Tralics, a LaTeX to XML translator

José Grimm

Jose.Grimm@sophia.inria.fr

INRIA

Outline

- What is the raweb?
- Why is a translator needed?
- Other tools for the raweb
- The DTD : simplier than DocBook, simplier than TEI
- LTX2X and Perl are good for a prototype
- Tralics is a C++ program
- The p element, and others
- Architecture
- Examples

The raweb

- Inria's Annual Activity Report (in French)
- Paper version via LaTeX since 1987
- On the web since 1994, see http://www.inria.fr
- SGML in 1994, Latex2html since 1995
- Last paper version in 1998
- On CRrom since 1999
- Uses XML since 2001 (experimental)

pub

Inria Sophia has 20 years

Sizes of the Raweb

Sizes (Mb)

year	total	html	ps.gz	pdf
1994		14	26	
1996		23	33	
1996		23	33	
1997		26	43	39
1998		46	53	40
1999	244		77	48
2000	222		68	50
2001	253		89	40
2002	338		147	114

Raweb2002

- 125 Teams
- 3000 pages, A4 size, 10pt font
- 3890 HTML pages
- XML version avalaible on the ftp server

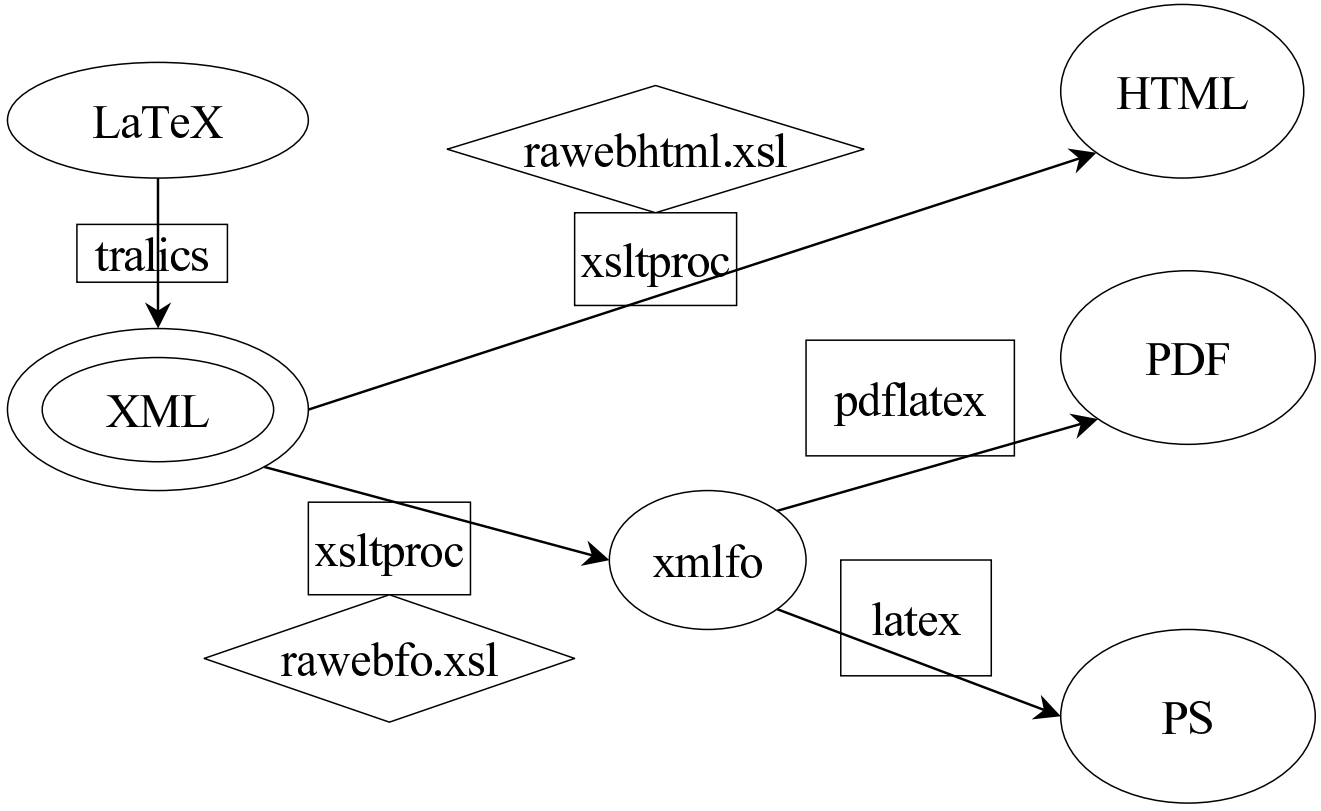
  ftp://ftp.inria.fr/scratch/RAWEB/RAWEB02/XML/

