Environnement collaboratif pour l'ingénierie des connaissances

Jean-Marc Trémeaux
jm.tremeaux@gmail.com

Université Lyon 2
Master Recherche en Informatique, spécialité ECD
Travail personnel commun aux cours Data Mining et Apprentissage
Année 2005-2006

The Web is more a social creation than a technical one. I designed it for a social effect – to help people work together – and not as a technical toy.
Tim Berners-Lee[1]

1 Introduction

Pendant près de 15 ans, le Web a été en grande partie à sens unique. Il y avait d'une part les auteurs publiant du contenu, et de l'autre des spectateurs. Ces dernières années ont vu apparaître de nouveaux standards, pratiques et outils (weblogs, wikis) facilitant la création et la gestion de l'information, lui permettant de circuler librement entre les sites Web et d'être agrégée (services Web, flux RSS) pour servir de support décisionnel. Dans ce contexte, le rôle de l'ingénierie des connaissances n'est pas seulement d'organiser l'information d'un système, mais d'offrir aux usagers de celui-ci une architecture de participation qui favorise l'intéraction, l'utilisation d'effets de réseau afin d'accroître son utilité.

Ce travail [1] est une étude critique de l'environnement collaboratif Wiki. J'énoncerai les principes qui régissent cet outil, notamment du point de vue de l'organisation de l'information. Puis je présenterai les différents usages qui peuvent en être faits, du plus spécifique (système de gestion de contenu) au plus général (encyclopédie universelle en ligne). Enfin, je décrirai en détail la mise en œuvre qui a été faite d'un environnement collaboratif par une équipe de recherche, et je concluerai par quelques remarques sur le déploiement éventuel d'un tel système pour le Master Recherche en Informatique spécialité ECD de l'Université Lyon 2.

[1] *Sujet* : « L'ingénierie des connaissances est un domaine vers lequel tendent les nouvelles approches de l'informatique. En effet, il s'agit de concevoir des environnements où les experts humains et les machines puissantes, capables de faire des choses que l'homme est incapable de réaliser par ses moyens propres en un temps limités, s'insèrent de manière à accroitre la puissance cognitive de la société dans sa globalité. Il existe à travers le monde de grands projets qui travaillent dans cette direction. Par exemple data Grid est un environnement de partage de ressources de calcul, de stockage, de données et d'expertises pour des domaines comme la santé. L'espace collaboratif autour de Wikipedia est une autre illustration. Dans ce travail, je demande à un binôme de monter autour du site du Master de Recherche ECD un environnement Wiki pour partager les connaissances, les logiciels, échanger, bref pour capitaliser et surtout accroître les connaissances accessibles aux étudiants du Master. »

2 Caractéristiques

Un Wiki est un ensemble de pages Web que n'importe qui peut éditer. Il s'agit d'un espace d'information virtuel, non structuré à priori, auquel tout utilisateur peut contribuer. L'idée est d'abaisser les barrières à l'entrée, en fournissant des outils d'édition simples et ne nécessitant pas de connaissances préalables, afin de faciliter la coopération des individus. On espère ainsi que la qualité de la base de connaissances produite irra en s'accroissant, allant même jusqu'à surpasser celle de ce qu'un expert ou un groupe d'experts aurait pu produire.

En ce sens, le Wiki représente une modélisation informatique du processus incrémental par lequel une équipe produit de la connaissance, à l'échelle de l'internet. En contrepartie de ce mode de fonctionnement très ouvert, ce type d'environnement collaboratif propose des outils d'assurance qualité inspirés du génie logiciel : systèmes d'authentification, de journalisation et de gestion (visualisation, comparaison, restauration) des différentes versions. Je présenterai dans ce chapitre ces différentes caractéristiques.

2.1 Organisation de l'information

L'information dans un Wiki est structurée sous forme de pages reliées par des liens hypertextes. L'organisation du Web s'applique ainsi à l'échelle d'un site unique, mais il est intéressant de noter que les liens sont ici bi-directionnels. Il est donc possible de déterminer, à partir d'une page donnée, vers quelles autres pages pointe celle-ci, mais également l'ensemble des pages pointant vers elle.

