Un modèle de bibliothèque numérique collaborative – ARMARIUS

Reim Doumat, Elöd Egyed-Zsigmond, Jean-Marie Pinon

LIRIS – INSA de Lyon
7avenue Jean Capelle
69100 Villeurbanne FRANCE

{reim.doumat, elod.egyed-zsigmond, jean-marie.pinon}@liris.cnrs.fr

RÉSUMÉ. *Les manuscrits anciens numérisés représentent un contenu spécifique pour les bibliothèques numériques. Les utilisateurs travaillant sur ce type de documents ont besoin de systèmes d'assistance et d'espaces de travail collectif pour interpréter, annoter et transcrire ces manuscrits. Dans cet article, nous présenterons un modèle de bibliothèque numérique spécialement conçu pour des manuscrits anciens numérisés : Armarius. Celui-ci fournit des interfaces d'annotation manuelle et semi-automatique. Il propose également un système d'assistance pour aider l'utilisateur à annoter et à exploiter les manuscrits. De plus, il contient un espace de travail collaboratif qui permet à un groupe d'utilisateurs de travailler sur un ensemble de documents.*

ABSTRACT. *The digitized ancient manuscripts are specific content for digital libraries. Users working on this type of documents need assistant systems and a collective work spaces to interpret, annotate and transcript these manuscripts. In this paper, we present a digital library model specially designed for digitized ancients' manuscripts: Armarius. It provides manual and semi-automatic annotation interfaces. It also proposes an assistant system to help users in annotating and consulting the manuscripts. Moreover, it contains a collaborative work space that permits user groups to work on a set of documents.*

MOTS-CLÉS : *modèle collaboratif, bibliothèque numérique, assistant d'annotation, manuscrits numérisés.*

KEYWORDS: *collaborative model, digital library, annotation assistant, digitized manuscripts.*

1. Introduction

Les bibliothèques numériques proposent la consultation à distance des documents électroniques répondant à la demande des utilisateurs. Il existe plusieurs types de bibliothèques numériques telles que : les bibliothèques sur l'héritage culturel (Gallica[1], PERSEE[2]…), les bibliothèques scientifiques (BND[3], BVH[4]...), pour ne parler que de quelques initiatives.

1. http://gallica.bnf.fr

Avec la prolifération des sources d'information, la gestion et l'exploitation des collections volumineuses est devenue fastidieuse. En effet, les utilisateurs ont besoin de systèmes d'assistance pour consulter et extraire des informations intéressantes de ces collections (Ignat *et al*., 2006), d'où la nécessité de concevoir une bibliothèque numérique intégrant des outils d'assistance, de traçage (Tamine *et al*., 2006) et de personnalisation (Chevalier *et al.*, 2007) afin de répondre aux besoins des utilisateurs. En effet, ces outils facilitent la récupération, la modification, l'organisation et l'enrichissement du contenu (Amous *et al*., 2005). De plus, avec le développement des systèmes d'information collaboratifs (Nguyen, 2006), les utilisateurs sollicitent un environnement de travail collaboratif pour partager leurs connaissances. Nous avons réalisé une étude des bibliothèques numériques (Doumat *et al.,* 2007) pour synthétiser les besoins de ces bibliothèques.

Dans cet article, nous présentons un modèle d'archive vivante spécialement conçu pour des manuscrits anciens en ligne qui intègre différents rôles d'utilisateurs, et offre un environnement de travail personnalisé, assisté et collaboratif qui facilite l'échange d'informations.

Le présent article est organisé comme suit: dans le chapitre suivant, nous exposerons un état de l'art des modèles de bibliothèques numériques et les travaux concernant des bibliothèques numériques de manuscrits et d'objets anciens, nous montrerons aussi les limites de ces bibliothèques. Dans la partie 3, nous proposerons notre modèle *Armarius* qui représente une archive vivante pour les manuscrits numérisés. Nous terminerons par une conclusion et quelques perspectives.

2. Etat de l'art

Les bibliothèques numériques sont des systèmes d'information qui stockent et gèrent les documents numériques. Le modèle de ces bibliothèques est défini par la structure qui rassemble les documents numériques et leurs métadonnées et de plus des services web pour la recherche, l'annotation (Pouliquen *et al.,* 2006) et la personnalisation (Bouzeghoub *et* Kostadinov, 2005).

