THESE pour obtenir le grade de Docteur
de l'université Lumière Lyon II
Discipline : Science de l'information et de la communication
par

Hala Kaileh

L'accès à distance aux manuscrits arabes numérisés en mode image

Sous la direction de Monsieur Richard Bouché
Soutenue le 28 Janvier 2004 devant la Commission d'Examen

Jury : M. Richard Bouché , Directeur de thèse M. Wahid Gdoura , Rapporteur M. François Deroche, Rapporteur M. Mohamed Hassoun M. Franck LeBourgeois M. Abdelaziz Abid

Table des matières

Remerciement

Je remercie infiniment M. Richard Bouché, mon directeur de thèse, pour son encadrement, pour son aide et ses conseils durant ce travail. Qu'il trouve ici le témoignage de ma gratitude et reconnaissance.

Mes remerciements à M. François Déroche et M. Wahid Gdoura pour avoir accepter d'évaluer ce travail en tant que rapporteurs.

Je tiens à remercier chaleureusement M. Franck LeBourgeois de son acceptation comme membre du jury et de sa coopération. Grâce à son soutien, notre travail concernant l'extraction semi-automatique des métadonnées à partire des images de manuscrits arabes a pû aboutir. Qu'il trouve ici le témoigne de ma reconnaissance.

Je remercie également M. Abdelaziz Abid et M. Mohamed Hassoun d'avoir accepté de participer au jury de thèse.

J'expresse ma gratitude aux personnels du service d'informatique à l'enssib pour leur aide dans le placement de la base de données sur le serveur de l'enssib.

Je remercie également M. GuillaumeBourgois, pour sa coopération dans la création de la base de données sur SDX.

J'express ma gratitude à Mm. Marie-Geneviève Guesdon pour son aide et son conseil durant ma recherche et mes études des manuscrits arabes, Mm. Claire Ponnatau pour le relecture de ma thèse.

Enfin, je remercie mes parents, mes frères et sœurs pour leur soutien et leur aide. Je remercie également mes amies pour leur soutien.

A ma famille et à mes aimes

Résumé

Notre travail de thèse s'inscrit dans le cadre de la numérisation d'un patrimoine rare et précieux notamment les manuscrits arabes anciens. Grâce aux nouvelles technologies, l'accès à distance à ces manuscrits exige tout une chaîne de processus qui commence par la numérisation (21 manuscrits sont déjà numérisées), suivi par la création de métadonnées propres aux manuscrits arabes (173 métadonnées ont été créées). Le format XML (*eXensible Markup Language*) a été utilisé pour définir la DTD par l'aide de l'éditeur XML spy. L'objectif de la DTD est de pouvoir définir un modèle de données formel. Les métadonnées proprement dites nous ont permis de créer une base de données avec le logiciel SDX (*Système Documentaire XML*). L'utilisation d'algorithme de reconnaissance de formes, permettant l'extraction semi-automatique des métadonnées à partir des images de manuscrits arabes, est un travail original prometteur. Il a été fait en coopération avec le laboratoire RFV-INSA.

Mots clès

Base de données ; DTD ; Manuscrits arabes ; Métadonnées ; Numérisation ; Reconnaissance de formes ; SDX ; Sauvegarde du patrimoine, XML,

Summary

Our PHD thesis is situated within the domain of culture heritage digitisation, the rare and precious document mainly the Arab manuscripts. Using new technologies of information, the remote access to these manuscripts requires a series of process, beginning with the technique of digitisation (21 manuscripts are already digitised), followed by the creation of metadata specific to the Arab manuscripts (173 metadata are created). The XML (eXensible Markup Language) format and especially the XML spy editor was used to define the DTD (Document Type Definition). The objective of the DTD is to define a formal data model. Based on these metadata we were able to create a database by using the SDX (Documentary System XML) software. A plateform for the image recognition, and in particular the semi-automatic metadata extraction from the Arab manuscripts images, is an original work which gives a promising result, was developped in co-operation with the RFV laboratory of the INSA Lyon.

Keywords :

Arabic manuscript ; Culture heritage conservation ; D TD ; Database ; Digitisation ; Forme recongnition ; Metadata ; Metadata semi-automatic extraction ; SDX ; XML

Introduction générale

La révolution électronique apparue ces dernières années a bouleversé d'une manière importante la vie quotidienne des gens dans plusieurs domaines. Les domaines de l'information et de la communication en générale, et de la bibliothèque en particulier, ont été révolutionnés par les nouvelles technologies. Des changements radicaux ont bouleversé la présentation physique des documents (du papier vers l'électronique) et leur moyen d'accès (de l'accès sur place vers l'accès à distance). Les documents imprimés ainsi que les documents manuscrits ont profité de cette révolution scientifique, surtout dans leur conservation. En conséquence, les moyens de conservation du patrimoine ont pris une autre dimension, notamment grâce à la naissance et à la mise en service de la technique de numérisation. Avec cette dernière, de nouveaux moyens ont rendu plus facile et plus rapide la sauvegarde, la préservation et la diffusion des informations contenues dans ces documents. La transformation de ces derniers en documents électroniques par la technique de la numérisation a rendu la consultation des documents de toutes sortes, accessible à plusieurs chercheurs simultanément grâce à l'Internet. A partir de son domicile, l'utilisateur peut consulter à la fois les différentes bases de données et le document en texte intégral.