Chaque page est identifiée à l'intérieur du Wiki par un nom de la forme *MasterRechercheECD*. La création d'une nouvelle page se fait simplement par l'ajout d'un lien vers celle-ci à partir d'une page existante. Une séparation stricte s'opère entre le contenu et la forme. La présentation en HTML à l'utilisateur s'effectue via une feuille de style s'appliquant à l'ensemble du site. Le contenu, quand à lui, est constitué de texte balisé dans un langage propre au Wiki, conçu pour être simple et lisible par un opérateur humain. Ainsi, un lien hypertexte s'écrit sous la forme *[[MasterRechercheECD]]*.

FIG. 1 – *Structures comparées de deux articles de Wikipedia*

Si l'on s'intéresse au contenu de Wikis existants, on constate l'émergence de structures communes dans des articles similaires. Par exemple, la figure 1 présente la table des matières de deux articles de Wikipedia concernant des universités lyonnaises. Les

deux pages ont les rubriques « Formations », « Chiffres », « Voir aussi » et « Coordonnées » en commun. Ces deux sommaires comportent cependant des rubriques qui les distinguent, une des universités dispensant par exemple des formations de science et de santé.

Contrairement à des projets conséquents comme Wikipedia, de plus petites organisations peuvent ne pas se permettre d'attendre que des standards de catégorisation émergent. Il est alors intéressant d'établir des gabarits de pages, pouvant servir de base à la création d'un article. Le rôle de l'ingénieur des connaissances est donc, plutôt que d'établir une taxonomie rigide, de catalyser les contributeurs afin de les orienter vers de bons standards de qualité.

2.2 Système d'édition

L'objectif du système d'édition est d'être intuitif, afin que la difficulté d'apprentissage soit la plus faible possible. En substance, un Wiki est une application Web et donc la connaissance de l'utilisation d'un navigateur Web est suffisante pour participer, la saisie se faisant dans un formulaire (*cf.* figure 2). L'écriture du contenu se fait sous forme de texte balisé, dont certains points sont présentés ici.

- **Structure** : les différents chapitres sont définis en encadrant le titre du chapitre par le caractère =, par exemple *=Titre de la page=*, *==Titre du chapitre==*, etc. Une ligne vide permet de séparer les paragraphes. Les éléments d'une liste sont précédés d'un astérisque ;
- **Liens hypertexte** : un lien s'effectue en encadrant le nom de la page liée par des doubles crochets, par exemple *[[PageLiée]]*. Le nom de la page liée devient alors le libellé du lien. Pour employer un autre libellé, il existe la notation alternative *[[Relation|Relation (mathématique)]]* ;
- **Mise en forme** : il est possible de mettre en valeur des parties du texte en les encadrants de doubles, triples, ou d'un nombre supérieurs d'apostrophes. D'une manière générale, les balises de mise en forme et les entités HTML sont aussi utilisables ;
- **Formules mathématiques** : une syntaxe similaire à LaTeX est disponible pour la saisie de formules mathématiques. Par exemple, la commande suivante : *$sum_{i=1}^n x_{i}$* produit la formule $\sum_{i=1}^{n} x_i$;
- **Signature** : les utilisateurs authentifiés peuvent signer leur texte en utilisant la commande triple tilde.