2.1. *Modèles de bibliothèques numériques*

Le modèle 5S (Flux, Structures, Espaces, Scenarios, Sociétés) est un modèle de base de bibliothèque numérique (Gonçalves *et al*., 2004).

Le projet DELOS est un réseau d'excellence sur les bibliothèques numériques supporté par la commission européenne (*DELOS*). Un des objectifs du projet est de

2. http://www.persee.fr
3. http://bnd.bn.pt (Bibliothèque Numérique Nationale)
4. http://www.cesr.univ-tours.fr

réaliser un modèle conceptuel d'une bibliothèque numérique (Formalizing the Design of Digital Libraries Based on UML, 2006). Le modèle de DELOS est basé sur le modèle de 5S.

Malheureusement, il n'existe pas de standard pour la structure d'une bibliothèque numérique centrée sur des images numérisées. Nous proposons donc un modèle basé sur le modèle 5S. Cependant, nous l'avons modifié selon nos besoins pour permettre à un groupe d'utilisateurs de travailler en collaboration sur un ensemble de documents qui les intéressent.

2.2. *Projets de bibliothèques numériques*

Nous nous intéressons à deux types de projets de bibliothèques numériques : les bibliothèques de manuscrits et celles avec un espace collaboratif.

En effet il existe plusieurs projets qui ont proposé des modèles de bibliothèques numériques de manuscrits et d'objets historiques telles que: BAMBI (Better Access to Manuscripts and Browsing of Images)(Calabretto *et al*., 1998), DEBORA (Digital AccEss to BOoks of theRenAissance)(Le Bourgeois *et al.,* 2001), ETANA-DL (Ravindranathan *et al*., 2004). Cependant ces projets sont limités dans la possibilité de travailler à distance, la personnalisation et l'assistance d'utilisateurs.

Il existe aussi des projets de bibliothèques numériques qui contiennent un espace collaboratif telles que: CYCLADES (Candela *et* Straccia, 2003), ADET (Alexandria Digital Earth Prototype Project)(Borgman, 2006) ou encore WiKiTUI(Wu *et al*., 2007). Le problème principal de ces applications est qu'elles ne sont pas adaptables à tous les types des documents. En effet, dans le cadre de notre travail, les documents que nous traitons sont des images scannées des manuscrits écrits à la main, rendant difficile l'utilisation des applications de reconnaissance de caractères (OCR). Pour cette raison les manuscrits doivent être essentiellement transcrits et annotés manuellement. Cependant, ce travail reste fastidieux pour une personne. D'où la nécessité d'un espace de travail collaboratif qui permet à un groupe d'utilisateurs de travailler sur le même document.

2.3. *Conclusion de l'état de l'art*

Au cours de notre étude sur les bibliothèques numériques, nous avons dégagé les problèmes de gestion des profils utilisateurs suivants:

– la plupart des projets d'annotation de manuscrits anciens ne gardent pas les traces des actions des utilisateurs. En traçant les actions des utilisateurs, on capitalise leur expérience qui peut être réutilisée par la suite par un système d'assistance;

– la plupart des bibliothèques numériques n'ont pas d'espace personnel qui permette aux utilisateurs réguliers de garder leurs recherches ou leurs collections;

– la majorité des projets de bibliothèques numériques ne contiennent pas d'espace de travail collaboratif. Cet espace permet de faciliter la communication entre différents utilisateurs, et de réaliser un travail difficilement faisable par une seule personne;

Considérant ces différentes limites, nous proposons dans la section suivante, notre modèle de bibliothèque numérique Armarius qui réunit les trois aspects suivants : la sauvegarde de traces, l'espace personnel et l'espace de travail collaboratif.

3. Armarius : modèle d'une bibliothèque numérique collaborative pour des manuscrits anciens

Le projet Armarius vise la mise en ligne d'une base de manuscrits anciens ; nous décrivons dans cette section l'architecture d'Armarius puis les interfaces de l'application et leurs fonctionnalités.

