Cahiers GUTenberg

STATIQUE ET DYNAMIQUE DE DOCUMENTS MATHÉMATIQUES

Laurent GUILLOPÉ

Cahiers GUTenberg, nº 32 (1999), p. 29-34.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1999___32_29_0>

Statique et dynamique de documents mathématiques

Laurent GUILLOPÉ

Cellule MathDoc, UMS 5638
Université J. Fourier–CNRS
&
Laboratoire de Mathématiques, UMR 6629
Université de Nantes–CNRS
laurent.guillope@mathdoc.ujf-grenoble.fr

Résumé. Divers prototypes visant à permettre l'exploration d'un ensemble critique en mathématiques sont décrits. Si aucun n'arrive à répondre aux contraintes contradictoires de ce type de montage numérique, des progrès notables sont néanmoins constatés.

Mots-clés : banque de données, lecture, formule, mathématique, PDF, HTML, Toile

Malgré la prolifération de journaux dits électroniques sur la Toile, la lecture de documents sur écran reste un problème ouvert. Souvent cet affichage n'est qu'une étape indiquant une transmission réussie, permettant un aperçu sur le document, préludant à une impression qui sera le support de la lecture et du travail sur le texte. La lecture-écran reste, malgré les avancées technologiques, éloignée de la lecture-livre.

Les banques de données n'échappent pas à ces observations, en dépit du gain d'efficacité considérable que peut apporter l'exploration de leur présentation numérique pour l'utilisateur. En outre, les documents dynamiques qu'elles génèrent n'ont pas en général le confort d'appropriation des documents statiques dont l'organisation typographique a été sémantiquement peaufinée.

Cette opposition *statique/dynamique* est celle du livre (et de ses bibliothèques, lieux de mémoire) d'une part, de l'univers numérique (où les formats n'ont aucune stabilité à 5 ans — la question de l'archivage reste actuellement sans réponse) d'autre part. Maîtriser ce caractère dynamique est un élément essentiel de l'organisation du cycle de production du document numérique, comme

l'expérience de la Toile le confirme : les documents statiques d'hier ne sont bien souvent aujourd'hui plus que des liens morts.

C'est dans ce cadre général que cet article relate des travaux portant sur des bases de données mathématiques. TEX, *lingua franca* de la communauté de la recherche mathématique (entre autres), y joue un rôle pivot. Néanmoins, TEX disparaît rapidement devant les contraintes contradictoires d'un tel montage et c'est leur résolution qui nous intéresse ici. Les choix initiaux, les contraintes privilégiées donnent des produits bien différents.

1. Le cadre général et une incarnation

Soit donc une banque de données de documents mathématiques : c'est un ensemble de notices, structurées en champs, dont le contenu peut contenir des formules mathématiques codées en TEX. Ainsi, le catalogue d'une bibliothèque[1], une bibliographie commentée[2], une banque de recensions[3], un annuaire de prépublications[4] ou de thèses, un dépôt de publications électroniques[5] ou des applications mêlant des types précédents[6]. Plusieurs phases (simultanées éventuellement) caractérisent la vie de cette banque : constitution des données, indexation, traitement de requêtes sous forme d'équations de recherches, affichage des réponses.

Ce sont les deux dernières phases qui sont explorées ici, à travers l'exemple de `SemProba`, un panorama critique de la trentaine de volumes du *Séminaire de Probabilités* (parus dans la série des *Lecture Notes in Mathematics* aux éditions Springer depuis 1967, avec comme éditeur principal P.-A. Meyer) : un millier de notices, décrivant chaque contribution du séminaire à travers des champs bibliographiques élémentaires (auteur, titre, pages, ...) et deux champs plus développés (résumé synthétique et commentaire critique) qui sont de véritables rédactions (avec paragraphes, formules de mathématiques complexes, renvois à d'autres notices ou références bibliographiques). Parallèlement au travail scientifique d'écriture (coordonné par l'équipe strasbourgeoise, et qui durera plusieurs années), différents prototypes de montages numériques de

1. Cf., par ex., `www-mathdoc.ujf-grenoble.fr/bibs/ouvrages.html`.

2. Le projet `SemProba` est discuté ci-dessous, voir `www-mathdoc.ujf-grenoble.fr/SemProba`.

3. Comme *Zentralblatt-MATH*, accès en mode démonstration sur `www-mathdoc.ujf-grenoble.fr/ZMATH`.

4. Ainsi de l'index international réparti MPRESS, `mathnet.preprints.org` et de sa base française `www-mathdoc.ujf-grenoble.fr/prepub.html`.

5. Cf. le serveur de Los Alamos `xxx.lanl.gov`.

6. Le projet EULER a pour but d'offrir un point d'accès unique à des banques de documents mathématiques de type différent, cf. `www-irma.u-strasbg.fr/EMIS/EULER`.

cet ensemble ont été réalisés à partir de 1996 : ils portent aujourd'hui sur un ensemble d'environ 500 notices. Il n'y a pas de sens à lire séquentiellement la centaine de pages de textes de résumés et d'analyses qu'elles représentent : la composition numérique doit aider le probabiliste intéressé à construire ses itinéraires de lecture.