Schedule of the raweb2002

11 Apr 2002: Decision to use XML
June 2002: Definition of the DTD
Aug 2002: Finalization of the tools
22 Aug 2002: Announcement
Oct-Nov 2002: Writing and review of the texts
15 Dec 2002: Dead line for first version
12 Jan 2003: Dead line for final version
14 Feb 2003: Creation of Html, PS, Pdf
21 Feb 2003: Creation of the master and cover of the CDrom
03 Mar 2003: Installation on the Web
14 Mar 2003: Issue of the CDrom

The XML route

Les étapes du traitement du Raweb



Requirements

- No change in the input
- No change in the PS output
- No change in the HTML output
- Must be available on Unix (Dec, Linux, Sun, ...)
- XML could be produced by hand (or Word...) $\Longrightarrow$ simple DTD
- Input split in “modules”, one HTML page per module
- Specific commands like `\pers`, env. like `glossaire`
- Need for beautiful math, lots of graphics, pseudo-verbatim code.
- Will be in English in 2003

Difficulties

- No direct route from XML to PDF
- Use of `xmltex` by Carlisle and others
- Use of `fotex` by Rahtz and others
- Use of `pstoimg` (part of latex2html) for converting images
- Use of LaTeX in the XML to HTML route for the math
- Use of a perl script to transform the XML

DTD

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!ENTITY amp "&#x26;#x26;" >
<!ENTITY lt  "&#x26;#x3C;" >

<!ENTITY  ier "<hi rend='sup'>er</hi>">    ...
<!ENTITY  numero "n<hi rend='sup'>o</hi>"> ...
<!ENTITY % mathml PUBLIC "mathml" "mathml2.dtd">
%mathml;

<!ENTITY % tei-aux "(cit | p | list | note | formula|table|figure)+" >
<!ENTITY % texte-general
   "ident | code | hi | term| ref| xref | formula | cit | label | list |
    note | figure | table | TeX | LaTeX" >

<!ENTITY % texte-restreint
  "ident | code |  hi | term | ref| xref | formula | TeX | LaTeX" >

<!ENTITY % tei-common-atts 'id ID #IMPLIED rend CDATA #IMPLIED'>

<!ELEMENT cit (ref) >
<!ATTLIST cit rend CDATA #IMPLIED >
```

```
<!ELEMENT list (head?, ((item)* | (label, item)+)) >
<!ATTLIST list
        %tei-common-atts;
        type (simple|gloss|ordered|description)  "simple" >

<!ELEMENT item (#PCDATA | %texte-general; | p | anchor)* >
<!ATTLIST item  id ID #IMPLIED rend CDATA #IMPLIED >

<!ELEMENT label (#PCDATA | %texte-restreint; | anchor)* >
<!ATTLIST label %tei-common-atts; >