3.1. *Architecture d'Armarius*

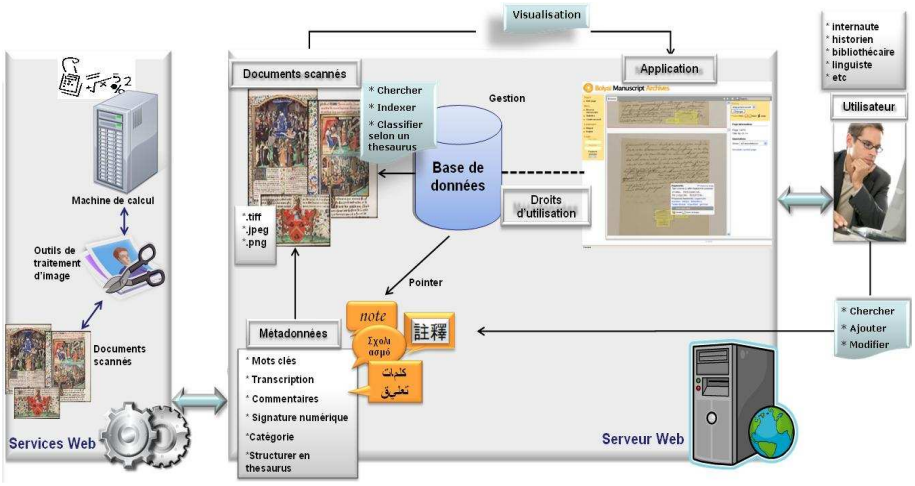
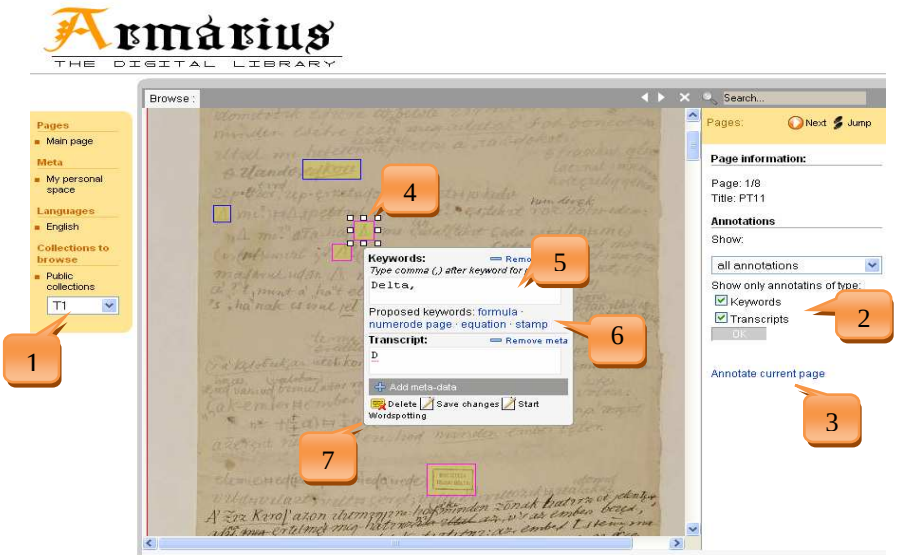


Figure 1. ***Vue globale du système Armarius***

L'architecture d'Armarius (Figure 1) est décrite à partir des composants suivants :

Collections : images numérisées des manuscrits, stockées dans une base de données relationnelle et sous différentes formes (PDF, JPEG, PNG, TIFF…). Les collections peuvent être organisées en *sous-collections* selon des critères différents (thème, date…). Les collections/sous-collections sont composées de pages et chaque page représente l'image numérisée d'une page manuscrit. Dans la base de données, nous gardons plusieurs versions d'une même image. Nous attachons des annotations à des *unités documentaires* qui peuvent désigner soit un fragment de page soit une page soit une collection.

Métadonnées: (mots-clés, commentaires, transcriptions, signatures numériques, métadonnées administratives ou descriptives…) sont crées par les utilisateurs et sont associées à une unité documentaire.



1- sélectionner la collection à visualiser; 2- filtre sur les annotations et les transcriptions visibles; 3- créer une nouvelle annotation sur la page courante; 4- dessiner un rectangle autour d'un fragment; 5- ajouter des diverses métadonnées; 6- utiliser les mots clés suggérés par le système d'assistance; 7- lancer une session wordspotting avec le fragment sélectionné.

