Cahiers GUTenberg

DOCUMENT ÉLECTRONIQUES : UNE APPLICATION

Philippe LOUARN

Cahiers GUTenberg, nº 19 (1995), p. 121-126.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1995___19_121_0>

Documents électroniques : une application*

Philippe Louarn

Irisa/Inria Rennes, Campus de Beaulieu, F-35042 Rennes
E-mail: Philippe.Louarn@irisa.fr

Résumé. Bien que saisi sous une forme électronique, le rapport d'activité de l'Inria n'avait jamais été traité sous cette forme. Cet article décrit la procédure mise en place, s'appuyant sur la norme SGML, pour exploiter par divers vecteurs (WWW, Minitel, ftp,...) l'important volume d'information contenu dans ce rapport. Nous évoquerons les problèmes rencontrés, les apports de ce nouveau système et conclurons sur les perspectives ouvertes par ce processus.

Abstract. *Each year, Inria produces an activity report. Althoug this report is typeset in an electronic form, it was never exploited in this way. This paper describes a new process, based on* SGML*, which allows users to access to the report by different ways (*WWW*, miniel, ftp,...). Advantages and disadvantages of this process will be shown and future developments will be presented.*

1. Introduction : le rapport d'activité de l'Inria

Chaque fin d'année, l'Inria (institut national de recherches en informatique et en automatique) doit produire le rapport d'activité de l'année passée. Ce rapport est d'abord dû à des obligations légales vis-à-vis des ministères de tutelle de l'institut mais il est également très largement diffusé auprès de nos partenaires scientifiques ou industriels. C'est en outre une excellente source d'information interne qui regroupe dans un même document l'ensemble des activité menées durant une année à l'Inria.

L'Inria comporte cinq centres de recherches dispersés sur le territoire français. À ces cinq unités de recherche s'ajoute une unité de communication et information scientifique. Les projets ou actions de recherche, répartis dans six programmes, sont au nombre de soixante-dix huit, auxquels s'ajoutent quelques actions de développement et une douzaine d'équipes de services.

*. Présentation réalisée à Nanterre, le 19 janvier 1995 lors de la journée *Diffusion des documents électroniques.*

Les annexes techniques du rapport d'activité de l'Inria comportent, dans l'édition 1993, environ deux mille pages réparties en sept volumes, un pour chacun des six programmes de recherche de l'institut plus un pour les activités annexes (formation, communication, développement, moyens informatiques,...).

La saisie de chaque rapport est délocalisée dans les projets, actions ou équipes. Depuis 1987, l'outil préconisé (et très largement employé) pour la saisie du rapport d'activité (plus exactement des annexes techniques, partie du rapport qui nous intéresse ici) est LaTeX. Le lecteur trouvera dans [Louarn 88] les informations quant à la saisie/compilation des rapports des divers projets.

2. De nouveaux outils : de nouveaux besoins

Si la version papier du rapport d'activité suffit à remplir les obligations légales, il s'est avéré que cette version n'apporte pas de réponse à la question « Comment exploiter au mieux cette mine d'information ? ». L'émergence de nouveaux outils informatiques, entre autres WWW ou les éditeurs de documents structurés a mis en évidence de nouvelles possibilités d'exploitation du rapport :

- d'abord en interne, vis à vis des divers responsables (par exemple, le directeur des relations industielles souhaiterait extraire l'ensemble des partenariats industriels dans tel ou tel domaine) et également inter projets (qui, au sein de l'Inria, possède des compétences sur tel ou tel sujet ?) ;
- et vers l'extérieur, industriels, partenaires académiques, étudiants et pour une moindre part, vers le grand public.

Pour réaliser ces divers objectifs (obtention de sous ensembles du rapport, éventuellement transversaux, d'une part, et accès électronique au document par diverses voies d'autre part) il a été décidé de constituer le rapport d'activité sous forme de document structuré, conforme à la norme ISO 8879 SGML [Iso 86].

3. Structuration

Une première étape a consisté à définir un modèle de document et convertir ce modèle en DTD SGML (*document type definition*). Les rapports des projets sont très fortement structurés. Ils se décomposent en huit sections :

1. composition de l'équipe,
2. présentation générale et objectifs,
3. actions de recherche,
4. actions industrielles,

5. actions nationales et internationales,
6. diffusion des résultats,
7. bibliographie,
8. résumé en anglais,

elles-mêmes découpées en sous-sections clairement identifiées.