<!ENTITY % div0-textp "(div1|div2|div3|div4)+">
<!ENTITY % div0-texts "(div1|div2|div3|div4)*">
<!ENTITY % div1-textp "(div2|div3|div4)+">
<!ENTITY % div1-texts "(div2|div3|div4)*">
<!ENTITY % div2-textp "(div3|div4)+">
<!ENTITY % div2-texts "(div3|div4)*">
<!ENTITY % div3-textp "(div4)+">
<!ENTITY % div3-texts "(div4)*">
<!ENTITY % particip "participant|participants|participante|participantes" >
<!ENTITY % ramodule-header "(moreinfo|keywords|%particip;|anchor)*">
<!ENTITY % tei-div-atts ’ %tei-common-atts; type CDATA #IMPLIED ’>
```

```
<!ELEMENT module
    (head, %ramodule-header;, (%div1-textp; | (%tei-aux;, %div1-texts;))) >

<!ATTLIST module %tei-div-atts; html CDATA #REQUIRED topic CDATA #IMPLIED >

<!ELEMENT div2 (head, %ramodule-header;,
  (%div2-textp; | (%tei-aux;, %div2-texts;))) >
<!ATTLIST div2 %tei-div-atts; >

<!ELEMENT div3 (head, %ramodule-header;,
  (%div3-textp; | (%tei-aux;, %div3-texts;))) >
<!ATTLIST div3 %tei-div-atts; >

<!ELEMENT div4 (head, %ramodule-header;, %tei-aux;) >
<!ATTLIST div4 %tei-div-atts; >

<!ELEMENT table ((head | anchor)*, (row, (anchor)*)+) >
<!ATTLIST table
        %tei-common-atts;
        rows NMTOKEN #IMPLIED
        cols NMTOKEN #IMPLIED >
```

```
<!ELEMENT row ((cell | table), (anchor)*)+ >
<!ATTLIST row
        %tei-common-atts;
        top-border (true|false) "false"
        bottom-border (true|false) "false"
        space-before CDATA #IMPLIED
        role CDATA "data" >

<!ELEMENT cell (#PCDATA | %texte-general; | anchor)* >
<!ATTLIST cell
        %tei-common-atts;
        role CDATA "data"
        rows NMTOKEN "1"
        cols NMTOKEN "1"
        right-border (true|false) "false"
        left-border (true|false) "false"
        halign CDATA #IMPLIED
        top-border (true|false) "false"
   >
```

```
<!ELEMENT figure
        ((anchor)*, (head, (anchor)*)?,
        (p, (anchor)*)*) >
<!ATTLIST figure
        id ID #IMPLIED
        rend CDATA #IMPLIED
        file CDATA #IMPLIED
        framed CDATA #IMPLIED
        width CDATA #IMPLIED
        height CDATA #IMPLIED
        scale CDATA #IMPLIED
        angle CDATA #IMPLIED>

<!ELEMENT formula (simplemath |math) >
<!ATTLIST formula
        %tei-common-atts;
        type (inline|display) "inline" >

<!ELEMENT simplemath (#PCDATA) >
```

```
<!ELEMENT keywords (term+) >
<!ATTLIST keywords %tei-common-atts;
        titre CDATA #FIXED "Keywords: " >

<!ELEMENT term (#PCDATA | %texte-restreint; | anchor)* >
<!ATTLIST term %tei-common-atts; type CDATA #IMPLIED >

<!ELEMENT p (#PCDATA | %texte-general; | anchor)* >

<!ATTLIST p
        %tei-common-atts;
        spacebefore CDATA #IMPLIED
        noindent CDATA #IMPLIED>

<!ELEMENT hi (#PCDATA | %texte-general; | anchor)* >
<!ATTLIST hi id ID #IMPLIED rend CDATA #REQUIRED >

<!ELEMENT ref (#PCDATA | %texte-general; | anchor)* >
<!ATTLIST ref %tei-common-atts;   target IDREF #IMPLIED >

<!ELEMENT xref (#PCDATA | %texte-general; | anchor)* >
<!ATTLIST xref  %tei-common-atts;  url CDATA #IMPLIED>
```

```
<!ELEMENT head (#PCDATA | %texte-general; | anchor)* >
<!ATTLIST head %tei-common-atts; type CDATA #IMPLIED >

