

Cahiers **GUT** *enberg*

☞ LATEX ET XML DANS LA CHAÎNE
ÉDITORIALE D'EDP SCIENCES

☞ Jean-Paul JORDA, Marie-Louise CHAIX, Ahmed MAHBOUB

Cahiers GUTenberg, n° 39-40 (2001), p. 167-179.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_2001__39-40_167_0>

© Association GUTenberg, 2001, tous droits réservés.

L'accès aux articles des *Cahiers GUTenberg*

(<http://cahiers.gutenberg.eu.org/>),

implique l'accord avec les conditions générales

d'utilisation (<http://cahiers.gutenberg.eu.org/legal.html>).

Toute utilisation commerciale ou impression systématique

est constitutive d'une infraction pénale. Toute copie ou impression

de ce fichier doit contenir la présente mention de copyright.

L^AT_EX et XML dans la chaîne éditoriale d'EDP Sciences

Jean-Paul JORDA, Marie-Louise CHAIX et Ahmed MAHBOUB

EDP Sciences
7 avenue du Hoggar
Parc d'Activités de Courtabœuf
F-91944 Les Ulis Cedex
<http://www.edpsciences.org>

Résumé. Dans cette présentation, nous faisons le point sur les évolutions intervenues récemment à EDP Sciences dans l'emploi de L^AT_EX et des outils associés. Nous présentons à titre d'exemple le processus de production de la revue *Astronomy and Astrophysics*. Nous décrivons également la production et l'utilisation de fichiers XML à partir des en-têtes de fichiers L^AT_EX.

Abstract. *Recent developments in the use of L^AT_EX and related tools at EDP Sciences are presented. The production process of the journal Astronomy and Astrophysics is described, along with the production and the use of XML files generated from L^AT_EX file headers.*

1. Évolutions récentes

En 1998, Marie-Louise CHAIX et Ahmed MAHBOUB présentaient, à la conférence EuroT_EX de Saint-Malo, notre expérience de T_EX dans la chaîne éditoriale [1]. Il nous a semblé intéressant de faire le point sur les évolutions intervenues dans la production des périodiques durant les trois années écoulées, tant au niveau éditorial qu'au niveau technique.

1.1. Au niveau éditorial

Deux évolutions majeures ont eu lieu récemment à EDP Sciences. La première est la diversification des domaines couverts par nos publications. Historiquement éditeur de livres et de périodiques de physique, EDP Sciences publie aujourd'hui une trentaine de revues couvrant des domaines scientifiques variés (agronomie, biologie, chimie, physique, astrophysique, mathématiques, ...). La deuxième évolution est l'augmentation considérable du nombre de pages publiées : nous produisons 45 000 pages par an depuis l'arrivée, au début de cette année, de la revue *Astronomy and Astrophysics*.

1.2. Au niveau technique

Environ la moitié des périodiques est composée avec $\text{\LaTeX} 2_{\epsilon}$. Des classes ont donc été développées pour les nouvelles revues.

Ces derniers mois ont été marqués par la mise à niveau des machines et des programmes (voir le tableau 1). Notre partenaire e-Press¹, qui compose 80% des pages, est maintenant équipé de PC sous Linux, avec des distributions basées sur $\text{\TeX}Live4$. Les bons-à-tirer définitifs sont réalisés chez EDP Sciences sur une station Sparc équipée d'une distribution identique.

TABLE 1 – Environnement de travail \TeX chez EDP Sciences.