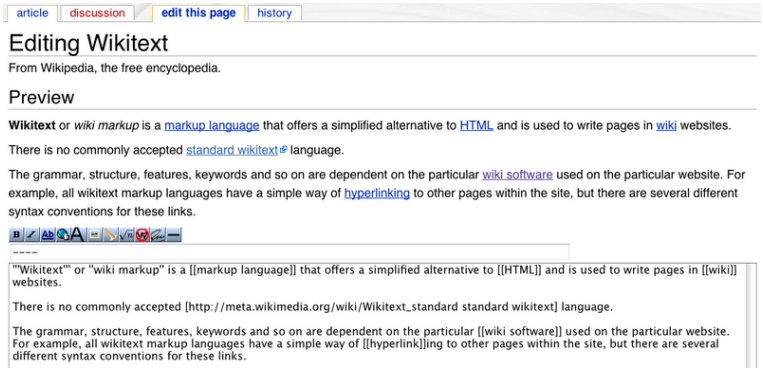

FIG. 2 – *Interface d'édition d'une page*

2.3 Gestion de versions

Afin de mettre en œuvre un procédé incrémental de production de connaissances, à la simplicité de création de contenu s'ajoute la simplicité de modification. L'intention est, là aussi, de faciliter la correction des erreurs plutôt que de rendre difficile leur introduction initiale. L'outil mis en œuvre à cette fin s'inspire des procédés de génie logiciel et s'appuie sur l'identification des acteurs et la tracabilité de leurs actions sous forme d'un système de révisions.

Chaque utilisateur est identifié par un nom d'utilisateur qu'il aura choisi préalablement lors de son inscription, ou par son adresse IP.

Chaque contribution apportée à une page donne lieu à un enregistrement daté et identifié des modifications effectuées. À chaque page est ainsi associé un journal des modifications permettant de visualiser, comparer (*cf.* figure 3) et éventuellement restaurer les versions antérieures de la page.

FIG. 3 – *Comparaison entre deux versions d'une page*

3 Différents usages

L'espace virtuel et les fonctionnalités décrites précédemment peuvent être utilisées à des fins multiples. J'identifie ici trois usages possibles, du plus spécifique au plus général : système de gestion de contenu (CMS), plate-forme collaborative en ligne et encyclopédie universelle.

3.1 Système de gestion de contenu

Un système de gestion de contenu permet la modélisation, la publication et la mise à jour de connaissances sur un support informatique, ici le Web. La publication se fait par une personne où une équipe rédactionnelle, en direction du public et sans contribution possible dans l'autre sens. On voit qu'un Wiki, restreint à un ensemble d'utilisateurs authentifiés et autorisés à effectuer des modifications, répond à cette problématique.

L'utilisation d'une application Wiki dispense l'auteur d'un site Web de connaissances particulières en langage HTML, et de programmation d'un outil spécifique

pour les opérations de publication, mise à jour et recherche d'information. Elle ne le dispense pas cependant d'un véritable travail sur la charte graphique du site, afin de la mettre en adéquation avec l'organisation représentée sur le Web. Cette mise en page, sur un Wiki, est réalisée par application d'une feuille de style à l'ensemble du site.

3.2 Plate-forme collaborative

Le deuxième usage identifié du Wiki est une véritable plate-forme collaborative, permettant aux utilisateurs d'œuvrer dans un but commun, en valorisant leurs échanges et leur permettant de tisser un véritable réseau social.

En effet, chaque utilisateur qui décide de s'inscrire sur le site dispose d'un identifiant et d'une page personnelle associée, sur laquelle il peut se présenter, décrire ses centres d'intérêt, etc. Il existe également une page décrivant dynamiquement chacune de ses contributions au cours du temps.

À chaque page du site est associée une *page de discussion*, permettant aux utilisateurs de commenter les contributions et de débattre ensemble sur les sujets sensibles. Il est à noter que chaque utilisateur contribuant à la rédaction d'une page peut ajouter celle-ci à une *liste de suivi* personnelle, lui permettant d'être notifié des interventions effectuées par d'autres personnes. Cette fonctionnalité permet non seulement de prévenir les actes de vandalisme (un auteur étant généralement soucieux des modifications apportées à son texte), mais de tracer les utilisateurs portant un intérêt à un sujet donné, afin d'établir des liens sociaux.

À ces fonctionnalités générales s'ajoutent les utilisations spécifiques qui peuvent être faites de l'espace virtuel : organisation de conférences et de réunions, échange de documents et de logiciels, forums de discussion, etc. L'environnement est alors utilisé comme un outil de communication asynchrone, tel le courrier électronique, permettant de rendre publique et de tracer l'activité.

