Universitat de Lleida
Escola Universitària Politècnica
Enginyeria Tècnica en Informàtica de Gestió

Treball Final de Carrera

Disseny i Implementació d'una eina de traducció de documents LaTeX a HTML

Autor: Albert Nadal i Garriga
Director: Josep Ma. Ribó i Balust
Juliol 2003

Part 1/3

Introducció

Característiques del LaTeX

- S'encarrega únicament de la composició tipogràfica de documents, no de la seva escriptura.
- Enumera automàticament els apartats, figures, equacions i taules.
- Disposa de mecanismes automàtics per incloure referències creuades i mantenir-les.
- Permet la composició de construccions matemàtiques complexes.

Característiques de l'HTML

(HyperText Markup Language)

- És un sistema de definició de documents estructurats, normalitzat i acceptat internacionalment.

- Permet la composició de taules, formularis, imatges flotants, etc...

- La informació s'enllaça entre si mitjançant *links* i anclatges.

Qué tenen en comú LaTeX i HTML?

- **Són llenguatges de tipus SGML** *(Standard Generalized Markup Languages).*

És a dir, els dos són llenguatges de disseny descriptiu basat en etiquetes. La funció de l'etiquetatge és aportar informació que defineixi l'estructura jeràrquica d'un document, de manera que el contingut pugui ser processat d'una forma o altra per un ordinador.

LaTeX

```
\section{Enumeraciones}
```

HTML

```
<H3>Enumeraciones</H3>
```

Es tracta doncs, d'analitzar la gramàtica d'un document LaTeX i traduir-la a la seva corresponent en HTML.

Algunes composicions en LaTeX tenen traducció directa a HTML:

```
\emph{Departamento de Lenguajes}
```

⇩

```
<EM>Departamento de Lenguajes</EM>
```

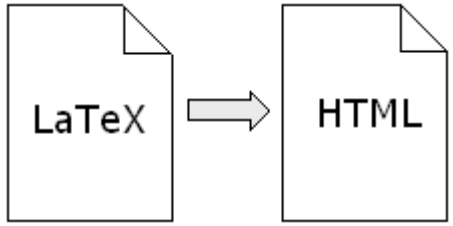
D'altres es tradueixen com la composició d'un conjunt:

```
\title{The Language Evolution}
\author{Marc D. Hauser \and C.Chomsky}
\date{November 2002}
\maketitle
```

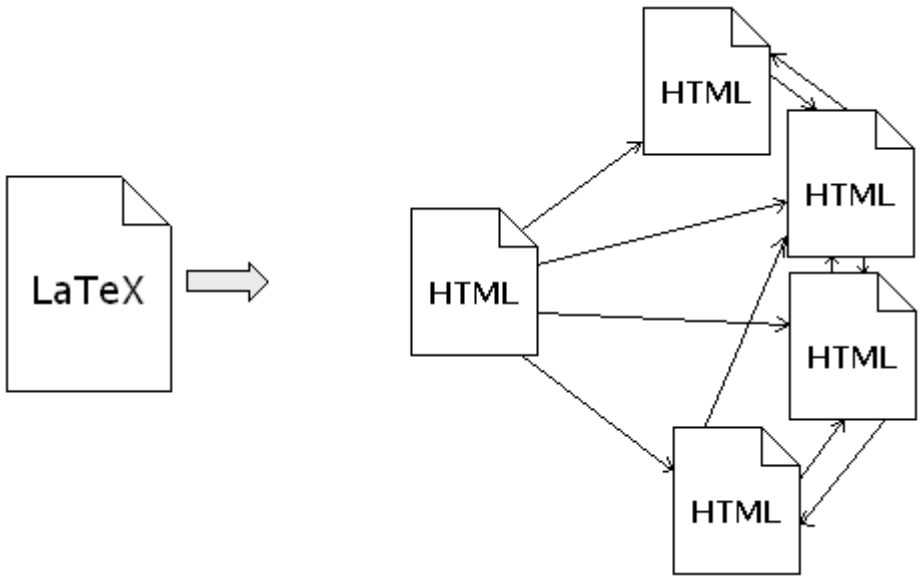
⇩

```
<h1 align=center>The Language Evolution</h1>
<h3 align=center>Marc D.Hauser&amp;nbsp;&amp;nbsp;C.Chomsky</h3>
<h3 align=center>November 2002</h3>
```

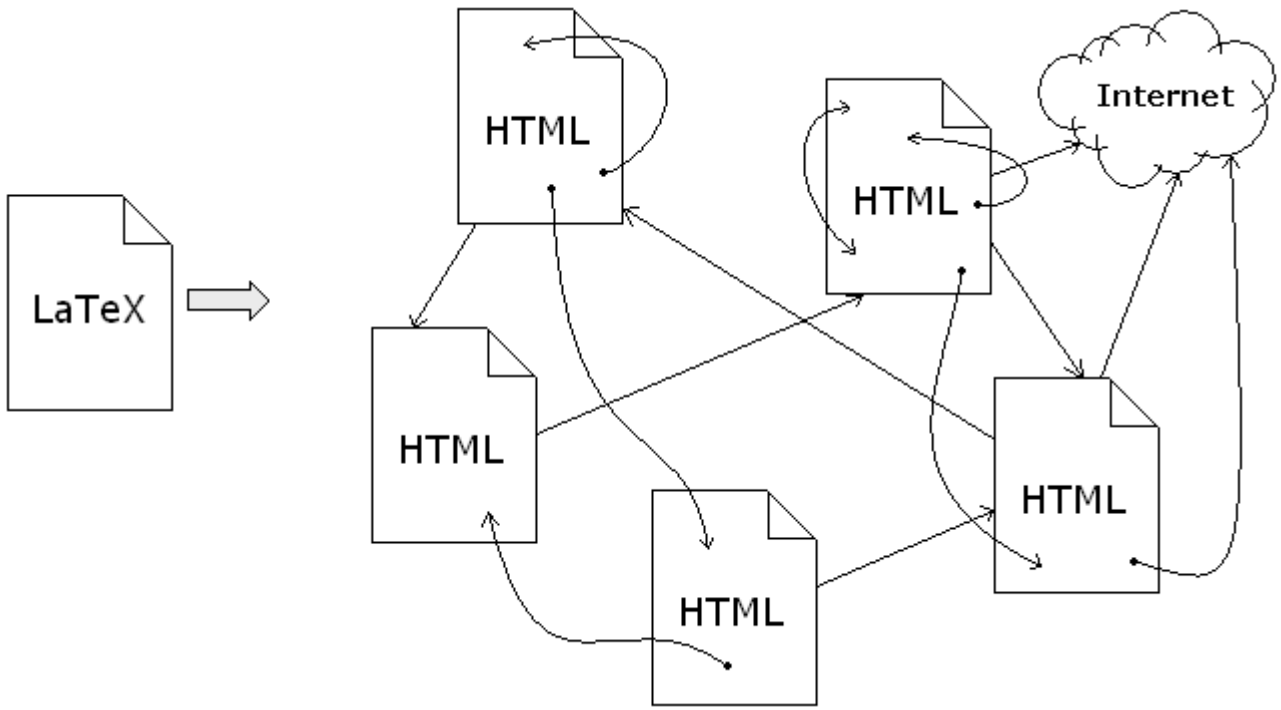
Traduïr la gramàtica no es suficient...



S'ha de fragmentar el document en nodes(apartats) i realitzar els adients enllaços entre els nodes...



I també s'han de considerar les referències creuades, notes a peu de pàgina, enllaços electrònics, hyperlinks, etc...



El desenvolupament del traductor ha implicat les següents fases:

• Determinar el conjunt de macros, entorns, símbols i accents que ha de sér capaç de traduïr.

• Especificar una gramàtica per reconèixer i analitzar documents realitzats en LaTeX.

• Implementar la gramàtica i els criteris i procediments de traducció.

