Callimaque [1]
une bibliothèque électronique sur Internet

-

Laurent Julliard, Laurent.Julliard@xerox.fr

Résumé

Callimaque est un projet de bibliothèque électronique qui propose un accès Internet à une collection de 3000 documents retraçant l'évolution des Mathématiques Appliquées et des Sciences de l'Ordinateur. Il intègre le traitement, la production de documents et la recherche d'information. Les publications sur support papier sont tout d'abord numérisées, puis stockées et indexées. Le système présente deux particularités : il mémorise la structure des documents pour une lecture plus efficace et il fournit des outils multilingues qui permettent au lecteur non francophone de mieux comprendre le résumé et d'interroger la base de données dans sa langue maternelle.

Introduction

L'origine du projet

Les collections qui sont l'objet du projet sont conservées à la Médiathèque de l'IMAG / INRIA Rhône-Alpes [2]. Elles représentent plus de quarante années des résultats de la recherche de l'IMAG. Plusieurs raisons ont présidé au choix de ces collections dans le cadre du projet Callimaque:

– les collections sont les résultats des travaux académiques : des rapports de recherche, des rapports techniques et des thèses éditées par le service de Reprographie de l'IMAG en dehors des circuits éditoriaux commerciaux. Leur diffusion ne pose donc pas de problème juridique majeur.

– les collections constituent un tout homogène. L'évolution de la recherche à l'IMAG et dans d'autres grands laboratoires mondiaux où certains chercheurs de l'IMAG ont essaimé est fidèlement traduite par les publications locales, par les thèses grenobloises (2200 titres) et par les rapports de recherche (800 documents) qui portent sur l'Analyse Numérique, les Statistiques, l'Algèbre et la Logique Combinatoire.

[1] Callimaque (320 avant JC), poète Alexandrien, grammairien et conservateur en chef de la fameuse Bibliothèque d'Alexandrie. Il a écrit plus de 800 ouvrages. Parmi les oeuvres connues qui nous sont parvenues l'une des plus importantes est le *Pinakes*, un gigantesque catalogue des oeuvres disponibles à la Bibliothèque d'Alexandrie.

[2] L'Institut d'Informatique et Mathématiques Appliquées, Grenoble et L'Institut National de Recherche en Informatique et en Automatique.

– la Médiathèque est présente sur les réseaux depuis les années quatre-vingts. Elle propose ses catalogues et certains documents au moyen de services d'information de l'Internet, notamment WAIS et le World-Wide Web, depuis que ceux-ci sont disponibles. Les collections numérisées de Callimaque seront bientôt le noyau de la bibliothèque électronique de l'IMAG. En effet, jusqu'en 1993 l'archivage des rapports de recherche et des thèses édités à Grenoble se faisait uniquement sous la forme de document papier. Leur diffusion n'étaient pas aisée et le mauvais état de certaines thèses anciennes (parfois manuscrites) ne facilitait pas la consultation. Depuis 1993, les documents de recherche de l'IMAG sont collectés directement auprès des usagers au format PostScript [3].

La constitution d'un partenariat

Le projet est développé en un partenariat qui rassemble un laboratoire de recherche privé et des institutions et des services publics de l'enseignement et de la recherche. Les cinq institutions de l'agglomération grenobloise participantes sont :

– l'IMAG (Institut d'Informatique et Mathématiques Appliquées, Grenoble)

– l'INRIA Rhône-Alpes (Institut National de Recherche en Informatique et en Automatique)

– le CICG (Centre Inter-Universitaire de Calcul de Grenoble)

– le Pôle Européen Universitaire et Scientifique

– RXRC, le Centre de Recherche Rank Xerox Grenoble

L'IMAG et l'INRIA Rhône-Alpes sont les deux instituts qui regroupent la majeure partie de la recherche en Mathématiques Appliquées et en Informatique à Grenoble, soit un millier de chercheurs. Le CICG est un centre de ressources informatiques commun aux universités de Grenoble et de Chambéry. Le Pôle Européen qui comprend douze partenaires (Universités, centres de recherche, collectivités territoriales) apporte son soutien à des actions de dimension européenne développées sur le site universitaire et scientifique grenoblois. Le Centre de Recherche Rank Xerox de Grenoble focalise une partie de ses activités de recherche et de développements avancés sur le thème des bibliothèques électroniques.