3.3 Encyclopedie en ligne

L'usage en tant que système de gestion de contenu impliquait un petit nombre de rédacteurs, et la plate-forme collaborative une équipe de petite ou moyenne taille. L'usage décrit ici implique un passage à très grande échelle, où tout utilisateur du Web est potentiellement un utilisateur du Wiki.

Le succès de l'encyclopédie libre Wikipedia a montré l'intérêt de tels projets, avec aujourd'hui plus de 3 millions d'articles toutes langues confondues, et plus de 800 000 utilisateurs de la version anglophone. Le mode de développement de Wikipedia est basé sur celui du logiciel libre : « étant donné suffisamment de paires d'yeux, tous les bogues feront surface » (*« Given enough eyeballs, all bugs are shallow »*, formulé par E. S. Raymond). En d'autre termes, la valeur de Wikipedia réside dans son nombre d'utilisateurs.

À une échelle moindre et dans des domaines plus ciblés, d'autres encyclopédies ont été déployées. Je m'intéresse ici aux encyclopédies concernant la recherche scientifique. Qwiki [3] est un environnement collaboratif sur le thème de la physique quantique. Complexity Zoo [6] se propose de recenser toutes les classes de complexité algorithmique. Quantiki [2], hébergé par l'Université de Cambridge, est une communauté autour de l'informatique quantique.

L'intéret de créer une copie de Wikipedia à petite échelle et pour un sujet donnée peut se poser, d'autant que ce type de projet n'est pas garanti de dépasser une masse critique d'utilisateurs permettant d'engendrer un effet de réseau. Les points évoqués par les auteurs de QWiki sont les suivants :

- Fédérer une communauté d'intérêt autour d'un sujet pointu ;
- Être plus flexible du fait de la plus petite taille de la communauté ;
- Identifier les auteurs, leurs contributions et leur permettre de bâtir une réputation fiable ;
- Aborder des informations de haute technicité et encourager des articles de meilleure qualité, plutôt que d'être exhaustif.

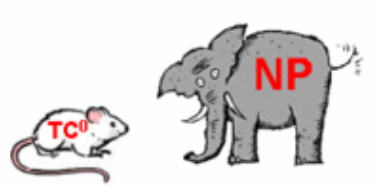

FIG. 4 – *Complexity zoo*

4 Étude de cas : TaoWebSite

L'équipe apprentissage et optimisation du LRI a, depuis Janvier 2005, remplacé son ancien site Web statique par une application en ligne de type Wiki. Dans ce chapitre, je décris les principales caractéristiques de ce nouveau [4] site Web.

L'outil utilisé est MediaWiki, un logiciel conçu et édité par la Fondation Wikimedia et utilisé par l'encyclopédie en ligne Wikipedia et de nombreux autres projets. Il s'agit d'un logiciel libre, s'exécutant sur la plate-forme LAMP (Linux, Apache, MySQL, PHP) très répandue pour les serveurs Web.

Le site Web est librement accessible sur l'internet et s'architecture autour de deux principales parties, l'une publique et l'autre restreinte à des utilisateurs authentifiés.

4.1 Partie publique

La première partie, d'accès public, a une structure fixe composée de 7 rubriques :

- page principale : présentation générale des thèmes de recherche, des collaborations industrielles et institutionnelles, des logiciels développés en interne, etc ;
- liste des membres de l'équipe, et liens vers leurs pages dédiées ;
- liste des sujets de recherche ;
- présentation détaillée des projets en cours et passés ;
- actualités ;
- offres de stages.
- liste des publications

4.2 Partie privée

La deuxième partie du site Web, en accès restreint à l'aide d'une authentification HTTP, tient lieu d'intranet de l'équipe TAO. Il s'agit là d'une véritable plate-forme collaborative. Celle-ci permet entre autres les usages suivants :

- présentation des travaux en cours en interne ;
- organisation des réunions de l'équipe : réservation des salles, planning, etc. ;
- outil de communication : discussions personnelles ou forum par petits groupes ;
- diffusion de l'information concernant les conférences ;

– comptes rendus de lecture de livres et de revues ;
– diffusion de logiciels ;
– outil de gestion administrative.