Machines (OS)	SUN sparc Entreprise (SunOs 5.6) PC (Linux 2.2)
Distribution \TeX	$\text{\TeX}Live4$ + classes maison + configuration maison + format à la « french » ($m\text{\TeX}+keyboard$)
Correcteur d'orthographe	ispell (+ Français-GUTenberg)
dvips	dvips 5.85 et 5.50
Fontes PostScript	CM : Bluesky
Ghostscript	Aladdin ghostscript 5.5 et 6.0
Convertisseur $\text{\LaTeX} \rightarrow \text{HTML}$	latex2html 98.1p1
Production PDF	Acrobat Distiller 3.01

2. Exemple de production d'une revue : *A&A*

Afin d'illustrer l'utilisation des outils liés à \LaTeX , nous présentons ici le processus de production de la revue *Astronomy & Astrophysics (A&A)*. L'arrivée de cette revue hebdomadaire (environ 380 pages par numéro, 19 000 pages par an) nous a poussés à mieux exploiter les fichiers \LaTeX afin d'automatiser un certain nombre de tâches.

1. e-Press, 197 bd Zerktouni, Casablanca, Maroc.

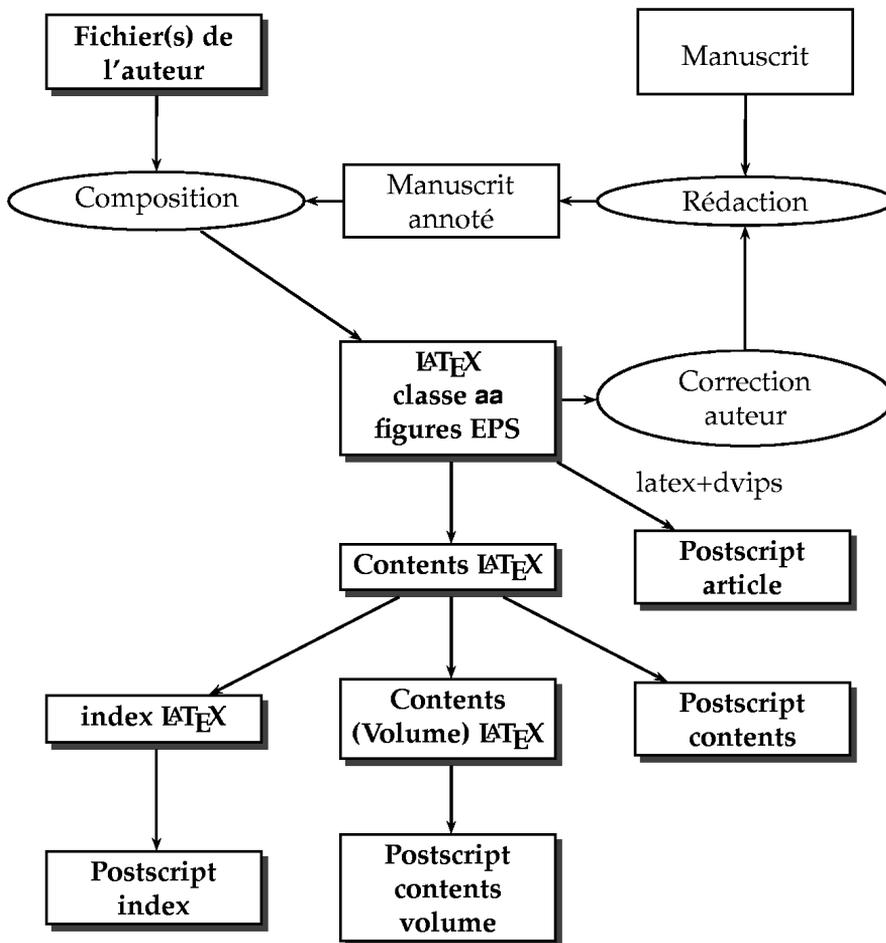


FIGURE 1 – Schéma de production de la version « papier » d'un numéro d'A&A.

2.1. Production de la version « papier »

Quand un article est accepté, le bureau éditorial nous transmet le manuscrit et les auteurs nous envoient leurs fichiers, par courrier électronique ou par FTP. Pour A&A, ce sont des fichiers \LaTeX à plus de 99,5 %. Environ 95 % des auteurs utilisent la classe aa.cls développée pour A&A. Pour leur bibliographie, les auteurs sont encouragés à utiliser le paquetage natbib.