Le projet s'articule autour du produit Xerox XDOD [4], le fruit d'une collaboration entre Xerox et Cornell University. Cette coopération a débuté en 1990 et portait sur des opérations de préservation de documents anciens conservés à la Bibliothèque de l'université. Aujourd'hui, les universités américaines de Cornell, d'Indianapolis et de Yale, la bibliothèque de la ville de New-York (New-York Public Library) et les bibliothèques grégoriennes pontificales (Rome) utilisent XDOD dans des expériences apparentées.

[3] PostScript est une marque déposée de la société Adobe.

[4] Xerox Document on Demand

Les résultats attendus

Bien que basé sur un produit standard, Callimaque est un prototype à bien des égards:

1. en ce qui concernela Médiathèque elle-même, l'étude des usages de Callimaque permettra une meilleure connaissance des utilisateurs et de leurs besoins ; qu'il s'agisse de l'identification du public, des thèmes recherchés ou de l'utilisation des services : lecture sur station, ergonomie de l'interface WEB, transferts de fichiers, impressions à la demande,…
2. les collections de Callimaque intéressent l'ensemble des communautés internationales scientifiques avec lesquelles les chercheurs grenoblois développent des relations étroites matérialisées par le nombre des échanges de publications : la totalité des transactions s'évalue aujourd'hui à près d'une dizaine de milliers par an. Il est probable que la commodité des accès électroniques via le WEB intensifiera la diffusion en général et celle des documents anciens en particulier.
3. le domaine des bibliothèques électroniques est en plein essor mais les exemples de réalisations sont encore rares. Callimaque permettra de tester en vraie grandeur, l'utilisabilité d'une bibliothèque électronique et sera aussi un champ d'expérimentation pour les nombreuses technologies afférentes au document développées par l'INRIA et Xerox.

Description du projet

XDOD

Le projet Callimaque s'appuie sur la plateforme XDOD composée de stations de scanérisation reliées par réseau à un ou plusieurs serveur de fichiers et d'impression.

Les stations sont des ordinateurs compatibles IBM-PC en environnement DOS/Windows équipés d'un scanner dont la capacité de balayage atteint 20 pages par minute pour une résolution de 600 dpi. Les images numérisées sont compressées en temps réel au moyen d'une carte d'interface spécialisée permettant d'atteindre des taux de compression de l'ordre de 40 avec la méthode CCITT Groupe IV-2D. Ainsi, une page A4 numérisée à 600 dpi n'occupe que 100 à 150 Koctets. Les pages numérisées peuvent être corrigées, nettoyées ou annotées avant leur sauvegarde sur le serveur de fichiers. Il est possible de composer de nouveaux documents à partir des documents déjà numérisés.

XDOD mémorise la structure des documents (leur découpage en table des matières, chapitres, sections, index, bibliographie, glossaire, etc…). Lorsque les documents ont été numérisés et structurés, ils sont archivés, puis indexés sur le serveur de fichiers à l'aide d'un système de base de données relationnelle qui permet de rechercher un document à partir du nom de l'auteur, du titre, des mots clés, de la date de publication, de la classification, etc...

A terme, 300 000 pages seront numérisées et la capacité de stockage de Callimaque atteindra 45 Gigaoctets de mémoire de masse rapide.

Les caractéristiques de Callimaque

Le fonds documentaire traité par Callimaque est écrit en français. Seuls 800 des 3000 documents proposent un résumé en français et un résumé en anglais. Des outils destinés à faciliter la