<!ELEMENT note (#PCDATA | %texte-general; |  p  | anchor)* >
<!ATTLIST note
        place CDATA "unspecified" ...>

<!ELEMENT anchor EMPTY >
<!ATTLIST anchor  ...>

<!ELEMENT raweb (accueil, moreinfo?, composition, presentation,
  fondements?,domaine?, logiciels?, resultats,contrats?,international?,
  diffusion?,biblio) >
<!ATTLIST raweb year CDATA #IMPLIED >
<!ATTLIST raweb language CDATA #IMPLIED >

<!ELEMENT composition (moreinfo?,catperso+)>
<!ATTLIST composition
           titre CDATA #FIXED "Members of the team"
           html CDATA #FIXED "composition"
           numero CDATA #FIXED "1"
           id ID #IMPLIED>
```

```
<!ELEMENT presentation (module+) >
<!ATTLIST presentation
          titre CDATA #FIXED "Presentation and general goals"
          numero CDATA #FIXED "2"
          id ID #IMPLIED>


<!ELEMENT fondements (module+) >
<!ATTLIST fondements
          titre CDATA #FIXED "Scientific foundations"
          numero CDATA #FIXED "3"
          id ID #IMPLIED>


<!ELEMENT domaine (module+) >
<!ATTLIST domaine
        titre CDATA #FIXED "Application fields"
        numero CDATA #FIXED "4"
        id ID #IMPLIED>
```

```
<!ELEMENT logiciels (module+) >
<!ATTLIST logiciels
        titre CDATA #FIXED "Software"
        numero CDATA #FIXED "5"
        id ID #IMPLIED>


<!ELEMENT resultats (module+) >
<!ATTLIST resultats
        titre CDATA #FIXED "New results"
        numero CDATA #FIXED "6"
        id ID #IMPLIED>


<!ELEMENT contrats (module+) >
<!ATTLIST contrats
        titre CDATA #FIXED "Industrial contracts"
        numero CDATA #FIXED "7"
        id ID #IMPLIED>
```

```
<!ELEMENT international (module+) >
<!ATTLIST international
        titre CDATA #FIXED "Regional, national and international initiatives"
        numero CDATA #FIXED "8"
        id ID #IMPLIED>


<!ELEMENT diffusion (module+) >
<!ATTLIST diffusion
        titre CDATA #FIXED "Diffusion of results"
        numero CDATA #FIXED "9"
        id ID #IMPLIED>


<!ELEMENT accueil (theme,projet,projetdeveloppe,UR, adresse?,topic*) >
<!ATTLIST accueil html CDATA #REQUIRED >
<!ATTLIST accueil isproject (true|false) "true">
<!ELEMENT addresse (#PCDATA) >
```

```
<!ELEMENT theme (#PCDATA)>
<!ELEMENT typeprojet (#PCDATA)>
<!ELEMENT projet (#PCDATA|hi)*>
<!ELEMENT projetdeveloppe (#PCDATA|hi)* >
<!ELEMENT UR (URSophia|URRocquencourt|URRhoneAlpes|URRennes|URLorraine|URFuturs)+ >
<!ELEMENT URSophia EMPTY>
<!ATTLIST URSophia
    url CDATA #FIXED "http://www.inria.fr/inria/organigramme/fiche_ur-sop.en.html"
    nom CDATA #FIXED "Sophia Antipolis" >
...

<!ELEMENT participants (pers)+ > ...
<!ATTLIST participants titre CDATA #FIXED "Contributed by"> ...