Eines que s'han utilitzat

- gnu/g++ Per a la implementació en c++ del traductor
- gnu/Flex++ Generador d'anàlitzadors lèxics per a c++
- gnu/Bison++ Generador d'analitzadors sintàctics per a c++
- LaTeX vtex i web2C Distribucions del compilador de LaTeX
- gnu/Linux Mandrake Distribució del sistema Linux

Part 2/3

Especificació d'una gramàtica

Necessitat d'implementar una gramàtica

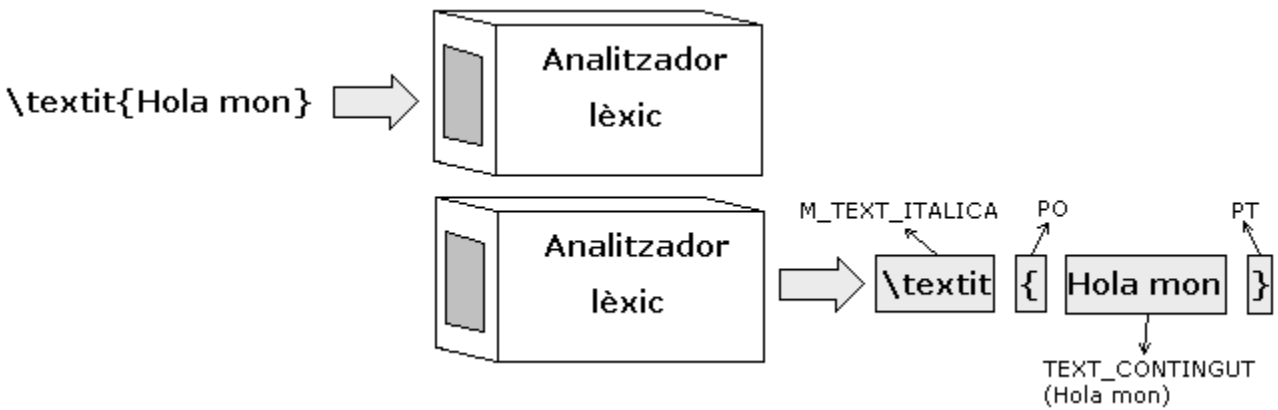
L'especificació d'una gramàtica permet al traductor validar si una determinada cadena pertany al llenguatge generat per la gramàtica. Per tal d'acceptar una seqüència de caràcters com una cadena generada per la gramàtica cal fer el següent:

- Analitzar-la i convertir-la en una cadena de símbols terminals de la gramàtica a partir d'una anàlisi lèxica.

- Analitzar la cadena de símbols terminals fins a acceptar-la com una cadena vàlida si supera l'anàlisis sintàctica.

Mitjançant una anàlisi lèxica descomposem una seqüència de caràcters en *tokens* o elements terminals. La descomposició la realitza l'analitzador lèxic efectuant comparacions entre el fluxe d'entrada i uns patrons prèviament definits.

Quan una cadena no compleix cap patró definit aleshores s'ha topat amb una macro no reconeguda pel traductor(tot i que això no comporta l'aturada de la traducció).



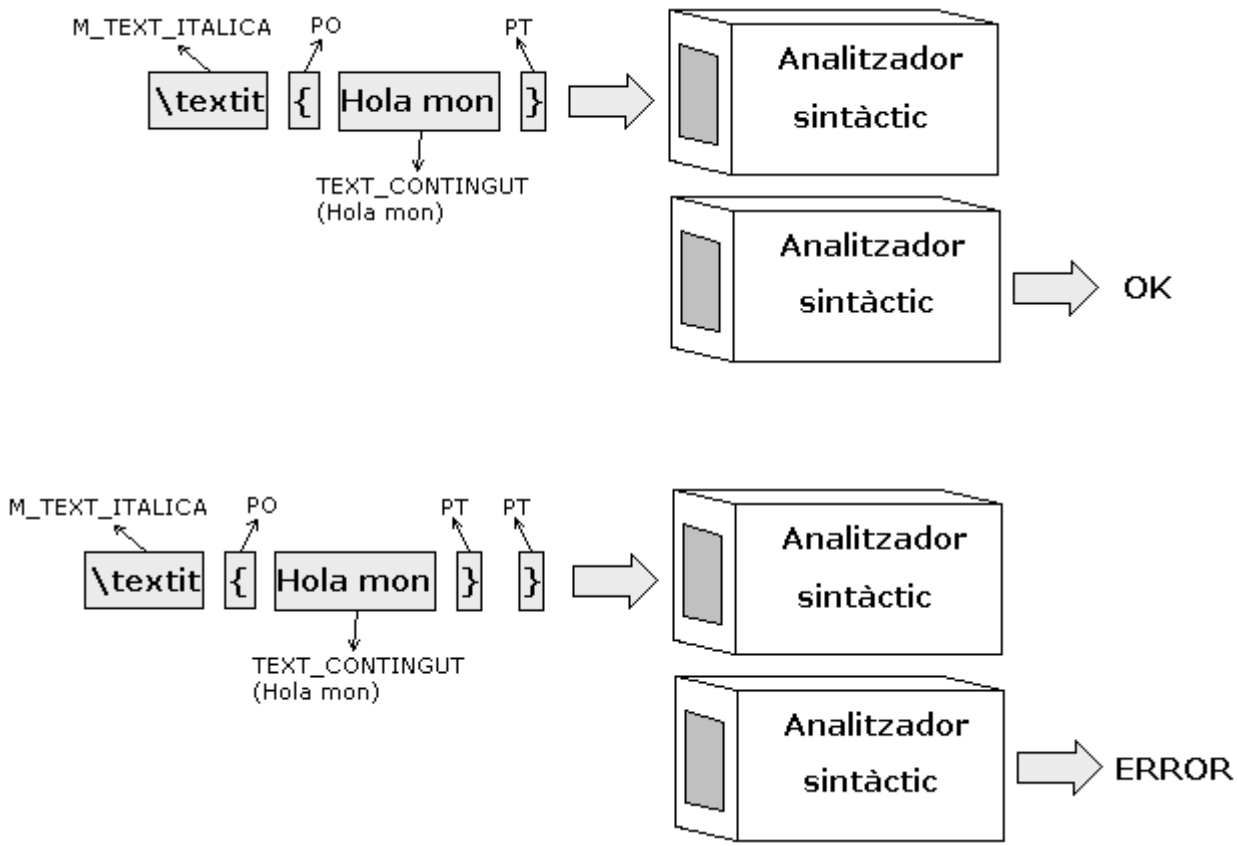
Fa falta implementar un analitzador lèxic?

No fa falta. *Flex++* és un aplicació generadora d'analitzadors lèxics per a c++

Genera *classes* que són capaces de realitzar anàlisis lèxiques:

Mitjançant una anàlisi sintàctica s'analitza cada un dels símbols terminals de la gramàtica fins a acceptar-la com a vàlida si supera l'anàlisi.

Si no supera l'anàlisi aleshores l'estructura sintàctica del document no és correcta i això implica la finalització de la traducció.



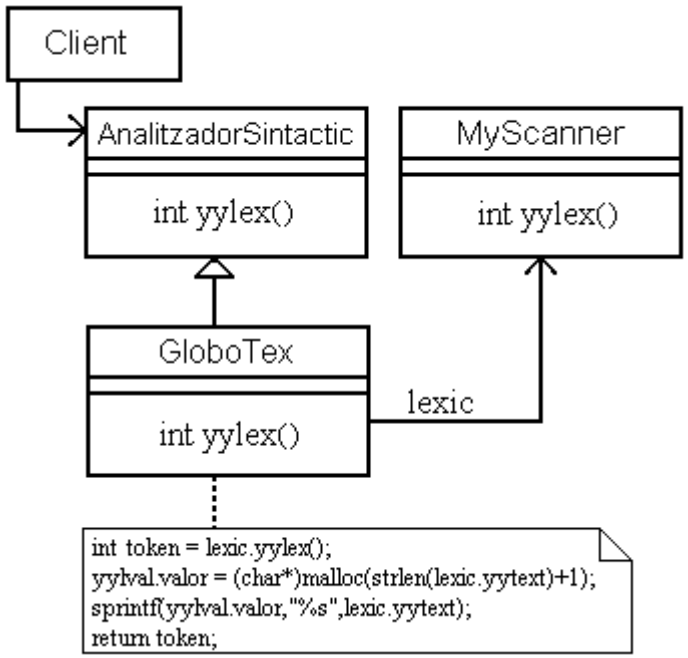
Fa falta implementar un analitzador sintàctic?