La figure 1 résume le schéma de traitement. Le bureau éditorial transmet également les données « administratives² » de l'article (coordonnées de l'auteur, dates d'acceptation, etc.). Celles-ci sont automatiquement insérées dans notre base de gestion de manuscrits (MMS) et un DOI (*Digital Object Identifier*) est alors attribué.

Le MMS permet aux différents partenaires (auteurs, éditeurs, compositeurs) de suivre l'article dans la chaîne éditoriale. Les points d'entrée dans la base sont différents suivant la personne qui consulte : un auteur n'a pas accès aux mêmes informations qu'un rapporteur. Il est développé sous Linux avec une base de données Postgres et une interface PHP.

Après relecture du manuscrit par la secrétaire de rédaction, les fichiers des auteurs sont transmis à notre atelier de composition.

Une épreuve PDF de l'article mis en page est ensuite envoyée à l'auteur pour relecture. Au bon-à-tirer final, les secrétaires de rédaction ajoutent les dernières informations dans le fichier \LaTeX : numéro du volume, numéro de page et numéro de la section. Les DOI sont insérés automatiquement à partir des informations de la base de données.

Le fichier \LaTeX du sommaire du numéro est généré automatiquement à partir des fichiers \LaTeX des articles. Les sommaires des volumes et les index sont générés à partir des fichiers \LaTeX des sommaires des numéros.

Les fichiers PostScript produits avec *dvips 5.85* sont envoyés à l'imprimeur. Pour l'essentiel, les problèmes d'impression proviennent des figures.

2.2. Les problèmes liés aux figures

Pour chaque article d'*A&A*, nous recevons des auteurs en moyenne 10 fichiers (E)PS contenant des figures réalisées avec des logiciels les plus divers (soit environ 400 fichiers par semaine, uniquement pour cette revue). Nous détaillons ci-dessous les problèmes fréquemment rencontrés avec ces figures.

2.2.1. Les fontes

Certaines fontes exotiques sont parfois utilisées, mais pas toujours incluses dans le fichier fourni par l'auteur : elle sont substituées par du *Courier* (il est à noter que l'original envoyé par l'auteur est bien souvent lui-même imprimé avec du *Courier*!).

2. Dans le monde XML on parle de métadonnées.

2.2.2. *Les traits fins*

En deçà d'un certain seuil, l'épaisseur des traits est insuffisante pour obtenir un résultat acceptable à l'impression. Un grand nombre de figures que nous recevons contiennent de tels traits. Une sortie laser ne permet malheureusement pas de détecter ce problème, puisqu'une épaisseur minimale est fixée par l'imprimante.

Nous avons développé un « détecteur de trait fins » en Perl qui interprète les informations fournies par *ghostscript 6.0* compilé avec `-DDEBUG`. Les traits trop fins sont donc maintenant détectés dès la réception des fichiers chez EDP Sciences.

2.2.3. *Les logiciels exotiques*

Certains logiciels génèrent des fichiers PostScript qui ne sont pas interprétés par les RIP des imprimeurs. Ces fichiers sont également détectés à la réception par le script Perl.

2.2.4. *Le codage des couleurs*

Une grande variété de problèmes provient du codage employé pour les couleurs (RGB ou CMJN). Pour la version papier, le codage est le CMJN ; pour la version en ligne, les couleurs doivent être codées en RGB, mieux approprié à une lecture à l'écran.

Les auteurs fournissent de plus en plus de figures couleurs bien adaptées à la version en ligne, mais qui donnent un résultat peu satisfaisant lors d'une impression en noir et blanc : le « noir » n'est pas noir et les différents tons de gris sont indistincts. Pour l'instant, nous n'avons pas de méthode satisfaisante pour traiter ce problème.