Figure 2. *Annoter une unité documentaire dans Armarius*

Droits de lecture et de *modification :* sont définis d'un coté sur les collections et leur contenu et de l'autre coté sur les groupes d'utilisateurs.

Application web : présente un ensemble de fonctionnalités en ligne (la recherche, la visualisation, l'annotation, la transcription manuelle, la gestion des utilisateurs et des droits etc.).

Base de données : utilisé pour la gestion des images, des collections, des métadonnées, des utilisateurs et leurs groupes et des droits d'accès.

Interface : basée sur des services web pour intégrer des outils de traitement automatique d'images (wordspotting).

Utilisateurs : peuvent être de trois types différents : administrateurs du système, internautes non identifiés et utilisateurs enregistrés qui sont classés en groupes et chaque utilisateur peut appartenir à un ou plusieurs groupes. Les utilisateurs du système Armarius peuvent créer un espace personnel qui contient des collections ou des pages choisies par l'utilisateur. Le travail des utilisateurs et leurs actions sont enregistrés dans l'historique qui sera utilisé dans le système de traçage et le système d'assistance.

Dans le système Armarius, les utilisateurs se connectent au serveur de l'application pour visualiser les images des manuscrits. Ils peuvent chercher une certaine image selon des métadonnées associées. Ensuite, les utilisateurs peuvent ajouter de nouvelles métadonnées au document, (Figure 2).

3.2. *Interfaces et fonctionnalités d'Armarius*

Une partie importante de la conception du projet Armarius est constitué par les interfaces homme-machine et leurs fonctionnalités. Les utilisateurs non identifiés peuvent visualiser une démo des collections. Les autres se connectent pour accéder à leur espace personnel. Presque toutes les fonctionnalités du système sont fournies dans l'espace personnel de l'utilisateur (chercher certaines collections ou pages et les visualiser, faire une liste des favoris, tracer les actions d'un utilisateur ou groupe, ajouter des pages pour créer une collection personnelle, gérer le profil et l'espace personnel…).

L'étape la plus intéressante dans le projet Armarius survient après la visualisation d'une page de collection. L'utilisateur peut visualiser les annotations sur la page selon ses droits d'accès: voir ses propres annotations (les privées) ou les annotations collectives de ses groupes ou les annotations publiques disponibles à tous le monde. En plus l'utilisateur peut annoter de nouveaux fragments d'image/document en créant un rectangle qui représente une unité documentaire (Figure 2). Une fois que l'utilisateur a précisé son unité documentaire, une boîte de dialogue apparait pour permettre à l'utilisateur de saisir ses annotations ou d'autres métadonnées sur l'unité documentaire.

Le *système d'assistance* propose à l'utilisateur des mots clés qui sont déjà utilisés par lui-même ou par d'autres utilisateurs afin de l'aider à mieux créer ses annotations. Un autre système d'assistance propose l'utilisation des outils de traitement d'images *wordspotting* qui cherche dans le document tous les fragments qui ressemblent au fragment précisé par l'utilisateur. Sur son espace personnel, il peut ensuite visualiser les résultats qui sont terminés et les valider. Ensuite, le

système de sécurité permet à l'utilisateur de déterminer le droit d'accès à ses métadonnées en définissant sa portée (privée, collective, publique). Une autre option offerte à l'utilisateur (s'il a le droit) est de choisir une unité documentaire déjà annotée et d'en modifier les mots clés. Notre système collaboratif fournit aux utilisateurs la possibilité d'ajouter des commentaires sur le travail personnel ou celui des autres personnes.

Dans Armarius le *système de traçage* vise à tracer les démarches des utilisateurs pendant leurs recherches de documents ou pendant leur travail sur les unités documentaires afin d'intégrer ces traces dans le système d'assistance.

4. Conclusion et perspectives

Nous avons présenté dans cet article, un modèle et un prototype d'archive numérique « vivante » pour des manuscrits anciens : Armarius. Notre modèle proposé peut être utilisé dans d'autres domaines (scientifiques, médicaux…). Armarius propose également l'intégration d'une assistance utilisateurs basée sur des traces d'utilisation et un cadre de travail collaboratif. L'assistance, la collaboration et la confrontation sont particulièrement importants lors de l'annotation des manuscrits. Pour ce type de documents, les interprétations sont multiples et les outils de traitement d'images peu efficaces.