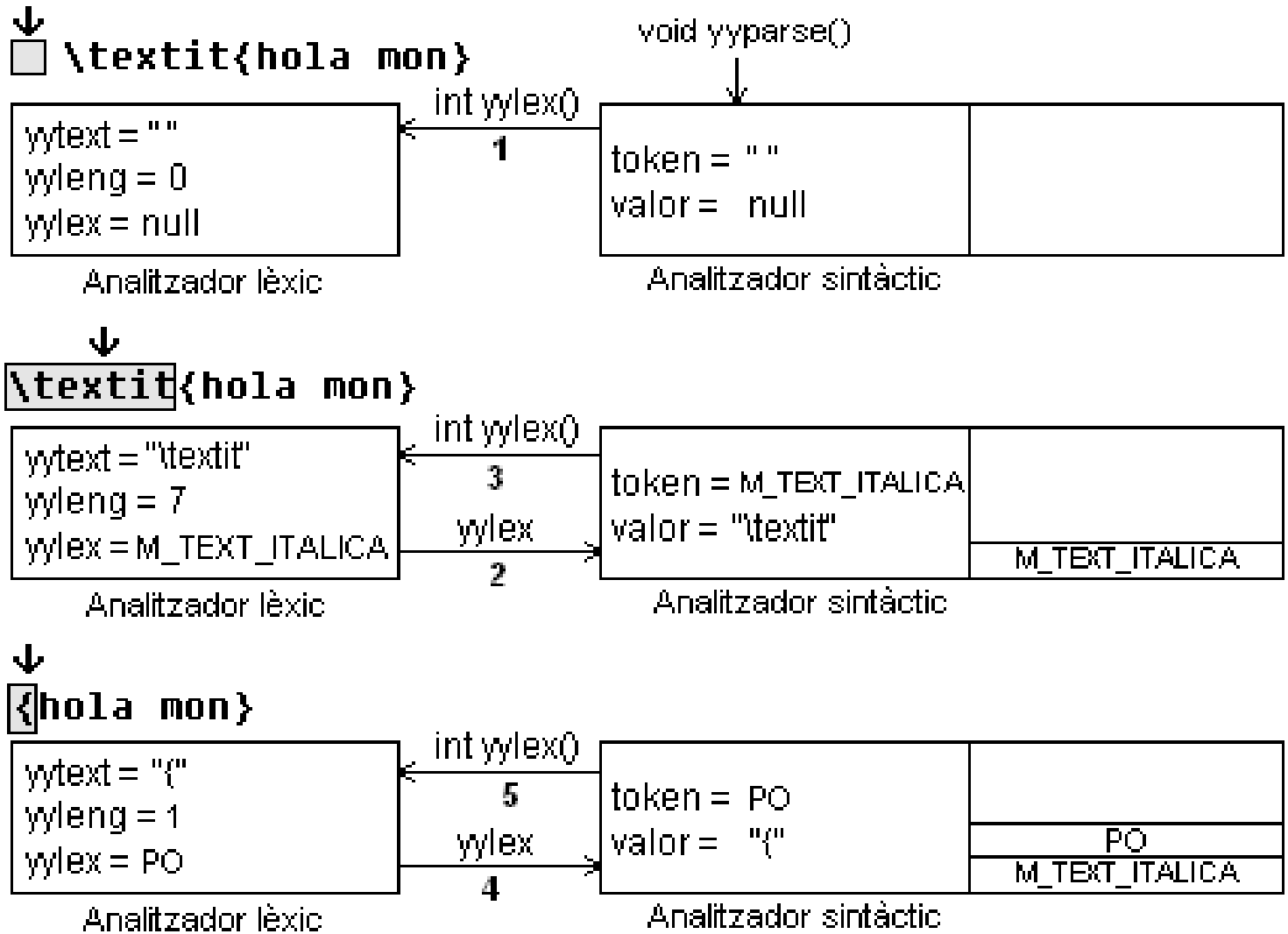
No fa falta. *Bison*++ és un aplicació generadora d'analitzadors sintàctics per a c++

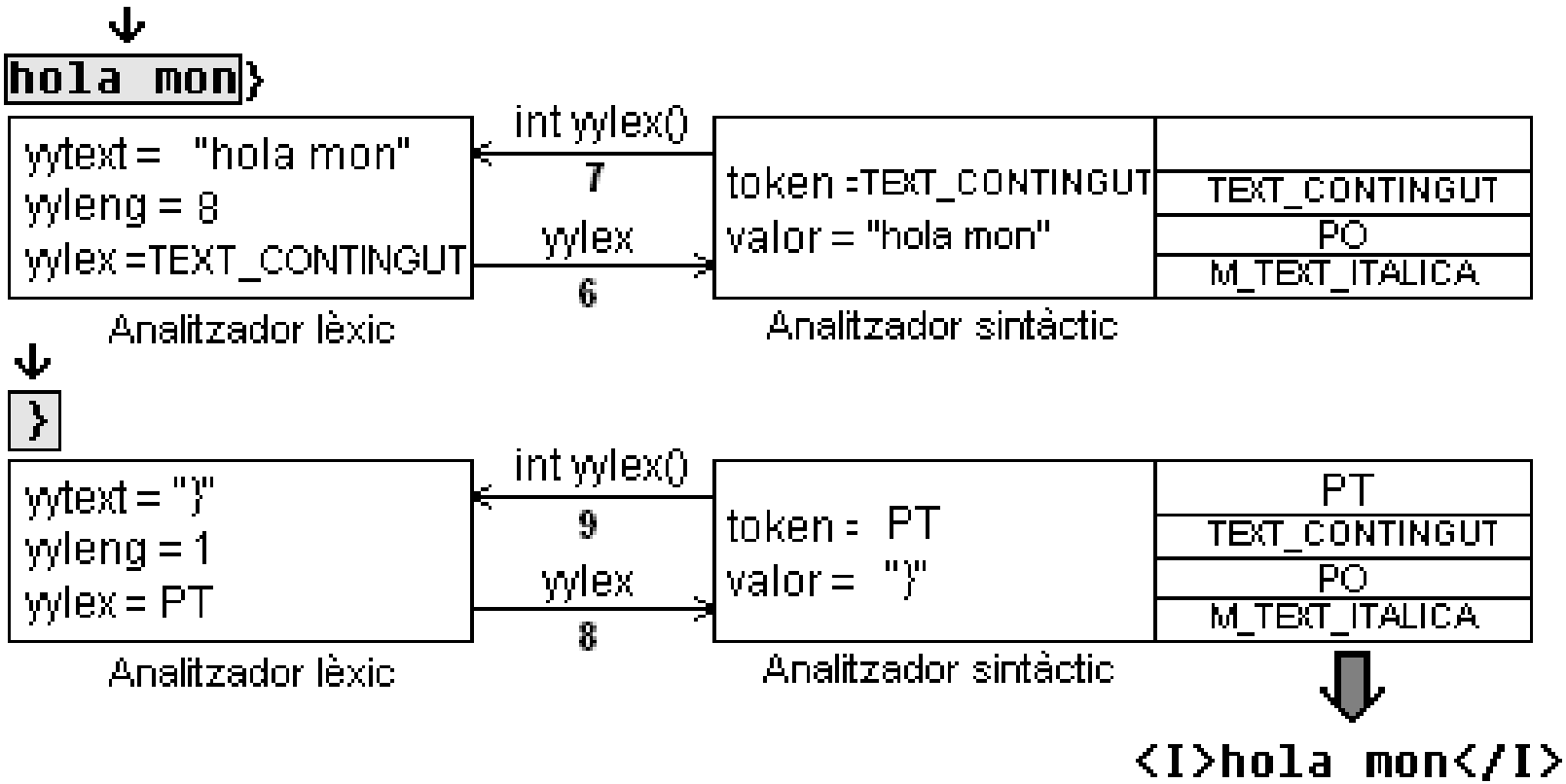
Genera *classes* que són capaces de realitzar anàlisis sintàctics:

Esquema d'interacció

El fluxe de sortida de l'analitzador lèxic és el fluxe d'entrada de l'analitzador sintàctic. Per tant s'han d'acoplar les dos *classes*...




