant à le décrire, et à l'aider à choisir la licence attachée à son document, la collection dans laquelle il désire le placer, les mots-clé sur la base d'un vocabulaire proposé (avec bien entendu la possibilité de proposer de nouveaux termes), et éventuellement une suggestion concernant le classement thématique.

Le gestionnaire de la collection sollicitée reçoit automatiquement un courrier électronique pour l'avertir. Un lien lui permet de visualiser les éléments de la nouvelle soumission. Il peut alors la valider d'un clic, ou bien la valider après quelques rectifications, notamment au niveau du classement, ou encore la refuser en adressant quelques mots d'explication à l'adresse de l'auteur.

Sitôt après, l'auteur reçoit à son tour un courrier électronique qui comporte, en cas d'acceptation, un identifiant qui deviendra celui de son document. Il devra alors renommer son document conformément à cet identifiant (ou mettre un alias). cette opération prouve l'acceptation par l'auteur du référencement de son document.

9 http://www.documentlibre.org/entree.php

10 http://www.documentlibre.org/servicecss.php

- **Mise à disposition des métadonnées**

Dès que le document est validé, une vignette est créée s'il s'agit d'une ressource graphique, et il est inscrit dans la base de données, il peut aussitôt être proposé aux internautes qui font une recherche. Un fichier de métadonnées au format XML est créé automatiquement et est disponible au bas de la fiche descriptive du document.

- **Suivi des mises à jour**

La présence en ligne des fichiers est testée quotidiennement par un automate. Si par exemple un fichier est modifié et que son indice de révision évolue, l'auteur reçoit automatiquement un formulaire pré-rempli pour recueillir les changements éventuels dans la description du document.

- **Service de requêtes hypertextuelles.**

Le service d'indexation répond à un certain nombre de requêtes types de sorte qu'il est possible d'insérer dans un document des liens hypertextes permettant par exemple de faire apparaître la production indexée d'un auteur ou des fiches bibliographiques.

- **Un vocabulaire qui s'adapte à l'utilisateur en fonction du public et du thème**

Bien que le procédé de traitement de l'indexation des documents soit identique sur l'ensemble des documents , la présentation et le vocabulaire utilisés sont adaptés à chaque situation. En réalité, tous les termes employés dans les dialogues avec l'utilisateur sont enregistrés en base de donnée, et peuvent être remplacés facilement. On peut de cette manière passer d'une langue à l'autre. Mais au delà du paramétrage de la langue, le système permet de recréer pour l'usager un univers aussi familier que possible, en se rapprochant des concepts utilisés lors de la fréquentation d'une bibliothèque au sens classique du terme. Ainsi, bien que la recherche soit fondée sur la valeur des éléments normalisés de métadonnées, l'utilisateur pourra rechercher comme dans sa bibliothèque préférée, et avec des termes propres à sa discipline, une « étagère », en fait une classe de documents ayant les mêmes attributs de sujet, de portée du sujet, et de public de destination. Ainsi une classe particulière de documents – une étagère – pourrait être appelée : « électronique analogique pour IUT/BTS ». La souplesse de ce dispositif permettra en fait aux documentalistes de le perfectionner en choisissant pour chaque champ les vocables les plus pertinents. Ainsi, l'élément « coverage » peut être appelé pays s'il s'agit de géopolitique, d'école s'il s'agit d'art plastique, d'époque s'il s'agit d'archéologie... Au total, chacun des 110 thèmes différents peut faire appel à un vocabulaire qui lui est propre.

Quand passerons-nous à la vitesse supérieure?

Si on veut bien prendre un chemin comparable à celui qu'empruntent avec succès les chercheurs, il est possible de tisser rapidement le réseau qui permet de mieux valoriser la production grise et ceci sans dénaturer la personnalité des sites qui sont déjà créés, sans perdre la maîtrise des choix de ce qu'on y présente, sans même devoir alourdir le système en place. Alors, est-ce que cela ne vaudrait pas la peine de s'y mettre ?