Le document, quel que soit son support (papier ou document multimédia), est un moyen de communication, de même qu'il est un moyen de conservation et de transmission des connaissances et du patrimoine humain.

Le terme « document électronique » est une nouvelle expression dans le domaine de l'information, dont l'apparition coïncide avec la naissance de l'Internet. Avec ces nouvelles

technologies, on assiste à l'émergence d'une nouvelle typologie de documents ayant des caractéristiques tout à fait différentes, de par leur support et leur présentation. Au cours de ces dernières années, une grande masse de documents électroniques a été diffusée grâce à l'Internet. Par conséquent, maints spécialistes de ce domaine ont dû concentrer leurs efforts sur la structuration de l'information afin de la rendre plus rapidement utilisable, sachant que les documents électroniques doivent être reformatés pour être accessibles à travers un système documentaire. Cette nouvelle structuration du document électronique a permis de le rendre plus accessible notamment grâce à l'intégration de champ spécifique telles que les métadonnées (données sur les données).

La bibliothèque qui possède des documents numériques a bien évidemment besoin des métadonnées pour gérer ces documents. Les métadonnées utilisées pour les documents numériques sont plus extensives et beaucoup plus diversifiées que celles utilisées pour la gestion de document sur papier.

Si le catalogage des manuscrits arabes sur papier a été problématique pendant le dernier siècle, celui des documents numérisés de même genre peut avoir le même problème. Cependant, bien que de nombreuses méthodes de catalogage aient été mises au point, une norme de catalogage standard pour les manuscrits arabes n'existe toujours pas. Chaque bibliothèque utilise ses propres normes. La question qui se pose ici, est la suivante : Est-il possible de créer des normes standardisées pour les manuscrits arabes numérisés? C'est dans ce contexte là que notre travail de thèse se situe. Notre problématique, qui traite de cette question, induit des problèmes tels que :

- Est-il possible de numériser les documents manuscrits arabes et avoir accès a la base de données qui contient ce genre de documents ?
- Pouvons-nous trouver facilement les informations à l'intérieur de ces documents mis en mode image ?
- Est-il possible de segmenter des informations à partir des manuscrits arabes numérisés en mode image?

Nous nous intéressons plus particulièrement à la numérisation des manuscrits arabes et à la création des métadonnées, en tenant compte des caractéristiques spécifiques de ces documents.

L'idée de cette thèse est née de l'étude que nous avons faite sur la technique de la numérisation dans le cadre de notre travail de DEA et, en particulier, pour les notes de synthèse. Les avantages de la numérisation nous ont fait réfléchir à un projet de numérisation des manuscrits arabes de Jérusalem comme un moyen de sauvegarder et d'avoir accès à ce patrimoine riche et précieux. Le bouclage imposé autour de Jérusalem a rendu l'accès à ce patrimoine très difficile, non seulement pour les chercheurs palestiniens de Cisjordanie et de Gaza mais aussi pour les chercheurs arabes et étrangers.

Pour réaliser ce projet, une étude de faisabilité a été faite dans le cadre d'un mémoire de DEA. Ce mémoire a donné lieu à deux articles : le premier a été publié en langue anglaise, dans le Journal *Manuscripta Orientalia* [1] et le deuxième en langue arabe, dans un livre, sous le titre *les bibliothèques de Jérusalem* [2] .

Le projet a pour but de numériser un fond de manuscrits arabes de Palestine et surtout ceux qui existent à Jérusalem. Selon le deuxième volume de l'étude sur les manuscrits musulmans dans le monde publiée par Al-Furqan (*Islamic Heritage Foundation*), sous le titre "*World survey of Islamic Manuscripts"* [3] , il y a environ ***21 instituts*** en Palestine qui possèdent des manuscrits islamiques en langue arabe.

Le nombre total de manuscrits dans tous ces instituts est d'environ ***11 275*** .

10 403 de ces manuscrits se trouvent à Jérusalem, parmi lesquels ***8 476*** sont gérés par les instituts palestiniens et ***1 927*** par les instituts israéliens.

Le reste (***872***) se trouve dans d'autres villes que Jérusalem : les villes de Abu Sinan, Acre, Burquayn, Hébron, Jaffa, Naplouse et Tel-Aviv.

Les objectifs du projet sont les suivants :

- La sauvegarde d'un patrimoine très riche qui est menacé du fait de la situation générale en Palestine depuis des siècles.
- La préservation de documents fragiles dont les conditions de conservation sont souvent mauvaises, la situation économique difficile des institutions propriétaires des manuscrits étant un facteur aggravant.
- Donner un accès à distance de ce patrimoine à une échelle internationale. A court terme, ceci permettra aux chercheurs palestiniens qui n'ont pas la possibilité de se déplacer, compte tenu du bouclage permanent des territoires en Palestine, de travailler sur les documents.
- Permettre l'accès à des documents qui, souvent, ne sont pas consultables pour des raisons de conservation.
- Valoriser le patrimoine palestinien en le faisant connaître dans le monde entier.