\textit{hola mon}
void yyparse()
int yylex()
1
yytext = ""
yyleng = 0
yylex = null
token = " "
valor = null
Analitzador lèxic
Analitzador sintàctic
\textit{hola mon}
int yylex()
3
yylex
2
yytext = "\textit"
yyleng = 7
yylex = M_TEXT_ITALICA
token = M_TEXT_ITALICA
valor = "\textit"
M_TEXT_ITALICA
Analitzador lèxic
Analitzador sintàctic
{hola mon}
int yylex()
5
yylex
4
yytext = "{"
yyleng = 1
yylex = PO
token = PO
valor = "{"
PO
M_TEXT_ITALICA
Analitzador lèxic
Analitzador sintàctic


hola mon}
yytext = "hola mon"
yyleng = 8
yylex =TEXT_CONTINGUT
int yylex()
7
yylex
6
token =TEXT_CONTINGUT
valor = "hola mon"
TEXT_CONTINGUT
PO
M_TEXT_ITALICA
Analitzador lèxic
Analitzador sintàctic
}
yytext = "}"
yyleng = 1
yylex = PT
int yylex()
9
yylex
8
token = PT
valor = "}"
PT
TEXT_CONTINGUT
PO
M_TEXT_ITALICA
Analitzador lèxic
Analitzador sintàctic
<I>hola mon</I>

Part 3/3

Procediments de traducció

Aspectes més importants

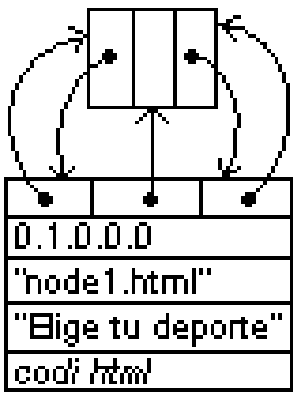
- Descomposició del document en nodes
- Necessitat d'una pila d'entorns
- Tractament de les referències creuades
- Tractament dels fragments matemàtics

Descomposició del document en nodes

Els documents es descomposen en nodes. Cada node representa un apartat(part, capítol, secció, subsecció o subsubsecció).

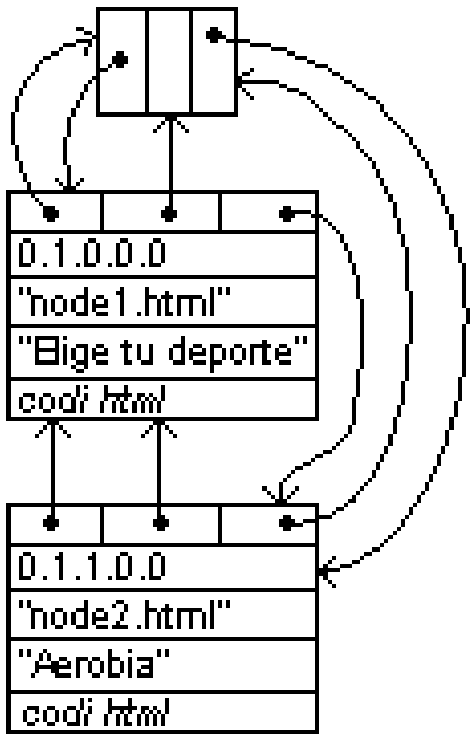
Els nodes s'emmagatzemen en una estructura arbòria que al finalitzar la traducció es materialitza en fitxers HTML.

L'estructura arbòria facilita la creació de panells de navegació i de la taula de continguts.



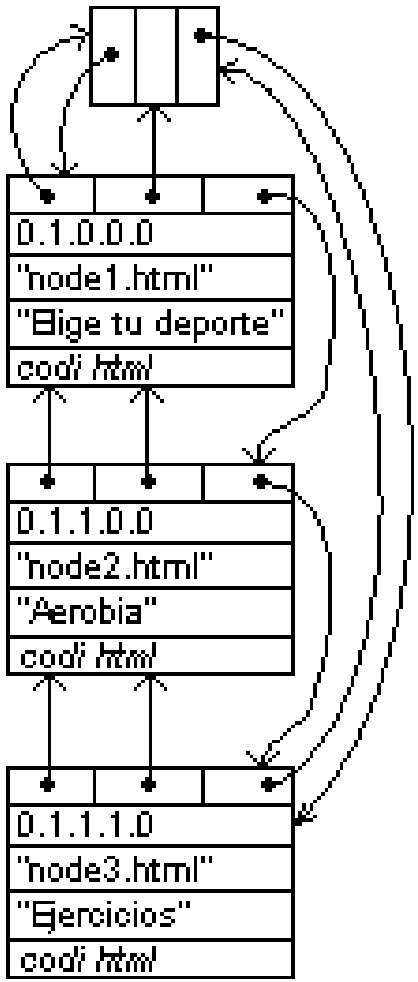
TEX

```
\begin{document}
\chapter{Elige tu deporte}
...
\section{Aerobia}
...
\subsection{Ejercicios}
...
\subsection{técnicas}
...
\section{Jogging}
...
\chapter{Alimentación}
...
\end{document}
```



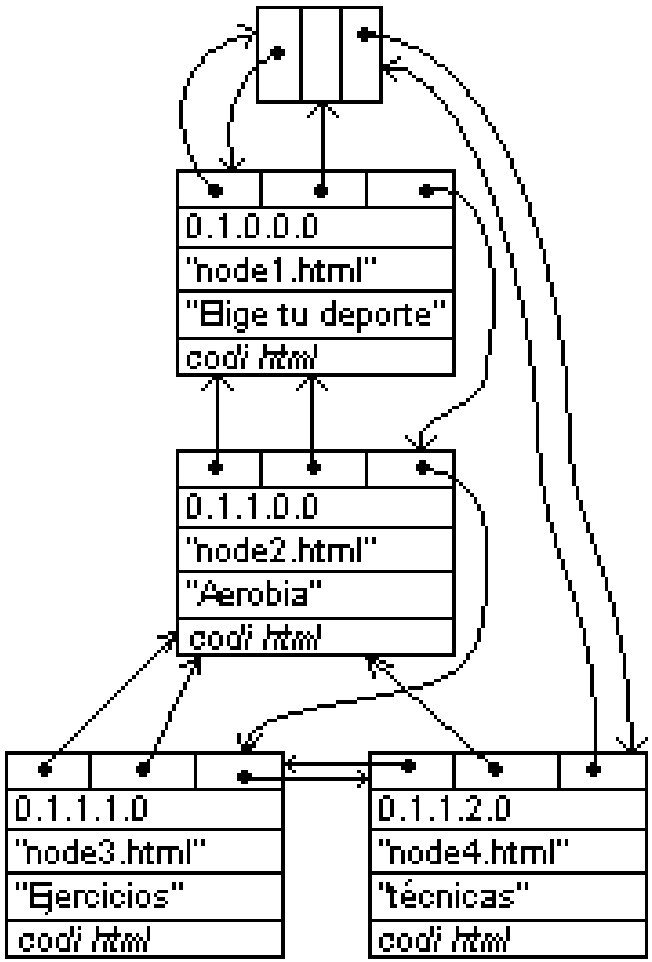
TEX

```
\begin{document}
\chapter{Elige tu deporte}
...
\section{Aerobia}
...
\subsection{Ejercicios}
...
\subsection{técnicas}
...
\section{Jogging}
...
\chapter{Alimentación}
...
\end{document}
```