Les contraintes essentielles (et contradictoires) de l'édition de cet appareil critique sont diverses : rendu des formules mathématiques, navigation multicritères, moteur de recherche, consultation en ligne ou sur des postes hors-réseau.

Les différents montages incluent en général divers index (pour les auteurs au nombre de 300 environ, pour des classifications ou suivant la nature de l'exposé), mais d'autres parcours sont à créer pour explorer ce travail critique de 250 pages à terme. Les données sont saisies en format `plain`, avec une économie de macros et chaque installation est faite par un script `perl` (d'autres choix auraient pu être faits) qui fait appel à un module de fonctions partagées : les nouvelles installations n'ont plus essentiellement qu'à traiter des problèmes de présentation (structures de la banque de données, affichage des vues partielles sélectionnées).

2. Diverses éditions

Les différentes approches, qui incorporent toutes des index (engendrés actuellement lors de l'installation des données), diffèrent principalement par l'existence d'un moteur de recherche et les formats d'affichage proposés.

α. Le couplage `latex2html` & `FREEwais-sf`/`SFgate`

Chaque notice est transformée à l'aide de `latex2html` en un document `html` (avec ses formules mathématiques rendues par des images), cette traduction conservant la structure en champs (au besoin par des séparateurs sous la forme de commentaires `<!--RES-->`...`<!--/RES-->` quasi-invisibles). L'indexeur `waisindex` constitue alors une base de données avec des index *ad hoc* (dont la création est pilotée par un document de structure utilisant des expressions régulières convenables) : cette base de type `FREEwais-sf` est alors interrogeable par champs, la médiation entre le formulaire `html` et le protocole `wais` (ancêtre du `Z39-50`) étant assurée par `SFgate`. Un mécanisme de filtres de sortie (les `converter` de `SFgate`) permet d'affiner l'affichage des réponses (par ex., en effaçant les en-têtes communs aux notices individuelles).

Ce montage montre de manière convaincante la souplesse des outils utilisés : l'intégration dans une base de données `FREEwais-sf` de documents produits par `latex2html`[7], accompagnée par la médiation de `SFgate`, est un pari réussi. Malgré tout, l'ensemble a les faiblesses de `latex2html` (ainsi du traitement des expressions en mode mathématique, restituées en italique ou sous forme d'images, ces images dont la hauteur jure gravement parfois avec les choix de taille de police faits par l'utilisateur) et celles de `SFgate` (les filtres de sortie sont d'action limitée). Le logiciel `FREEwais-sf` lui-même est problématique à moyen terme : difficulté d'installation suivant la plate-forme et la version, incertitude quant à la maintenance et l'avenir de ce progiciel. Enfin, ce montage n'est pas envisageable sur des systèmes de type personnel (i.e. avec les systèmes d'exploitation `mac-os` ou `windows-xx`; bien que `linux` puisse être considéré désormais aussi comme un système d'exploitation individuel, le montage reste difficile).

β. Feuilletage

Pour répondre à la critique concluant le montage précédent (prérequis logiciels importants) et suivre une contrainte de grande portabilité (par exemple livraison de l'ensemble `SemProba` sur un cédérom, immédiatement consultable par l'entremise d'un navigateur), l'exigence de moteur de recherche a été abandonnée. Comme substitut, des liens hypertextuels ont été, systématiquement et sans vergogne, introduits. Ainsi, à l'intérieur de deux présentations de l'ensemble des notices (une classée par auteur, l'autre chronologiquement), des cycles de liens ont été introduits dans les différents champs à valeurs finies ; les atomes des différents index pointent vers les notices de la présentation par auteur. Par exemple, pour le mot-clé *Markov chains*, on a le cycle des notices repérées par leur numéro $2\#05 \to 8\#03 \to 8\#13 \to 10\#14 \to 10\#24 \to 12\#22 \to 14\#36 \to 2\#05$, cycle qu'on pourra pénétrer à partir de l'index des mots-clés ou l'une de ces notices dans la présentation globale par auteur. La construction de ces cheminements proposés au lecteur voudrait donner une impression de mobilité à l'intérieur de l'ensemble `SemProba` : le caractère dynamique reste néanmoins illusoire et le modèle montre rapidement ses limites, ne serait-ce qu'en révélant les insuffisances de mémoire du navigateur (une segmentation des données est possible, mais elle augmente la complexité du graphe des liens à mettre en place).

Ce montage a été réalisé d'abord en `html` brut (tous les codages TEX, et notamment les $, demeurent tel que) : au lieu d'améliorer la lisibilité de l'affichage par le recours à `latex2html`, un montage en `pdf` a été réalisé. La structure est

7. L'usage d'un autre convertisseur, tel `tex4ht` ou `HeVeA`, aurait amené à des résultats et insuffisances comparables.

la même, mais un outil supplémentaire (`pdflatex`, complété par l'extension `hyperref`) est requis pour la production et au niveau de l'affichage, l'utilisateur est supposé disposer d'un navigateur `pdf` (*a priori* largement disponible : pourtant, la première présentation publique se révéla infructueuse et les installations de visualiseur de `pdf` en *plug-in* de navigateur sont somme toute encore assez exceptionnelles dans les laboratoires visités par l'auteur).