<!ELEMENT catperso (head,pers+)>

<!ELEMENT pers (#PCDATA|hi|note|xref)* >
<!ATTLIST pers prenom CDATA #REQUIRED nom CDATA #REQUIRED>

<!ELEMENT moreinfo (p+) >
```

```
<!ELEMENT topic (t_titre) >
<!ELEMENT t_titre (#PCDATA) >
<!ATTLIST topic num CDATA #IMPLIED>

<!ELEMENT TeX EMPTY>
<!ELEMENT LaTeX EMPTY>

<!-- specific to RR -->
<!ELEMENT RRstart (UR,title, etitle, projet, theme, motcle, keyword,
   resume, abstract, author,date, RRnumber)>
<!ELEMENT title (#PCDATA|hi|LaTeX)* > ...
<!ELEMENT author (auth)* >
<!ELEMENT auth (#PCDATA)* >

<!ELEMENT div1
    (head, (%div1-textp; | (%tei-aux;, %div1-texts;))) >
<!ELEMENT div0
    (head, (%div0-textp; | (%tei-aux;, %div0-texts;))) >
<!ELEMENT rr (RRstart,div0*)>
<!ATTLIST rr language CDATA #IMPLIED type CDATA #IMPLIED>
```

```
<!ELEMENT biblio (citation)* >
<!ATTLIST biblio
        html CDATA #FIXED "bibliography"
        titre CDATA #FIXED "Bibliography"
        numero CDATA #FIXED "10">


<!ENTITY % bibliostuff "bnote|bauteurs|bediteur|btitle|borganization|
     bschool|byear|bmonth|xref|bseries|bnumber|bvolume|bedition|
     binstitution|baddress|bpages|bhowpublished|bbooktitle
     |bpublisher|bjournal|bchapter|btype">

<!ELEMENT citation (%bibliostuff;)*>
<!ATTLIST citation key CDATA #REQUIRED
                  id ID #REQUIRED
                  type (book|booklet|proceedings|phdthesis|article|inbook|
         incollection|inproceedings|conference|manual|techreport|coursenotes
        |unpublished |misc|masterthesis|mastersthesis) #REQUIRED
                  from (year|foot|refer) #REQUIRED >
<!ELEMENT borganization (#PCDATA) >
<!ATTLIST borganization bname CDATA #FIXED "organisation" >
```

```
...
<!ELEMENT btitle (#PCDATA|hi|TeX|LaTeX|formula)* >
<!ATTLIST btitle bname CDATA #FIXED "title" >
<!ELEMENT bauteurs (bpers|etal)* >
<!ATTLIST bauteurs bname CDATA #FIXED "authors" >
<!ELEMENT beditеur (bpers|etal)* >
<!ATTLIST beditеur bname CDATA #FIXED "editors" >

<!ELEMENT etal EMPTY>
<!ATTLIST etal nom CDATA #FIXED "et al." >

<!ELEMENT bpers EMPTY>
<!ATTLIST bpers prenom CDATA #REQUIRED
                part CDATA #IMPLIED
                nom CDATA #REQUIRED
                junior CDATA #IMPLIED>
```

DTD

```
div0   (=section)
 |
div1   (=module)
 |
div2   (=\subsubsection)
 |
div3   (=\paragraph)
 |
div4   (=\subparagraph)
 |
 p   +  Dformula +  Dtable + Dfigure
 |
text +  Iformula + Itable + Ifigure + ...
```

Raweb = header + spec.mod. + some div0 + biblio

I=inline, D=display or float

Analyse Avancée Appliquée à l’optimisation des codes

A3

Rapport d’activité de l’année 2002

Rocquencourt

Thème : 1A

Page de présentation du projet - Rapport d’activité au format PostScript ou PDF

Before Tralics

- Preliminary version (Pdf only, 2001)
- Using ltx2x (for the text) and Ω (for the math) and Perl (for macro-expansion)
- 5500 lines of Perl code (1/3 of latex2html)
- latex2html has problems with tokenisation and spaces
- ltx2x is OK if the syntax is regular
- perl version OK unless `\let\oltit\it \def\it{\rm\oldit}`

Tralics in General

- Is written in C++ (25 000 lines)
- Compiled on linux, SunOS, MacOSX, Windows + cygwin
- Home page: http://www-sop.inria.fr/miaou/tralics
- TEX syntax, LATEX semantics

The p element

- Is the cornerstone of Tralics
- leave-v-mode opens, leave-h-mode closes a par.
- `\par` sometimes ignored, no `\long`, `\outer`
- Environments center, quote, flushleft, flushright
- Commands `\par`, `\\`, `\indent`, `\noindent`.
- Font changes are local.

Structure of the program

- Table of equivalents (catcodes, integer registers, etc)
- Other registers: glue, token list, boxes
- Execution stack and others (semantic nest)
- All the commmands
- The XML pointers stack, and tree
- The label-reference tables
- The citation-bibitem tables
- Simple language switch
- The *titlepage* data structure

Structure of the program

- No founts (metrics, ligatures, kerns)
- No hyphenation (T1 vs OT1)
- No dvi
- ISO 8859-1 input, Unicode output (`\char300` is OK)
- hbox=vbox= anonymous XML element
- A box is any XML element

Execution model

- TEX reads and expand macros
- TEX creates boxes, splits paragraphs
- TEX creates pages and outputs them
- TEX and AMS produce high quality math
- Tralics reads and expand macros
- Tralics creates XML elements, adds them to the tree
- Tralics checks and outputs the final tree
- Tralics is limited by MathML
- Tralics integrates bibtex

Example: moreinfo

- ltx2x code