Dans nos recherches à venir, nous visons à intégrer des technologies type « push » et RSS pour suivre l'évolution de certains documents, thèmes, etc. Nous nous intéressons aussi au développement du module de collaboration pour permettre aux utilisateurs d'échanger des messages pour discuter sur le contenu. Un autre axe de recherche concerne l'enrichissement des outils de traitement d'images.

5. Bibliographie

Borbinha J., Gil J., Pedrosa G., Penas K., «The case of the digitized works at a national digital library», *Second International Conference on Document Image Analysis for Libraries, DIAL '06*, 2006, Lyon France, p. 116-125.

Borgman C. «What can Studies of e-Learning Teach us about Collaboration in e-Research? Some Findings from Digital Library Studies». *CSCW 2006*. vol.15, n°4. p. 359-383.

Bouzeghoub M., Kostadinov D., «Personalisation de l'information: apercu de l'état de l'art et definition d'un modèle flexible de profiles», *CORIA 2005*, p. 201-218.

Calabretto S., Pinon J.M., Bozzi A., « BAMBI : système de manuscrits anciens pour historiens », *Les bibliothèques numériques*, vol. 02, n° 3-4, 1998, p. 31-50.

Candela L., Straccia U. «The Personalized, Collaborative Digital Library Environment CYCLADES and Its Collections Management». *SIGIR 2003*. vol. 2924, p. 156-172.

Chevalier M., Julien C., Soulé-Dupuy C., Vallès-Paralangeau N., « Personalized information access through flexible and interoperable profiles », *Web Information Systems Engineering – WISE 2007 Workshops, Springer Berlin,* 2007, p. 374-385.

DELOS. http://www.delos.info/ (accédé Décembre, 2007).

Doumat R., Egyed-Zsigmond E., Pinon J.M., « Online ancient documents in European national libraries, a survey », *Colloque International sur le Document Electronique (CIDE10)* , 2007, p. 151-162.

Formalizing the design of Digital Libraries based on UML, présentation de DELOS, 2006, http:// old.hki.uni-koeln.de/people/herrmann/forschung/Formalizing_Design.ppt.

Gonçalves M.A., Edward Fox A., Watson L., Lipp N., « Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries», *ACM Transactions on Information Systems (TOIS)*, vol. 22, n° 2, 2004, p. 270-312.

Ignat C., Pouliquen B., Steinberger R., Erjavec T., «A tool set for the quick and efficient exploration of large document collections». *CoRR*, 2006

Ikram A., Jedidi A., Sèdes F., «A contribution to multimedia document modeling and querying», *Multimedia Tools and Applications, SpringerLink*, vol. 25, n° 3, 2005, p.391-404.

Kiamoto A., Onishi M., Ikezaki T., Deuff D., Meyer E.; Sato S., Muramatsu T., Kamida R., Yamamoto T., Ono K., «Digital bleaching and content extraction for the digital archive of rare books», *Second International Conference on Document Image Analysis for Libraries, DIAL '06* , 2006, Lyon France, p. 133-144.

Lebourgeois F., Emptoz H., Trinh E., Duong J., «Networking Digital Document Images», *Sixth International Conference on Document Analysis and Recognition (ICDAR'01), IEEE Computer Society*, Seattle, WA, USA, 2001, p. 379-383.

Nguyen A., COCoFil2 : Un nouveau système de filtrage collaboratif basé sur le modèle des espaces de communautés. Université Joseph Fourier, 2006.

Pouliquen B., Streinberger R., Ignat C., «Automatic annotation of multilingual text collections with a conceptual thesaurus», *CoRR*, 2006.

Ravindranathan U., Shen R., Gonçalves M.A., Fan W., Fox E.A., Flanagan J.W., «ETANA-DL: a digital library for integrated handling of heterogeneous archaeological data», *International Conference on Digital Libraries, ACM* , Tuscon, AZ, USA , 2004, p. 76-77.

Tamine L., Boughanem M., Zemirli N., «Inferring the user interests using the search history», *LWA 2006*. p. 108-110.

Wu A., Robinson S.J., et Mazalek A. «WikiTUI: leaving digital traces in physical books». *ACE 2007*. p. 264-265