γ. Le progiciel `edbm`

Les deux montages précédents confirment l'intérêt des contraintes initiales : un moteur de recherche avec une interface d'interrogation sélective, un format d'affichage restituant convenablement le contenu, des liens hypertextuels, une compatibilité avec des clients en environnements divers. . .

Le progiciel `edbm`[8] apparaît tout à fait propice à ce genre d'installation. Sa composante principale `edbm/w3`, à placer à l'écoute d'un serveur `http`, est un amalgame d'un moteur de recherche (basé sur la bibliothèque `db` de l'Université de Berkeley) et d'un interpréteur `python`, ce qui permet aisément l'insertion de filtres d'entrées et de sorties réalisés *ad libitum* par des scripts `python` (ces filtres pouvant, en utilisant les échanges gérés par l'interpréteur `python`, dialoguer avec des programmes complémentaires). Grâce à cette architecture du noyau, des micro-scripts peuvent être associés à chaque champ pour action éventuelle à différentes étapes du traitement de la requête d'interrogation. Le module `edbm/w3`, complété par un indexeur autonome (en cours de développement), est ainsi bien adapté à des bases relativement simples et statiques (i.e. sans mise à jour en temps réel).

Un premier montage consiste en une navigation sur index, questionnement de la base (par un formulaire guidé ou sous forme d'une interrogation libre suivant une syntaxe booléenne normalisée) et action de liens immergés dans les réponses. Des bifurcations vers des formats graphiques (`ps`, `dvi`, `pdf`, pertinents suivant l'environnement de l'usager) ou textuels spécifiques (ainsi d'un format `bibtex`, pour réutilisation dans des textes propres à l'utilisateur) sont

8. L'ensemble logiciel `edbm` (*european database manager for mathematics*) a été développé par C. GOUTORBE (Cellule MathDoc, Grenoble). Il représente une contribution majeure dans l'action de coopération franco-allemande visant à transformer la banque de données *Zentralblatt-MATH* (aujourd'hui en partenariat entre le *Fachinfomationszentrum* de Karlsruhe, les éditions Springer et la Société Mathématique Européenne) en une infrastructure à assise européenne. Comme les *Mathematical Reviews*, son concurrent créé par la Société Mathématique Américaine en 1941, l'organe de recension *Zentralblatt-MATH* répertorie depuis 1931 toutes les publications de recherche en mathématiques, soit actuellement 60 000 documents chaque année. Le module `edbm/w3` est librement disponible sous forme binaire sur les unix courants pour les abonnés de *Zentralblatt-MATH*, il est prévu ultérieurement de diffuser l'ensemble suivant les us et coutumes du *logiciel libre*.

proposées. Un deuxième montage, piloté par les mêmes outils, se déroule dans un context purement `pdf`, où réapparaissent les index, formulaires, activations de liens… Les fonctionnalités des deux montages sont évidemment légèrement différentes : la juxtaposition de formulaires naturelle dans une page `html`, n'est pas possible dans une page `pdf` ; l'affichage `pdf` du premier montage est à fin d'affichage écran et impression papier, alors que dans le second cas la lecture-écran est privilégiée…

ω. L'étape suivante ?

Dans quelques mois, un moteur TEX produisant du MathML sera certainement disponible. Il sera alors aisé de transformer le modèle précédent par ex. en une interface `XML`. Qu'en sera-t-il de la diffusion de navigateurs capables de traiter le `XML` ? Il faudra encore sans doute un peu de patience pour y compter vraiment, ainsi il y a place pour des améliorations sur les modèles précédents et d'autres prototypes…

3. Quelques points techniques

La principale difficulté est de synchroniser des environnements de nature différente et des outils qui ne sont pas habitués à coopérer. Ainsi chacun a ses propres caractères spéciaux. L'esperluette séparateur des requêtes d'interrogation `http cgi-bin/edbm_sp/SemProba/spfr.pdf?au=Weil&type=html&format=short` est à traiter soigneusement dans un texte à compiler par `pdflatex`. De même, si le # est utilisé pour marquer le numéro de volume d'une notice, c'est un caractère spécial de TEX (de catcode 6) aussi bien que pour `perl` (amorce une ligne de commentaire), ce qui demande du soin quand il intervient dans des expressions régulières. Le codage des caractères pour des données provenant d'un Mac doit être converti soigneusement. Ces détails n'ont guère d'intérêt ici. Il faut seulement remarquer que, même après les avoir réglés, l'on n'est pas assuré que l'application soit accessible à travers les navigateurs dominants du marché (ainsi surprise en août 1998 : `IE` n'avait pas de *plug-in* `pdf` en état de marche). Des tests sont toujours nécessaires.

Remerciements : le premier montage n'aurait pu avoir lieu sans l'expertise d'E. CHERHAL sur `wais` & `SFgate`, le dernier repose entièrement sur la généreuse collaboration de C. GOUTORBE. En outre, T. BOUCHE et S. RAHTZ ont apporté des éclairages décisifs durant la progression de ce projet.