Dans le cadre de notre thèse, nous avons travaillé sur la conception d'une chaîne de numérisation de manuscrits arabes anciens pour aboutir à un prototype de bases de données permettant l'accès à distance aux documents. Un des problèmes importants sur lequel nous avons fait porter nos efforts est la définition des métadonnées et d'une DTD (Document Type Definition) permettant la description formelle des manuscrits et la prise en compte d'éléments de contenu. Un usage constant et cohérent du format XML (eXensible Markup Language*)* a présidé à la création de la base de données. Pour terminer, la conception d'algorithme de reconnaissance de formes a permis de développer des processus d'extraction semi-automatique de métadonnées à partir des

[1] Kaileh, Hala. Feasibility study for the digitization of arabic manuscripts. *The international journal for oriental manuscripts research (manuscripta orientalia)*.vol.5, no.4, December 1999. Pp. 47-57

[2] Kaileh Hala "Digitisation as a way to safeguard the manuscript of Jerusalem*» in Al-Quds Libraries.* Tunis: AFLI (Arab Federation For Libraries & Information), 2002, 146pages (article pages 87-126) (the article was published in Arabic)

[3] " World survey of Islamic manuscipts " , Geoffrey Roper. General Editor. Vol.2. London : Al-Furqan Islamic Heritage Foundation manuscripts, 1993

images de manuscrits arabes.

Le résultat d'un processus complet de numérisation peut prendre deux formes : un ensemble d'images de page (le mode image) ou un texte (le mode texte), après reconnaissance et identification des caractères. En mode image, la copie numérique reste très fidèle au document original ; dans l'autre mode, toutes les fonctionnalités des traitements de texte peuvent être mises en oeuvre sur le document. Sous le titre « environnement de recherche », la première partie de thèse traite des différentes techniques de la numérisation, des formats de compression des images, des formats de stockage des documents numériques. Les travaux effectués dans le cadre de projets de numérisation de documents anciens sont présentés et montrent les solutions mises en œuvres par différentes équipes.

Les nécessités d'un accès le plus précis possible à l'information exigent qu'on dépasse les techniques traditionnelles du catalogage pour arriver à une description plus fine des données par le moyen de métadonnées (c'est-à-dire des données sur les données). Nous avons procédé à une étude détaillée de manuscrits arabes, ce qui nous a permis d'extraire les éléments importants pour la définition des métadonnées propres à ce type de documents.

La deuxième partie de notre thèse est plus généralement consacrée à une étude des manuscrits arabes en général. D'une part, elle repose sur une étude plus ciblée d'un échantillon de vingt et un manuscrits arabes collectés dans différentes bibliothèques palestiniennes et françaises. Cette collection doit servir à la construction d'une base de données prototype. D'autre part, les avis de spécialistes de manuscrits ont été recueillis dans le cadre d'un questionnaire qui leur a été soumis. Une analyse des projets mentionnés dans la première partie apporte également un éclairage intéressant.

La troisième partie de notre thèse traite plus particulièrement de la définition et le signalement des métadonnées nécessaires à la conception de la base de données.

On y trouve la présentation des différents éléments de description, avec leur situation hiérarchique et les relations mutuelles qui les relient.

Le mode image de présentation des manuscrits numérisés ne permet pas l'accès direct à l'information par le chercheur. En coopération avec le laboratoire de reconnaissance des formes de l'INSA de Lyon, des algorithmes ont été développés qui permettent de reconnaître automatiquement les titres de chapitres, les illustrations, les décors, les cachets, etc. Les données récupérées renseignent alors les métadonnées dans la base de données des images de pages. Au lieu d'avoir à feuilleter le manuscrit, le chercheur a accès directement aux éléments qui l'intéressent : cachets ou illustrations par exemple.

La quatrième partie est plus applicative : nous présentons SDX (Système Documentaire XML) qui présente l'intérêt d'utiliser le format de données XML et de pouvoir mieux intégrer les futurs développements des traitements d'images de l'INSA de Lyon. Nous décrivons la base de données prototype, la prise en compte des 173 métadonnées définies, le chaînage avec les images de pages et l'interface d'accès par l'Internet.

Dans la conclusion finale, nous résumons le travail effectué est nous présentons des perspectives pour la généralisation du prototype.

Thèse au format PDF

Première partie. L’environnement de la recherche

kaileh_h_premiere_partie.pdf

Deuxième partie. Les manuscrits arabes

kaileh_h_deuxieme_partie.pdf

Troisième partie. Analyse des besoins des utilisateurs

kaileh_h_troisieme_partie.pdf

Quatrième Partie. L’accès à distance aux manuscrits

kaileh_h_quatrieme_partie.pdf

5. Conclusion générale et perspective

kaileh_h_cinquieme_partie.pdf

Bibliographie

kaileh_h_bibliographie.pdf

Annexes

kaileh_h_annexes.pdf