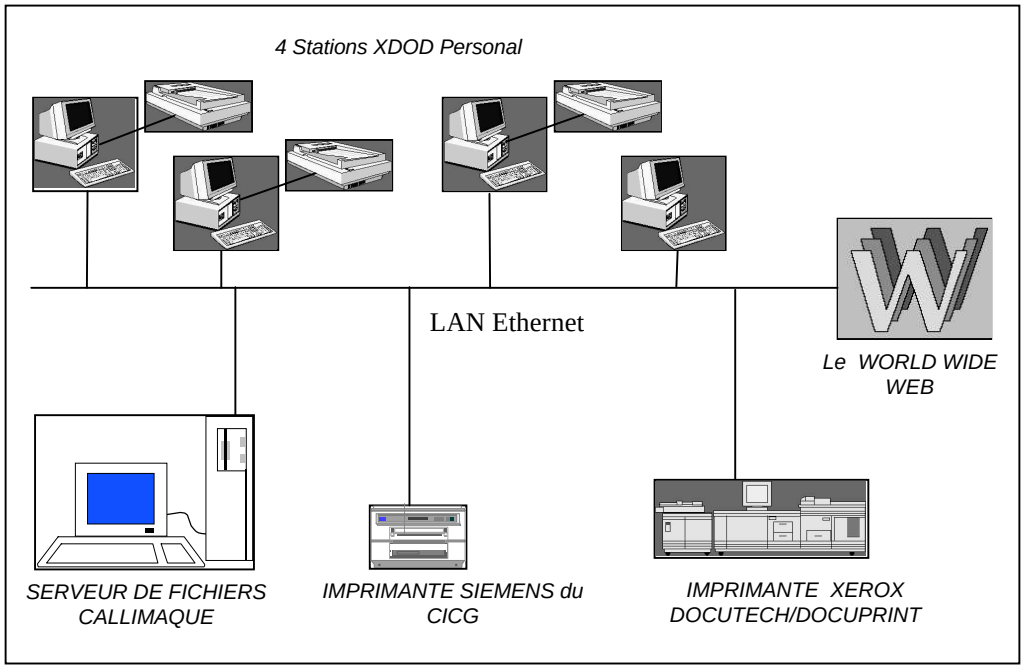

FIG. 1 – *La plateforme Callimaque*

compréhension des 2200 documents restants assisteront les usagers non francophone de l'Internet. La plateforme "Callimaque" permet à l'équipe du Centre de Recherches Rank Xerox de tester en vraie grandeur les outils d'aide à la traduction et à la compréhension de texte qui sont actuellement en cours de développement.

Un ensemble de services multilingues bâtis autour du serveur Tans (Translation Aided Network Services) seront proposés sur le serveur WEB de XDOD:

- l'aide à la traduction du Français vers l'Anglais sur les informations textuelles (résumé, titre, mots-clés). Une analyse de la phrase précisera le rôle de chaque mot : sujet, verbe, ... et une traduction tenant compte du contexte sera proposée.
- l'interrogation multilingue sera transparente (par exemple la recherche de documents contenant "network" ou "réseau" donnera un résultat identique)
- un traitement plus "naturel" des mots-clés utilisés pour effectuer une recherche dans la base (par exemple tous les verbes seront recherchés sous leur forme conjuguée, les noms et adjectifs sous leur formes "pluriel" et "singulier" et la recherche portera aussi sur une liste de mots).

La mise en place de services orientés vers la compréhension du langage fait appel à des techniques de traitement (analyse morphologique et "désambiguisation") et d'indexation de portions de textes que la plupart des bases de données relationnelles ne savent pas résoudre. Dans sa version standard, XDOD s'appuie sur une base de donnée de ce type. Pour contourner ces difficultés, les partenaires de Callimaque ont choisi de s'appuyer sur une implémentation du standard WAIS (freewais-sf) pour indexer les champs dont le contenu est textuel. Dans l'avenir la base de donnée TDB (Text database) développée par Xerox PARC pourrait se substituer a WAIS.

XDOD s'ouvre sur l'Internet

L'interface DocuWEB

Les Universités américaines de Cornell, Yale et Indianapolis qui ont adopté XDOD durant les trois dernières années ont rapidement émis le souhait que les documents numérisés deviennent accessibles au plus grand nombre. Cet accès devait présenter un caractère d'universalité et ne pas nécessiter d'investissement spécifique de la part de l'utilisateur final. L'idée d'adjoindre un serveur WEB au serveur de document XDOD s'est donc imposée naturellement.

Le serveur WEB de Callimaque est opérationnel depuis Mai 1995 et propose une interface utilisateur en français ou en anglais. Il offre des fonctions de recherche par mots-clés sur tous les attributs d'un document (titre, auteur, date de publication, mots-clés,etc…). Plusieurs "vues" d'un document sont proposées:

- une vue "catalogue" où sont présentés toutes les informations attachées au document numérisé,
- une vue de la structure du document. C'est une table des matières qui contient des liens hypertextes vers les pages numérisées,

– une vue page par page du document,

– et enfin une vue d'ensemble des pages au format timbre-poste.