2.2.5. *Quelques solutions aux problèmes des figures*

Pour obtenir des fichiers corrects, nous utilisons une palette de solutions parmi lesquelles :

- la modification directe des fichiers PostScript : c'est souvent le plus rapide pour résoudre des problèmes de traits fins ou de fontes ;
- le passage par Adobe Illustrator, notamment pour les conversions RGB-CMJN ;
- la numérisation de la figure, qui fournit une réponse radicale à bien des problèmes. Cette solution n'est cependant pas toujours possible, car la qualité des images scannées est parfois fortement dégradée.

2.3. Version « en ligne »

Dès que le bon-à-tirer est envoyé à l'imprimeur, nous commençons la préparation de la version Web du numéro.

2.3.1. Les résumés et le sommaire en HTML

Les pages HTML des résumés et des sommaires sont générées à partir des fichiers XML des articles (eux-mêmes produits à partir des fichiers \LaTeX , voir la section 3). Ces pages Web donnent accès aux versions complètes des articles, réservées aux abonnés. Cette version est disponible aux formats PostScript, PDF et HTML.

2.3.2. Le PostScript et le PDF

Le fichier PostScript est celui fourni à l'imprimeur. Le PDF est produit avec Distiller d'Adobe à partir du PS. Là encore, il est courant que les figures nous posent encore des problèmes. Cependant la préparation des résumés HTML et des PDF est en général rapide et ces versions sont accessibles en ligne plusieurs jours avant la distribution de la revue papier.

2.3.3. La version complète en HTML

Pour la version HTML, nous utilisons $\LaTeX2HTML$ configuré pour traiter les commandes et les environnements de la classe aa. Cependant, $\LaTeX2HTML$ a des limites et la production est loin d'être automatique. Elle nécessite un travail de préparation du fichier \LaTeX , une relecture et, bien souvent, des modifications du fichier HTML.

La version HTML de l'article peut contenir une section *Online Material*, contenant des tables et des figures non disponibles dans la version papier. Cette section est présente dans le fichier \LaTeX , mais n'est pas utilisée pour la production du PostScript.

2.3.4. Les liens avec le CDS et la NASA

Un des intérêts d'une version en ligne est la possibilité d'accéder rapidement aux documents référencés dans l'article. Pour *A&A*, nous collaborons avec le CDS (Centre de Données astronomiques de Strasbourg [4]) pour construire des liens vers leur base d'objets stellaires (SIMBAD) et la base bibliographique de la NASA. Pour construire ces liens, nous utilisons le système de nomenclature des *ref_code* développé par ces deux organismes [1, 5].

La page Web du résumé de l'article comprend un lien vers la liste d'objets stellaires cités dans l'article. Cette liste est préparée par le CDS à partir de la version papier. D'autre part, chaque occurrence d'un objet dans le résumé ou la version complète est un lien vers cet objet au CDS. Pour construire ces liens, nous exploitons le contenu des commandes `\object{}` insérées par les auteurs de l'article.

Pour plusieurs autres revues, nous collaborons avec des éditeurs pour fournir des liens réciproques dans les références. C'est notamment le cas pour *The European Physical Journal B, D, E* et *AP* et *Europhysics Letters*, pour lesquelles nous construisons des liens vers les revues de l'APS (*American Physical Society*).

3. L^AT_EX et XML chez EDP Sciences

Nous éditons plusieurs dizaines de revues dont les mises en page, les fontes et les formats sont très divers.

Le langage de balises XML nous est apparu comme une bonne solution pour stocker et manipuler de façon standard les principales informations des articles que nous publions.

Depuis janvier 2000, EDP Sciences produit donc, pour chaque article de recherche, un fichier XML contenant les informations essentielles sur la revue (titre, issn, volume, ...) et l'article lui-même (page, auteurs, résumé en anglais et/ou en français). Grâce aux nombreux outils maintenant disponibles, l'exploitation de ces fichiers bien structurés est relativement aisée. Nous nous en servons pour générer du HTML et pour alimenter une base de données (voir la section 3.2). Leur production à partir de L^AT_EX est bien plus complexe.