Les éditeurs de document SGML tels que Grif [Grif 93], [Quint 94], produit issu de l'Inria, permettent la manipulation aisée de ce document et permettent la réalisation de la première partie du cahiers des charges (obtention de sous-ensembles, création d'index, coupes transversales,...).

4. Conversions

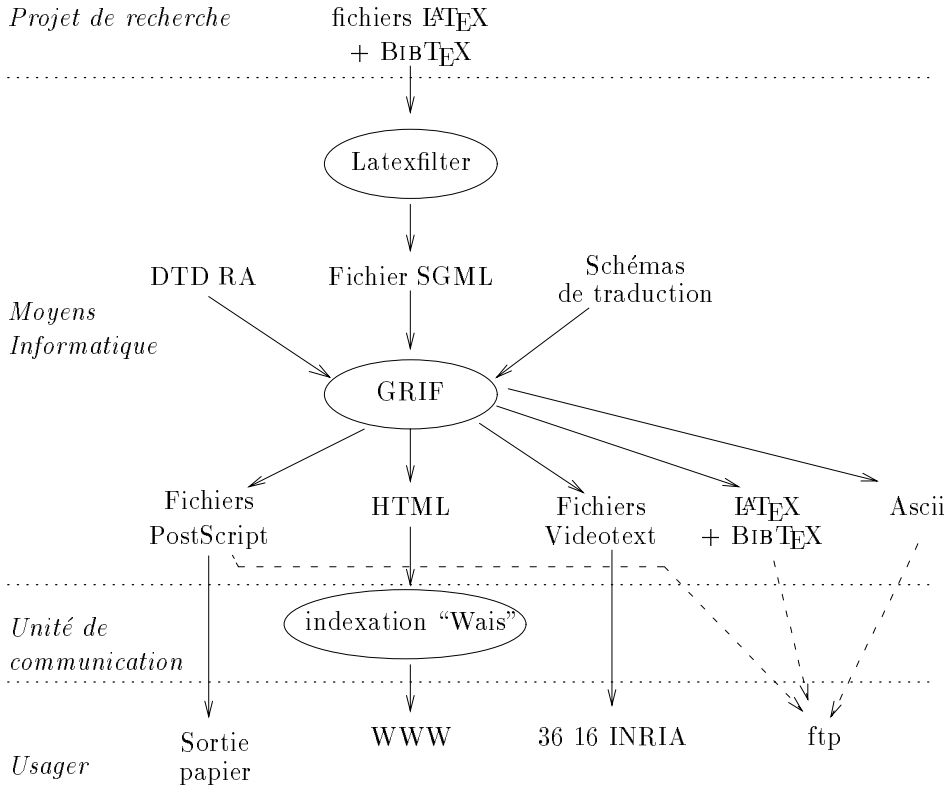

FIGURE 1 - *Architecture des processus de conversion du rapport d'activité*

Un problème rencontré est l'impossibilité de demander aux auteurs une saisie directe de leur contribution en SGML : la plupart des chercheurs de l'institut utilisent LaTeX et très peu manipulent un éditeur SGML ; de plus, les rapports 1993 ont été saisis en LaTeX... et servent fréquemment de point de départ à la rédaction du rapport de l'année suivante. C'est pourquoi un outil de conversion de LaTeX/BibTeX vers SGML est nécessaire. Nous avons demandé à la société Grif SA, qui avait déjà travaillé sur ce problème dans le cadre du projet européen Euromath de nous fournir un tel outil. Le programme latexfilter (basé sur des scripts perl et un programme lex) réalise cette conversion.

Grif dispose d'un module d'exportation de documents SGML vers tout autre langage. Divers schémas de traduction ont été réalisés : vers HTML, Ascii, Vidéotext et... LaTeX.

La passerelle LaTeX sert à valider la traduction dans l'autre sens, et aussi à refournir des fichiers .tex et .bib si des mises à jour ont eu lieu lors de l'édition SGML.

L'export HTML permet de constituer la base de documents hypertexte du rapport d'activité sur le serveur WWW de l'Inria [1]. Cette base sera ensuite indexée en mode *texte intégral* pour permettre une recherche utilisant le protocole Wais. Par ailleurs, les mots-clefs explicitement mis par les auteurs dans leur texte et les listes des participants aux diverses actions de recherche sont également indexés.