```
TYPE= BEGIN_ENV
NAME= moreinfo
    START_TAG= "</par><MoreInfo><par>"
END_TYPE
TYPE= END_ENV
NAME= moreinfo
    START_TAG= "</par></MoreInfo>\n<par>"
END_TYPE
```

- Tralics code

```
primitive("moreinfo",moreinfo_cmd);
primitive("endmoreinfo",end_moreinfo_cmd);

 case moreinfo_cmd:
    leave_h_mode();       // execute \par
    stack.push1(X,X);     // emit <moreinfo>

 case end_moreinfo_cmd:
    leave_h_mode();       // execute \par
    stack.pop(X);         // skip over </moreinfo>
    stack.add_nl();       // add a linebreak
```

- Perl code

```
sub TR_beg_env {
   $y = $::my_envs{$z}; # $emacs
  ...
  return TM_push_par_hack("moreinfo","moreinfo","
        if $y == 1011; # $emacs
  ...
}

sub TR_end_env {
   ...
    return TM_pop_par_hack("moreinfo")
     if $y == 1011; # $emacs
   ...
}
```

Example: glo command

- ltx2x code

```
TYPE= COMMAND
NAME= \glo
    START_TAG= "</par><glo>"
    END_TAG= "</glo><par>"
    REQPARAMS= 1
    START_TAG_1=
    END_TAG_1=
END_TYPE
```

- Perl code

```
sub TM_glo {
   my $res = TM_pop("p");
   $res .= TMR_arg1("label");
   $res .= TM_push1("item","item","");
   $res .= TM_push_par;
   $res .= TM_next_arg;
   $res . TM_pop_par_hack("item");}
```

- Target

```
</p><label>#1of\glo</label>
<item><p>argument#2 of \glo</p></item><p>
```

glo command (continued)

```
void Parser::T_glo ()
{
  leave_h_mode();              // translate \par
  stack.set_arg_mode();        // switch to arg. mode
  T_arg1("label");             // <label> #1 </label>
  stack.push1("item","item");  // start <item>
  stack.set_v_mode();          // switch to vert. mode
  T_next_arg();                // translate next arg
  leave_h_mode();              // translate \par
  stack.pop("item");           // finish the </item>
  stack.set_no_mode();         // switch to undef mode
}
```

Example: module

- ltx2xcode

```
TYPE= BEGIN_ENV
NAME= RAmodule
    START_TAG= "</par><Module "
    END_TAG=
    REQPARAMS= 4
    START_TAG_1= "projet=""
    START_TAG_2=
     RESET_BUFFER: 7
     RESET_BUFFER: 8
    PRINT_P2= TO_BUFFER 7
    PRINT_P3= TO_BUFFER 8
    START_TAG_4= "type=""
     SOURCE: BUFFER 7
     STRING: "" html=""
     SOURCE: BUFFER 8
     STRING: ""> <sname>"
    END_TAG_4=  "</sname>\n<par>"
END_TYPE
```

- Target