3.1. La production des fichiers XML à partir de L^AT_EX

Les fichiers XML sont produits à partir de fichiers L^AT_EX à l'aide d'outils développés chez EDP Sciences. Si la revue est composée en L^AT_EX, ce sont les fichiers originaux des articles qui sont utilisés. Sinon, les en-têtes des articles sont recomposés en L^AT_EX par copier-coller.

Le principe général de la transformation est présenté sur le schéma de la figure 2. Le point clé est une analyse du fichier L^AT_EX en tant que structure arborescente de données.

Cette génération nécessite parfois une intervention manuelle sur les fichiers L^AT_EX, quand des macros compliquées ont été utilisées dans le titre ou le résumé ou quand certaines règles de composition n'ont pas été respectées.

L'expression des formules mathématiques et chimiques en XML (puis en HTML) est une question ouverte. Pour l'instant, nous les laissons en \LaTeX dans des champs CDATA.

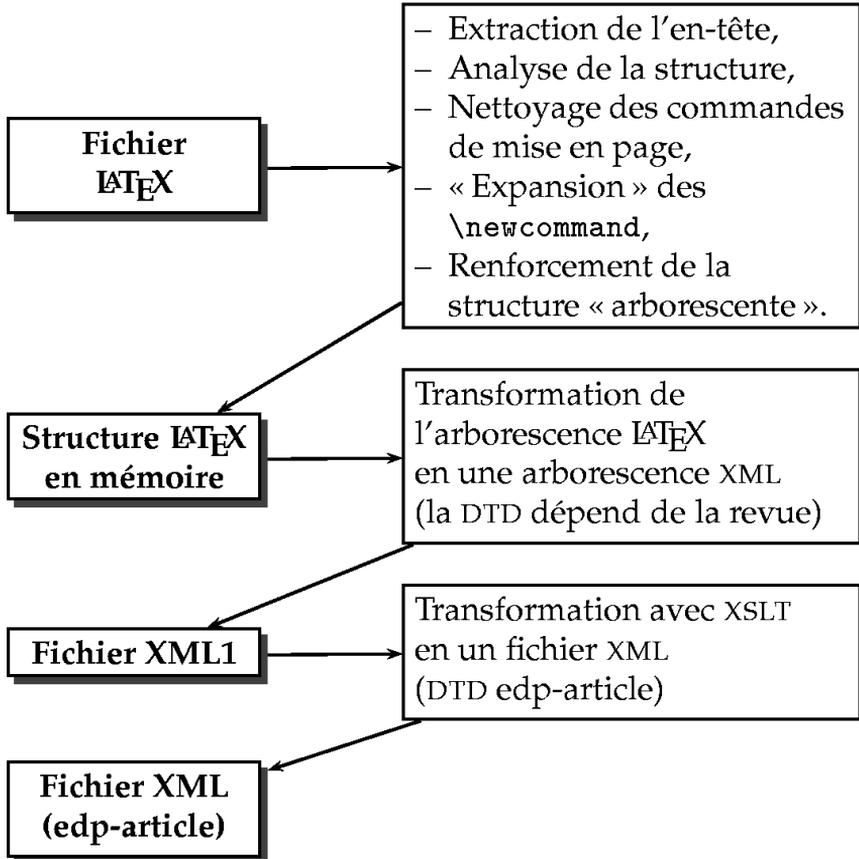


FIGURE 2 – Principe de la transformation de \LaTeX en XML.

3.2. Utilisation des fichiers XML

Les fichiers XML nous servent :

- à produire les versions HTML des résumés des articles, en accès libre ;
- à produire les fichiers HTML des sommaires ;
- à alimenter les bases de données de notre outil de recherche.