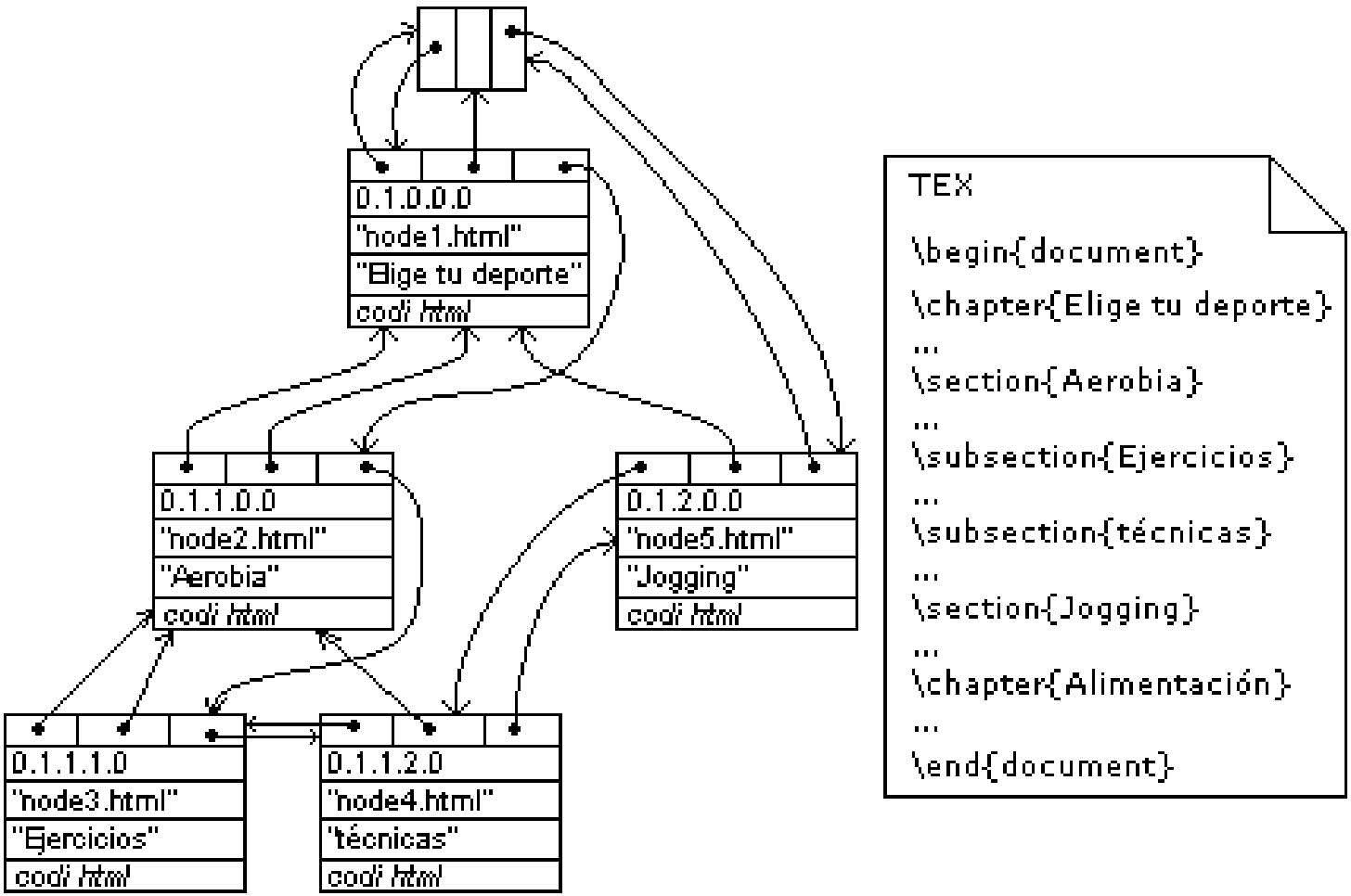
TEX

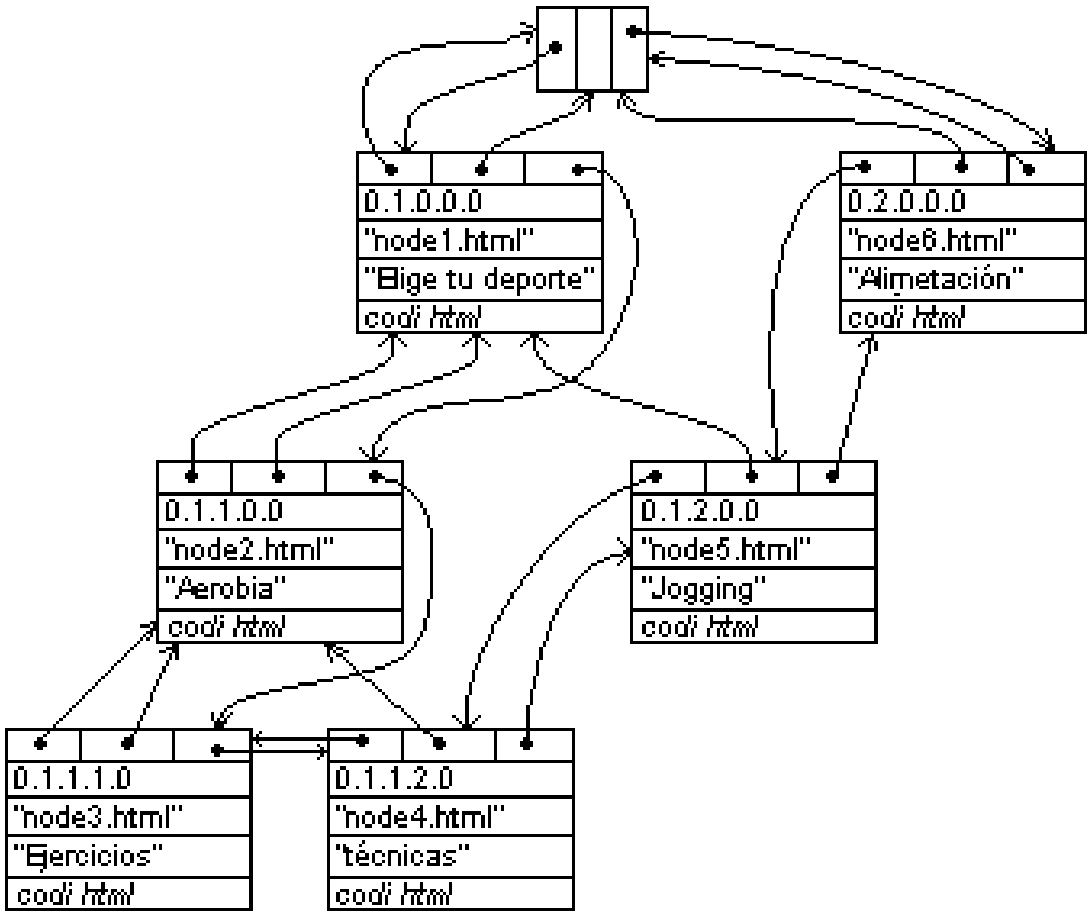
```
\begin{document}
\chapter{Elige tu deporte}
...
\section{Aerobia}
...
\subsection{Ejercicios}
...
\subsection{técnicas}
...
\section{Jogging}
...
\chapter{Alimentación}
...
\end{document}
```