De plus le serveur WEB propose un service d'impression à la demande vers des imprimantes de production assurant l'impression et la finition du document (agrafage, collage, pose d'une couverture). Enfin le serveur WEB de XDOD est administré depuis le WEB lui même via un URL spécifique dont l'accès est contrôlé par un mot de passe.

Les limites de l'interface WEB

Avec l'essor de services de plus en plus sophistiqués, les limites de l'interface utilisateur proposé par les clients WEB classiques commencent à se faire sentir. Cette constatation est particulièrement vraie dans le domaine des bibliothèques électroniques où l'interaction entre l'homme et la machine ne doit pas entraver la démarche intellectuelle du lecteur. On peut mentionner quelques unes des améliorations souhaitées:

– la transcription des formules mathématiques des résumés des thèses et rapports techniques n'est pas possible dans la version 2.0 du langage HTML. Cette lacune est comblée par la version 3.0.

– la possibilité de définir dynamiquement à la souris une zone d'une page numérisée pour en transmettre les coordonnées au serveur et lui demander d'effectuer un traitement particulier sur cette zone (par exemple procéder à une reconnaissance de caractères ou surligner une partie du texte avec une certaine couleur).

– pouvoir créer une fenêtre sur un client distant et synchroniser 2 fenêtres entre elles. Toutes ces fonctionnalités sont nécessaires si l'on veut accéder à plusieurs "vues" d'un même document et les maintenir en phase (par exemple avoir simultanément une vue d'un document sous forme de texte structuré et une autre sous forme numérisé).

– pouvoir annoter une partie d'une page WEB. Pour le moment La notion de notation est non seulement locale et mais elle est attachée à l'URL dans son ensemble.

Les versions ultérieures de HTML, l'évolution du protocole HTTP et l'avènement de langages de programmation tel que Java de SUN et son compagnon HotJava devraient combler quelques unes de ces lacunes et améliorer l'ergonomie d'ensemble sans développement spécifique aux bibliothèques électroniques.

Augmenter la bande passante

L'ergonomie n'est pas le seul problème auquel se heurtent les bibliothèques électroniques. Ces nouvelles bibliothèques sont étroitement liés aux réseaux locaux ou métropolitains puisque c'est le média par lequel elle diffuse l'information. Les documents conservés sur XDOD sont numérisés à

600 dpi, une définition qui permet de restituer pleinement la qualité d'un document. L'exploitation de ces images haute définition se heurte à 2 obstacles:

1. la définition des écrans cathodiques disponibles aujourd'hui dépasse rarement les 100 dpi [5] et il est quasiment impossible d'avoir une lecture soutenue du document dans ces conditions.

2. chaque page numérisée, même compressée, a une taille moyenne de 150 Koctets. Les débits actuels des lignes Internet ne permettent pas de transférer une telle quantité d'information sans faire chuter le temps de réponse.

Le premier point est en passe d'être résolu puisque des écrans à matrice active d'une définition de plus de 300 dpi seront utilisés en 1996 dans la cadre du projet Callimaque [6]. En ce qui concerne le deuxième point et en supposant que le passage d'une page à une autre ne doit pas prendre plus d'une seconde , la transmission de 150 Koctets en 0,5 seconde (ce qui laisse une autre demi seconde pour la décompression et l'affichage) requiert une bande passante utile de 2,4 Mbit/s soit un débit brut d'environ 3 Mbit/s. De la même façon, l'impression à la demande sur une imprimante de production est aussi un service très gourmand en bande passante. L'impression d'un document de 300 pages nécessite le transfert d'un fichier de près de 45 Moctets. Pour que le transfert se fasse en une minute il faut un débit minimum 6 à 8 Mbit/s.

Une telle bande passante reste encore assez peu courante dans les réseaux métropolitains et même si elle est disponible il ne faut pas oublier que nous parlons ici de 3 à 8 Mbit par seconde et par utilisateur. Même des liaisons ATM natives à 155 ou 622 Mbit/s pourraient devenir insuffisantes si l'usage des bibliothèques électroniques venait à se répandre. Un certain nombre de dispositifs peuvent être mis en place pour donner l'impression de feuilleter un livre même sans lorsqu'on ne dispose pas d'une bande passante de 3 Mbit/s:

– pendant que le lecteur s'attarde sur la page courante, il est possible d'anticiper le chargement des pages suivantes et/ou précédentes.