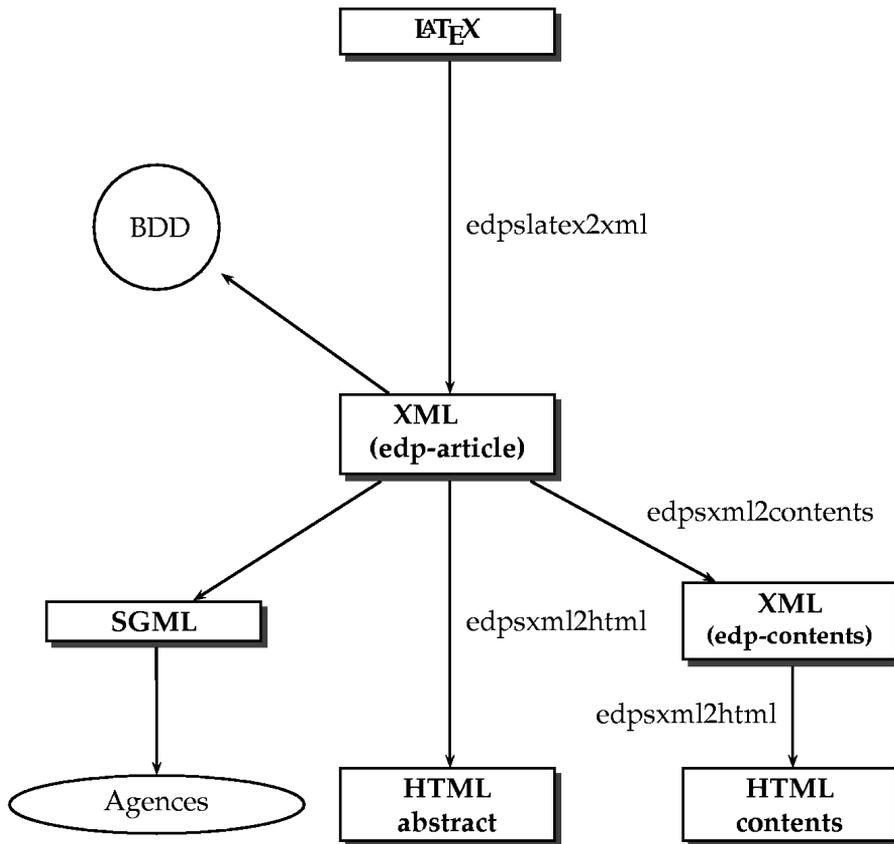


FIGURE 3 – Utilisation des fichiers XML.

D'autre part, les fichiers XML sont fournis à diverses agences indexant nos revues. À plus long terme, l'existence de ces fichiers XML nous permettra de générer rapidement des fichiers HTML avec une autre présentation. Quand les butineurs Web le permettront, ils pourront également être utilisés directement en lieu et place des fichiers HTML.

Les fichiers HTML des résumés et des sommaires sont produits à partir des fichiers XML des articles. Nous utilisons essentiellement un interpréteur XSLT (Saxon [2]). Les formules sont transformées en HTML ou en images GIF à l'aide de *L^AT_EX2HTML*, puis réintroduites dans le document HTML.

```

\documentclass{aa}
\usepackage{graphicx}
\idline{A&A 366, 558--564 (2001)}{558}
\doi{20000228}
\AASection{6}
\begin{document}

\title{The early-type multiple system QZ~Carinae}
\thanks{Based on observations collected at the European Southern
  Observatory,
  La Silla, Chile.}}

\author{P. Mayer\inst{1}
\and R. Lorenz\inst{2}
\and H. Drechsel\inst{2}
\and A. Abseim\inst{1}
}

\offprints{P. Mayer}

\institute{Astronomical Institute, Charles University,
  V Hole\v sovi\v ck'\{a}ch 2, 180 00 Praha 8, Czech
  Republic
\and Dr. Remeis-Sternwarte, Astronomisches Institut der
  Universit\ "at Erlangen-N\ "urnberg, Sternwartstra\ss e 7,
  \\\96049 Bamberg, Germany }

\date{Received 27 June 2000 / Accepted 7 November 2000}

\titlerunning{Multiple system QZ Carinae}

\abstract{
We present an analysis of the early-type quadruple system QZ~Car,
consisting of an eclipsing and a non-eclipsing binary.
[...]
An ephemeris for the eclipsing binary QZ~Car valid at present was derived
Prim.\, Min. = hel. JD\,\, $2448687.16 + 5\fd9991 \cdot E$. The
relative orbit of the two binary constituents [...]

\keywords{binaries: eclipsing -- binaries: spectroscopic --
  stars: early-type -- stars: individual: QZ Car}}

\end{document}