TEX

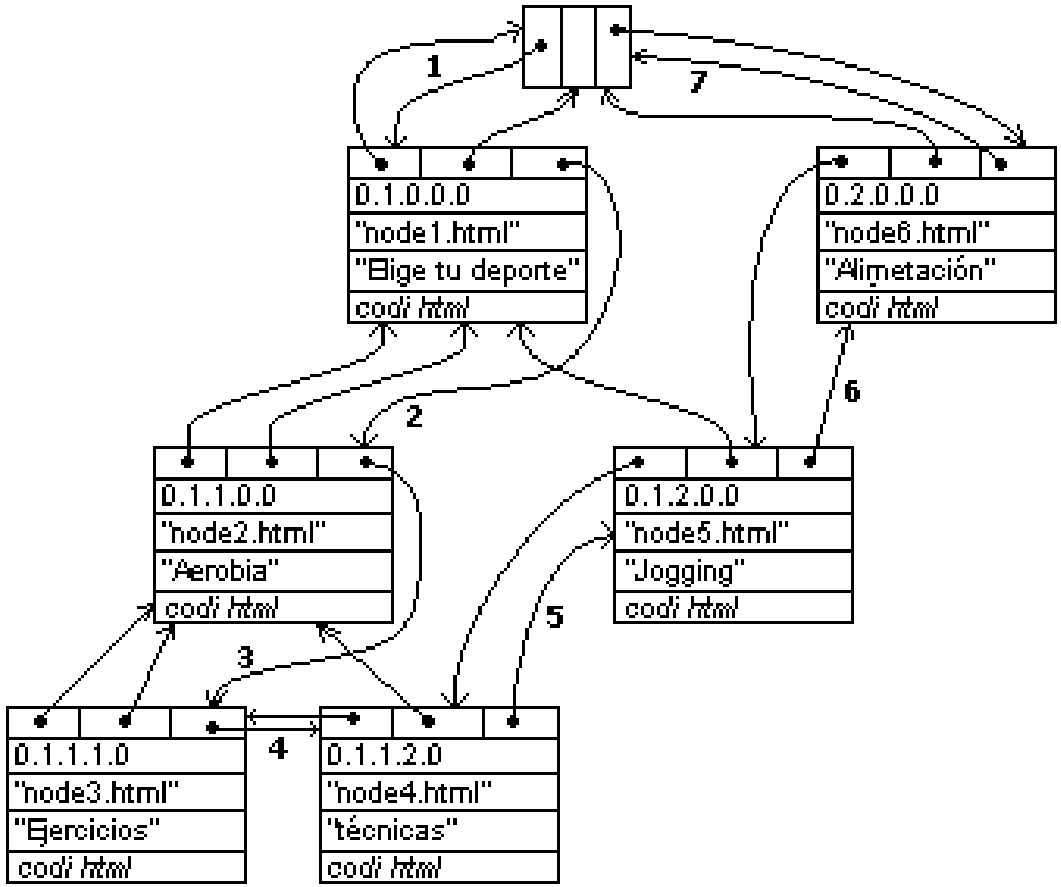
```
\begin{document}
\chapter{Elige tu deporte}
...
\section{Aerobia}
...
\subsection{Ejercicios}
...
\subsection{técnicas}
...
\section{Jogging}
...
\chapter{Alimentación}
...
\end{document}
```


0.1.0.0.0
"node1.html"
"Elige tu deporte"
codi html
0.1.1.0.0
"node2.html"
"Aerobia"
codi html
0.1.2.0.0
"node5.html"
"Jogging"
codi html
0.1.1.1.0
"node3.html"
"Ejercicios"
codi html
0.1.1.2.0
"node4.html"
"técnicas"
codi html
TEX
\begin{document}
\chapter{Elige tu deporte}
...
\section{Aerobia}
...
\subsection{Ejercicios}
...
\subsection{técnicas}
...
\section{Jogging}
...
\chapter{Alimentación}
...
\end{document}



TEX

```
\begin{document}
\chapter{Elige tu deporte}
...
\section{Aerobia}
...
\subsection{Ejercicios}
...
\subsection{técnicas}
...
\section{Jogging}
...
\chapter{Alimentación}
...
\end{document}
```



1 [<UL>
<LI><A HREF=node1.html#sec0_1_0_0_0 target=principal>Elige tu deporte</A>

2 [<UL>
<LI><A HREF=node2.html#sec0_1_1_0_0 target=principal>Aerobia</A>

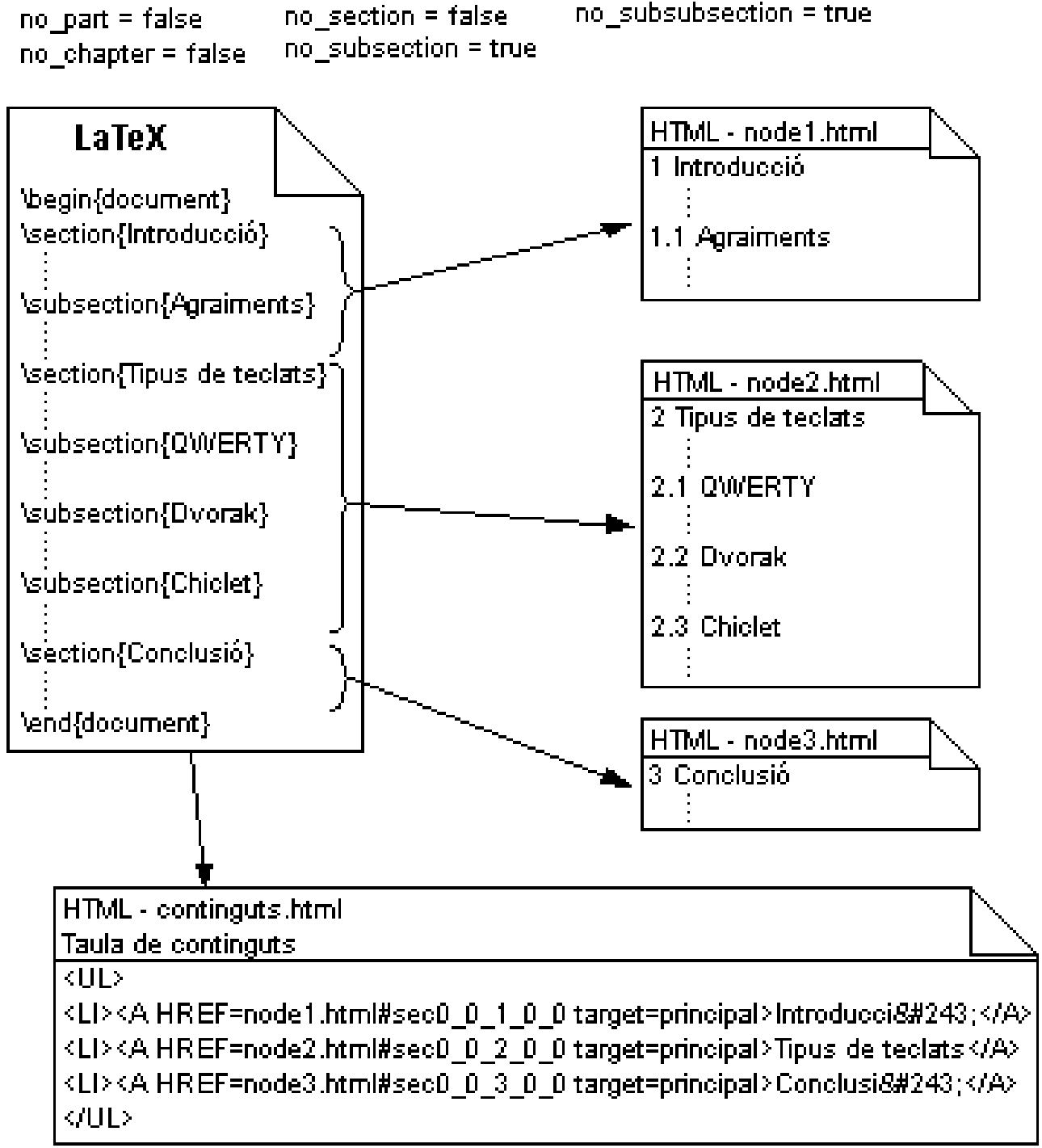
3 [<UL>
<LI><A HREF=node3.html#sec0_1_1_1_0 target=principal>Ejercicios</A>