Cette partie du site se caractérise par sa structure assez désordonnée, la participation se faisant sur le mode du « tableau noir », où chaque acteur apporte une contribution suivant ses besoins. La première page est praticulièrement caractéristique du désordre apparent, et on voit assez peu de structure communes émerger. En effet, chaque personne travaille principallement sur ses pages personnelles alors que les pages communes tendent à être encombrées car personne ne prend la responsabilité d'effectuer leur nettoyage et leur organisation.

4.3 Sécurité

Pour les deux parties du site, l'édition est en accès libre. Cela a donné lieu à des actes de vandalisme perpétrés par des personnes cherchant à améliorer le référencement de leur site Web. Ceux-cis utilisent des robots afin de chercher des pages Web en écriture libre (commentaires ouverts sur les weblogs, forums, wikis) et d'y inscrire automatiquement des mots-clés afin de mieux référencer leur site dans les moteurs de recherche. On peut noter par exemple la page « Master Computer Science », ayant été vandalisée un grand nombre de fois au courant du mois de Janvier, par plusieurs robots se disputant successivement l'espace virtuel disponible. À l'heure actuelle, la plupart des pages du site public sont encombrées de liens (cachés ou non) vers des sites commerciaux.

Les réponses à apporter à ce phénomène sont multiples. À l'échelle de Wikipedia, il est considéré que les modifications non-sollicitées sont identifiées par les auteurs des articles, qui restaurent en un temps rapide une version antérieure valide. Cette stratégie de sécurité n'est pas applicable pour un site de petite taille, où chaque article possède peut-être un unique auteur, et il est donc nécessaire de restreindre les accès en écriture aux seuls utilisateurs authentifiés. Une autre voie employée est de vérifier que l'utilisateur est un agent humain par l'exécution d'une tâche simple comme la reconnaissance d'un petit texte déformé [5].

La dernière remarque concernant la sécurité du site porte sur la présence de données sensibles et nominatives dans sa partie privée. Ces données, engageant les responsables du site au regard de la loi, sont exposées aux failles de sécurité potentielles du système (logiciel MediaWiki, PHP et serveur Web) et à d'éventuelles fuites des codes d'authentification. C'est pourquoi ce type d'usage devrait être restreint au réseau local de l'organisation.

5 Conclusion

J'ai présenté dans ce travail les caractéristiques des architectures de collaboration en ligne, leurs usages possibles et une application à un type d'organisation. Un déploiement de ce type d'architecture basé sur le logiciel MediaWiki est actuellement envisagé dans le cadre du Master de Recherche en Informatique spécialité ECD de l'Université Lyon 2. Les deux premiers usages étudiés peuvent être envisagés pour une petite équipe, en tenant compte des remarques émises concernant l'organisation de l'information et la sécurité. L'usage du Wiki en tant qu'encyclopédie est beaucoup plus ambitieux, et nécessite afin d'être efficace de fédérer une vaste communauté d'utilisateurs.

Références

[1] Tim Berners-Lee. *Weaving the Web : The original design and ultimate destiny of the World Wide Web*. Collins, 2000.

[2] Quantiki. http ://cam.qubit.org/wiki/.

[3] Qwiki. http ://qwiki.caltech.edu/wiki/.

[4] TaoWebSite. http ://tao.lri.fr/.

[5] L. von Ahn, M. Blum, N. Hopper, and J. Langford. CAPTCHA : Using hard AI problems for security. In *Proceedings of Eurocrypt*, pages 294–311, 2003.

[6] Complexity Zoo. http ://qwiki.caltech.edu/wiki/complexity_zoo.