```

FIGURE 4 – Exemple d’un en-tête de fichier \LaTeX pour un article d’*A&A*.

```

<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<!DOCTYPE edp-article SYSTEM "/home2/web/XML/DTD/edp-article/edp-article2.dtd">
<edp-article>
  <journal-id>
    <j-title>Astronomy and Astrophysics</j-title>
    <j-shorttitle>A&#038;A</j-shorttitle>
    <j-edpsname>aa</j-edpsname>
    <issn-paper>0004-6361</issn-paper>
    <issn-elec>1432-0756</issn-elec>
  </journal-id>
  <issu-id>
    <volume>366</volume> <issue>2</issue> <year>2001</year>
    <month>FebruaryI</month>
  </issu-id>
  <article-id>
    <edps-ref>aah2289</edps-ref> <doi>10.1051/0004-6361:20000228</doi>
    <history>Received 27 June 2000 / Accepted 7 November 2000</history>
    <section>6</section>
    <first-page>558</first-page> <last-page>564</last-page>
    <subject>
      <keyword>binaries: eclipsing </keyword>
      <keyword>binaries: spectroscopic </keyword>
      <keyword>stars: early-type</keyword>
      <keyword>stars: individual: QZ Car</keyword>
    </subject> <copyright>ESU</copyright>
  </article-id>
  <article-title>
    <title>The early-type multiple system QZ Carinae</title>
    <other-tricks>Based on observations collected at the European Southern
      Observatory, La Silla, Chile.</other-tricks>
  </article-title>
  <author>
    <author-name>P. Mayer</author-name>
    <address>Astronomical Institute, Charles University,
      V Hole&#353;ovi&#269;k&#225;ch 2, 180 00 Praha 8, Czech Republic</address>
  </author>
  <author>
    <author-name>R. Lorenz</author-name>
    <address>Dr. Remis-Sternwarte, Astronomisches Institut der
      Universit&#228;t Erlangen-N&#252;rnb&#229;rg, Sternwartstra&#223;e 7,
      96049 Bamberg, Germany </address>
  </author>
  [...]
  <correspondence>P. Mayer</correspondence>
  <abstract>
    We present an analysis of the early-type quadruple system QZ Car,
    consisting of an eclipsing and a non-eclipsing binary. [...]
    An ephemeris for the eclipsing binary QZ Car valid at present was derived
    Prim. Min. = hel. JD <formula notation="TeX">E$2448687.16 + 5\fd9991 \cdot E$</formula>.
    The relative orbit of the two binary constituents [...]
  </abstract>
</edp-article>

```

FIGURE 5 – Exemple de fichier XML pour un article d'A&A.