```
</par><Module projet="#1" type="#2" html="#3">
<sname>#4</sname><par>
```

Example: module

- Perl code

```
sub TR_start_RAmodule{
  my($res,$aux,$att);
  $res = ""; $att = "";
  TM_next_arg;  # projet
  TM_next_arg;  # section
  $aux = TM_next_arg;  # nom du module
  $att .= " html=\"$aux\"";
  $att .= " id=\"mod:$aux\"";
  $aux = TM_push1("module", "module", $att);
  $aux .= TMR_arg1_head;
  $aux . TM_push_par;
}
```

- Target

```
</par><module html="#3" id="mod:#3">
<head>#4</head><par>
```

module

- Tralics code

```
void Parser::T_start_ramodule ()
{
  ignore_next_arg();            // skip project name
  ignore_next_arg();            // skip section name
  ostring aux = sT_next_arg();  // fetch module name
  stack.push1("module", module_name); // emit <module
  stack.add_att_to_last("html",aux); // html='#3'?
  stack.ok_for_label();         // \def\@currentlabel
  see_new_id_spec(true);        // NoTitle hack
  add_id("mod:" +aux);          // \label{mod:#3}
  stack.set_arg_mode();         // switch to arg mode
  T_arg1_head();                // translate title
  stack.set_v_mode();           // start in vert. mode
}
```

- How to translates this ?

```
\label{this is a funny label}
```

center cmd

- ltx2x code

```
TYPE= BEGIN_ENV
NAME= center
    START_TAG= "</par><center><par>"
END_TYPE

TYPE= END_ENV
NAME= center
    START_TAG= "</par></center><par>"
END_TYPE
```

- Tralics code

```
  primitive("center",    center_cmd,center_code);
  primitive("quote",     center_cmd,quote_code);
  primitive("quotation", center_cmd,quote_code);
  primitive("flushleft", center_cmd,flushleft_code);
  primitive("flushright",center_cmd,flushright_code);
case center_cmd:
    leave_h_mode();    // execute \par
    word_define(incentering_code,cur_chr,false);
case end_center_cmd:
    leave_h_mode();   // execute \par
```

References

[2000] Carlisle, David, M. Goossens, and S. Rahtz. “De XML à PDF avec `xmltex` et PassiveTEX”. In *Cahiers Gutenberg*, number 35-36, pages 79–114. 2000.

[2001] Carlisle, David, P. Ion, R. Miner, and N. Poppelier. “Mathematical Markup Language (MathML) Version 2.0”. http://www.w3.org/TR/MathML2/, 2001.

[1993] Goossens, Michel, F. Mittelbach, and A. Samarin. *The LaTeX companion*. Addison Wesley, 1993.

[2002] Grimm, José. “Outils pour la manipulation du rapport d’activité”. Technical Report RT-0265, Inria, 2002.

[1999] Haralambous, Yannis and J. Place. “Produire du MathML et autres ...ML à partir d’Ω : Ω se généralise”. In *Cahiers Gutenberg*, number 33-34, pages 173–182. 1999.

[1984] Knuth, Donald E. *The TEXbook*. Addison Wesley, 1984.

[1988] Louarn, Philippe. “Une expérience d’utilisation de LaTeX : le Rapport d’activité de l’INRIA”. *Cahiers Gutenberg* (0), 17–24, 1988.

[2003] Rahtz, Sebastian. “Passive TEX”. http://www.tei-c.org.uk/Software/passivetex/, 2003.

[2000] The Unicode Consortium. *The Unicode Standard, version 3.0*. Addison Wesley, 2000.

[1999] Walsh, Norman and L. Muelner. *Docbook. The Definite Guide*. O’Reilly & Associates, Inc, 1999.