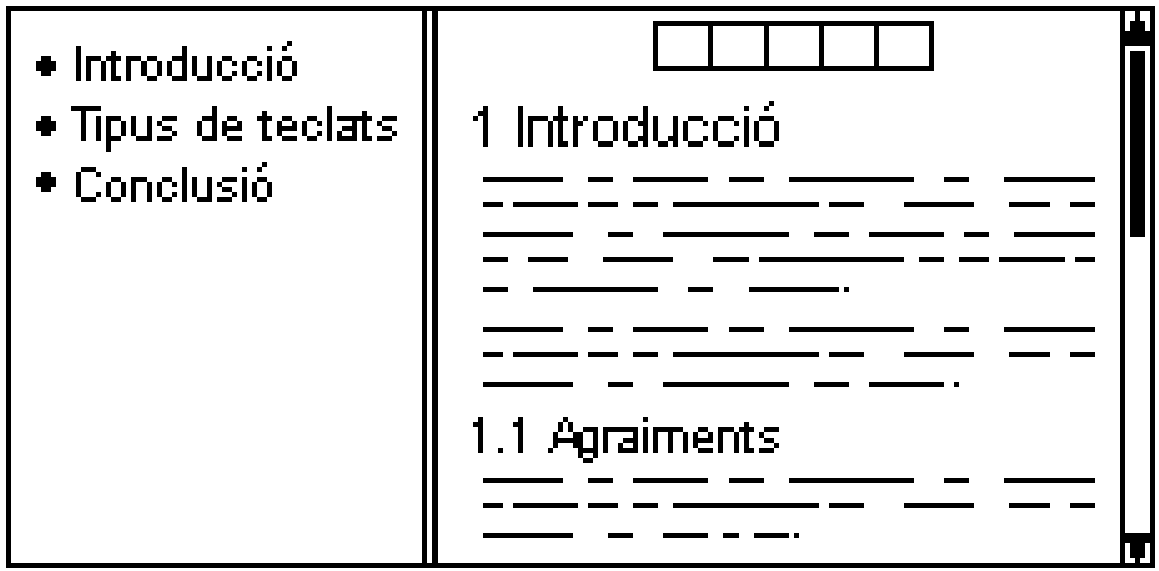
4 [<LI><A HREF=node4.html#sec0_1_1_2_0 target=principal>técnicas</A>

5 [</UL>
<LI><A HREF=node5.html#sec0_1_2_0_0 target=principal>Jogging</A>

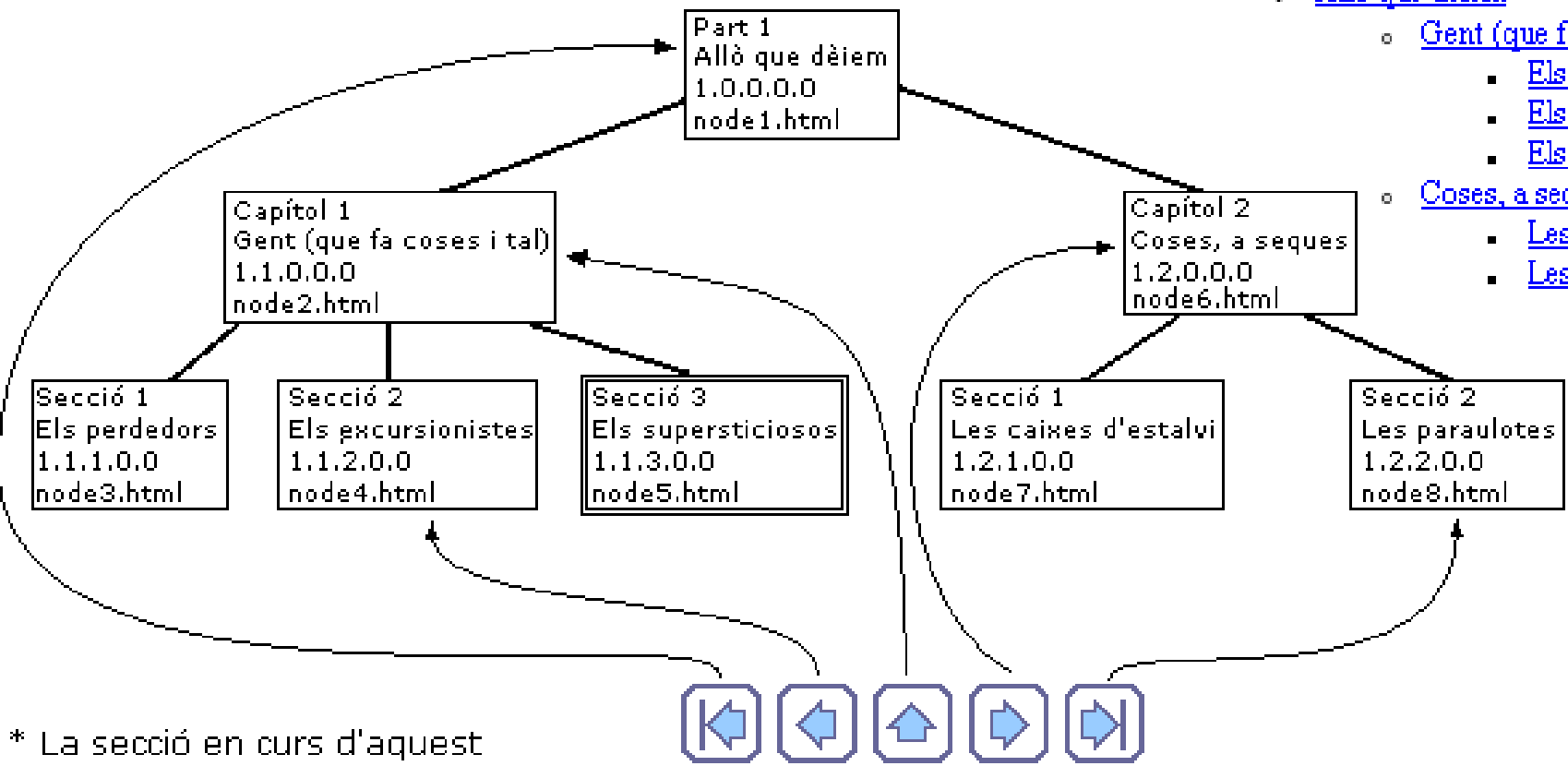
6 [</UL>
<LI><A HREF=node6.html#sec0_2_0_0_0 target=principal>Alimentación</A>

7 [</UL>


no_part = false
no_chapter = false
no_section = false
no_subsection = true
no_subsubsection = true
LaTeX
\begin{document}
\section{Introducció}
\subsection{Agraiments}
\section{Tipus de teclats}
\subsection{QWERTY}
\subsection{Dvorak}
\subsection{Chiclet}
\section{Conclusió}
\end{document}
HTML - node1.html
1 Introducció
1.1 Agraiments
HTML - node2.html
2 Tipus de teclats
2.1 QWERTY
2.2 Dvorak
2.3 Chiclet
HTML - node3.html
3 Conclusió
HTML - continguts.html
Taula de continguts
<UL>
<LI><A HREF=node1.html#sec0_0_1_0_0 target=principal>Introducció</A>
<LI><A HREF=node2.html#sec0_0_2_0_0 target=principal>Tipus de teclats</A>
<LI><A HREF=node3.html#sec0_0_3_0_0 target=principal>Conclusió</A>
</UL>