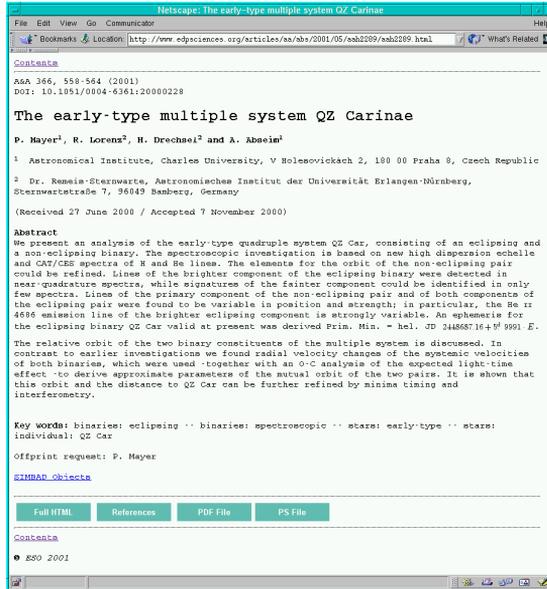


FIGURE 6 – Page HTML pour un article de *A&A*, produite à partir du fichier XML de la figure 5.

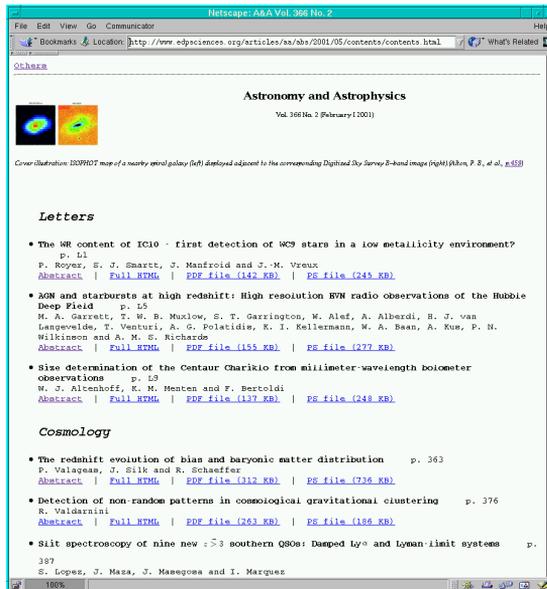


FIGURE 7 – Page HTML pour un sommaire de *A&A*, produite à partir des fichiers XML.

Conclusion

Nous avons essayé d'illustrer qu'un fichier L^AT_EX n'est plus seulement exploité pour fabriquer un document imprimé de très grande qualité : il est de plus en plus utilisé comme un langage de balises permettant de structurer des informations. Ceci a plusieurs conséquences :

- certaines règles de composition doivent être respectées même si elles ne sont pas nécessaires pour obtenir un document papier correct (auteurs séparés par des `\and`, par ex.) ;
- le fichier contient des commandes qui n'ont aucun effet sur la version papier, mais qui sont utiles à la production des formats XML ou HTML (`\object{}`, `\AASection{}`, ...).

Ceci modifie les points de repère des différents acteurs qui participent à la composition du fichier L^AT_EX :

les auteurs qui doivent réfléchir à la version papier *et* à la version en ligne de leur article ;

les secrétaires de rédaction qui doivent intégrer la version en ligne dans leur champ de préoccupations et ajouter des informations dans le fichier L^AT_EX ;

les compositeurs qui ne peuvent plus se baser uniquement sur la sortie papier pour vérifier que leur fichier est correct ;

les développeurs qui doivent faire évoluer les classes L^AT_EX pour que le document soit convenablement structuré dans l'optique de sa transformation en XML.

Bibliographie

- [1] Marie-Louise MUNIER et Ahmed MAHBOUB, « Expérience de T_EX (L^AT_EX) dans la chaîne éditoriale », *Actes de la X^e conférence T_EX européenne, Cahiers GUTenberg*, n° 28-29, mars 1998.
- [2] <http://users.iclway.co.uk/mhkay/saxon/index.html>
- [3] Version en ligne de la revue *Astronomy and Astrophysics* : <http://www.edpsciences.org/aa/>
- [4] Le site du Centre de Données astronomiques de Strasbourg (CDS) : <http://cdsweb.u-strasbg.fr/>
- [5] *NED and SIMBAD Conventions for Bibliographic Reference Coding* <http://cdsweb.u-strasbg.fr/simbad/refcode/refcode-paper.html>