La sortie Vidéotext utilise la structuration du rapport pour n'exporter que les sections pertinentes (il serait impossible de consulter les 2 000 pages des 7 annexes techniques avec un minitel !) qui sont alors accessibles depuis le 36*16 INRIA. Cette consultation par minitel est fortement utilisée par les usagers français non (encore) connectés aux réseaux.

Par ailleurs, les versions Ascii, LaTeX et PostScript du rapport d'activité sont mises en accès *anonyme* sur le serveur ftp de l'institut [2].

5. Problèmes rencontrés

Le principal problème est lié à la trop grande rigidité de la DTD. En effet, malgré les consignes données aux auteurs, il arrive fréquement que ceux-ci prennent des libertés avec la structure du rapport : ajout de sections non prévues, inexistence de parties requises, éléments placés à un endroit du texte où il n'est pas attendu, etc. Dans ce cas, la traduction en SGML qui s'appuie sur la DTD peut mal se passer. Il faut alors modifier manuellement le source LaTeX afin de le rendre conforme au modèle.

Dans la version actuelle des outils, nous nous heurtons à quelques limitations : par exemple Grif limite la taille des clefs pour les références croisées à 6 caractères ;

1. http://zenon.inria.fr:8003/
2. ftp.inria.fr

or rien n'interdit dans le source LaTeX d'en utiliser plus pour un `\label` : seuls les 6 premiers caractères sont alors pris en compte lors de la traduction... et en découlent les erreurs que le lecteur devinera aisément. Une collaboration très active avec Grif SA permet de faire évoluer les outils et ces *bugs* devraient disparaître de la prochaine version.

HTML ne définit ni les tableaux, ni les formules mathématiques. Il faut donc créer une image lorsqu'un de ces éléments est rencontré. De plus, jusqu'à peu, seul le format GIF, gourmand en espace disque, était traité par les clients WWW. Nous avons donc du faire face à un « envahissement » de notre espace mémoire afin de stocker ces fichiers. Ce problème sera en partie résolu par l'emploi de formats d'images plus compressés, tel que JPEG.

Notons aussi que seule la version 2.09 de LaTeX est prise en compte : il faut donc mettre à jour les outils d'importation et d'exportation pour les rendre compatibles avec LaTeX2$_\varepsilon$.

6. Conclusion et prespectives

Une première expérimentation grandeur nature sur le rapport d'activité 1993 nous a convaincu de l'intérêt de l'utilisation de SGML comme langage *pivot* entre les divers systèmes de consultation/édition du rapport d'activité. L'utilisation du processus est en cours pour le rapport 1994.

Parmi les futurs développements à réaliser, nous noterons une meilleure prise en compte dans la DTD des rapports des équipes de services, qui ne suivent pas exactement le structure des rapports de projets. Des outils permettant la consutation de sous-ensemble du rapport doivent être réalisés. Par ailleurs, il faudrait affiner certaines parties de la DTD : par exemple, il est actuellement impossible dans la partie *relations internationales* de distinguer les relations avec l'Union Européenne ou les pays en voie de développement ou l'Amérique du Nord...

Un autre effort, et non des moindres, sera de convaincre les utilisateurs d'utiliser Grif. Même si ce produit n'est pas encore très répandu, les diverses passerelles, notamment vers LaTeX, permettent de rester compatible avec les standards de fait des organismes de recherche dans les domaines de l'informatique et de l'automatique.

Références bibliographiques

[Grif 93] GRIF SA, *Grif éditeur SGML*, manuel de référence du logiciel Grif, Saint Quentin en Yvelines, 1993.

[Iso 86] ISO, *Information processing - Text and office systems - Standard Generalized Markup Language*, ISO 8879, octobre 1986.

[Louarn 88] Ph. LOUARN, « Une expérience d'utilisation de LaTeX : le rapport d'activité de l'Inria », in *Les Cahiers GUTenberg* , numéro zéro, avril 1988, p. 17-24.

[Quint 94] V. QUINT, « Édition de documents structurés », in *Le traitement électronique du document*, école Inria, 3-7 octobre 1994, Aix-en-Provence, ADBS Éditions, 1994.