– on peut aussi avoir recours a un "serveur-cache" qui mémorisera les pages lues. Le serveur HTTP du CERN peut d'ores et déjà être utilisé comme un serveur proxy avec une capacité variable de cache. Cependant le protocole HTTP ne prévoit rien quant à l'échange d'information telle que la date de dernière mise à jour d'un URL. Aussi le serveur proxy se base-t-il sur un ensemble d'heuristiques locales pour actualiser le cache. D'autre part le serveur du CERN ne cache jamais le résultat d'un script. Ceci pose un réel problème pour des applications telles que le serveur WEB de XDOD qui s'appuie quasi-exclusivement sur des scripts pour générer dynamiquement les pages HTML.

Authentification, Autorisation, Comptabilité

L'apparition des bibliothèques électroniques et leur ouverture sur Internet ne va pas sans poser de nombreux problèmes. Le plus fréquent est sans doute celui de la gestion des droits d'auteurs et le

[5] Certains moniteurs monochromes haut de gamme atteignent 150 dpi.

[6] Ces écrans sont développés par Xerox au laboratoire Palo Alto Research Center.

respect des droits de recopie. Ce problème ne date certes pas d'aujourd'hui mais l'expansion galopante d'Internet amène les éditeurs et les bibliothécaires à se pencher sérieusement sur la question. Le projet Callimaque échappe pour l'instant à cette problématique mais l'INRIA Grenoble utilisera bientôt XDOD [7] pour numériser une partie des périodiques qu'elle reçoit et la question du contrôle des accès deviendra une préoccupation centrale.

On peut résumer les besoins en la matière par 3 mots: Authentification, Autorisation et Comptabilité. L'Authentification est le moyen par lequel on s'assure de l'identité du demandeur de service. Cette authentification repose sur un ensemble de techniques comme la transmission de signature contenant une clé publique (système PEM ou PGP) ou la transmission d'un ticket comme dans le cas du système Kerberos développé par le MIT. Une fois connue l'identité du client on pourra passer à la phase d'Autorisation à l'issue de laquelle le client se verra attribuer des droits (par exemple l'accès à un sous ensemble de collections, le téléchargement de la totalité d'un document, etc…) et aura accès à certaines ressources (impression à la demande avec finition sur une imprimante de production, faxer un document depuis le serveur, etc…). L'aspect Comptabilité permettra non seulement de garder une trace des transactions que ce soit dans un but d'audit, d'études statistiques ou de facturation. Cette facturation devra tenir compte de la valeur du bien et/ou du service délivrés (par exemple le prix d'un livre) mais aussi de la qualité de ce service. Ceci s'applique tout particulièrement dans le cadre de réseaux ATM où la nature du débit utilisé pour la transaction [8] influe sur la qualité du service rendu au client.

Le Xerox PARC travaille à la réalisation d'une "boîte noire" qui, une fois connectée à un réseau, assurera les trois fonctions d'Authentification, d'Autorisation et de Comptabilité soit au niveau TCP/IP, soit au niveau de la couche d'adaptation AAL5 d'ATM.

Conclusion

Les projets Callimaque et Calliope seront un champ d'expérimentation intéressant à bien des égards tant pour l'INRIA que pour Xerox. Les bibliothèques électroniques font appel à un très large éventail de technologies (numérisation, reconnaissance de caractères et de structures, outil d'interrogation en langage naturelle, assistance à la compréhension, visualisation de grandes quantités d'informations, interface utilisateur, réseaux à haut débit, etc…). Certaines sont plus matures que d'autres mais toutes gagneront à être confrontées à ce nouveau mode de diffusion de notre savoir qui s'apparente fortement à une autre révolution qui a eu lieu cinq siècles plus tôt: celle de l'imprimerie.

[7] Il s'agit du projet Calliope dirigé par Isabelle Allegret, Pierre Fontanille et Francois Rechenmann de l'INRIA Grenoble

[8] Voici une description outrageusement simplifiée des trois types de débits disponibles sur ATM: CBR (Constant Bit Rate) est un débit constant, VBR (Variable Bit Rate) est un débit variable qui s'adapte à la demande de l'utilisateur et ABR (Available Bit Rate) est un débit qui se contente de la bande passante disponible.