• Introducció
• Tipus de teclats
• Conclusió
1 Introducció
1.1 Agraiments

Esquema arbori



* La secció en curs d'aquest exemple és la secció 3 del capítol 1

Taula de continguts

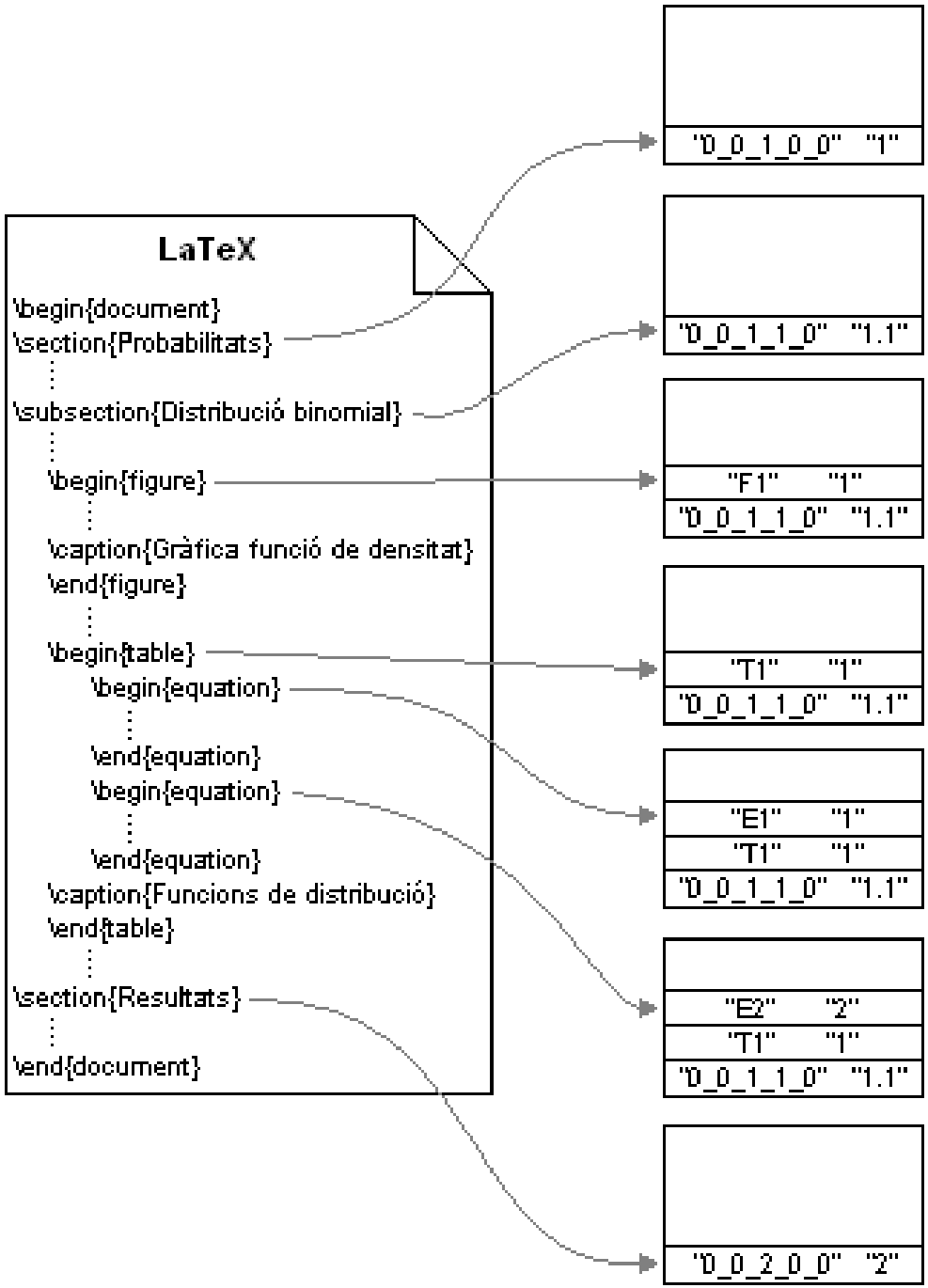
Necessitat d'una pila d'entorns

La complexitat d'un document LaTeX recau en l'aniuació d'entorns, és a dir, la composició d'entorns que es composen dins d'altres entorns.

El problema sorgeix en aquelles situacions en les quals interessa conèixer determinades dades de l'entorn actual en curs i més tard, recuperar les dades de l'entorn del nivell o nivells inferiors.

Aquestes situacions es donen en els entorns comptables: **figures**, **taules**, **equacions** o **apartats**

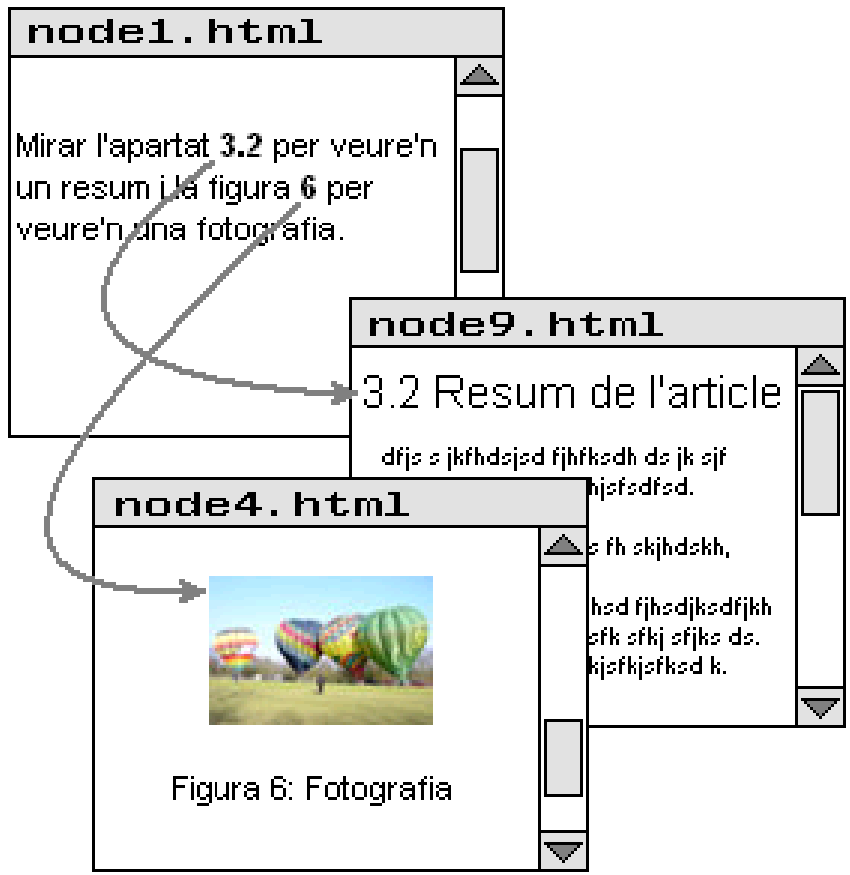
La solució és una pila d'entorns.


LaTeX
\begin{document}
\section{Probabilitats}
\subsection{Distribució binomial}
\begin{figure}
\caption{Gràfica funció de densitat}
\end{figure}
\begin{table}
\begin{equation}
\end{equation}
\begin{equation}
\end{equation}
\caption{Funcions de distribució}
\end{table}
\section{Resultats}
\end{document}
"0_0_1_0_0" "1"
"0_0_1_1_0" "1.1"
"F1" "1"
"0_0_1_1_0" "1.1"
"T1" "1"
"0_0_1_1_0" "1.1"
"E1" "1"
"T1" "1"
"0_0_1_1_0" "1.1"
"E2" "2"
"T1" "1"
"0_0_1_1_0" "1.1"
"0_0_2_0_0" "2"

Tractament de les referències creuades

Les macros *label* i *ref* serveixen per incloure referències a punts concrets del document LaTeX. La macro *label* etiqueta una zona comptable del document(apartat, figura, equació o taula) i la macro *ref* fa referència a una zona en concret.

```
...
\section{Resum de l'article}
\label{resum}
...
\begin{figure}
\caption{Fotografia}
\label{foto}
\end{figure}
...
Mirar l'apartat \ref{resum} per veure'n un resum i la figura \ref{foto} per veure'n una fotografia.
```



Tractament de les referències creuades

Les referències creuades es tradueixen en dos passos o escombrades

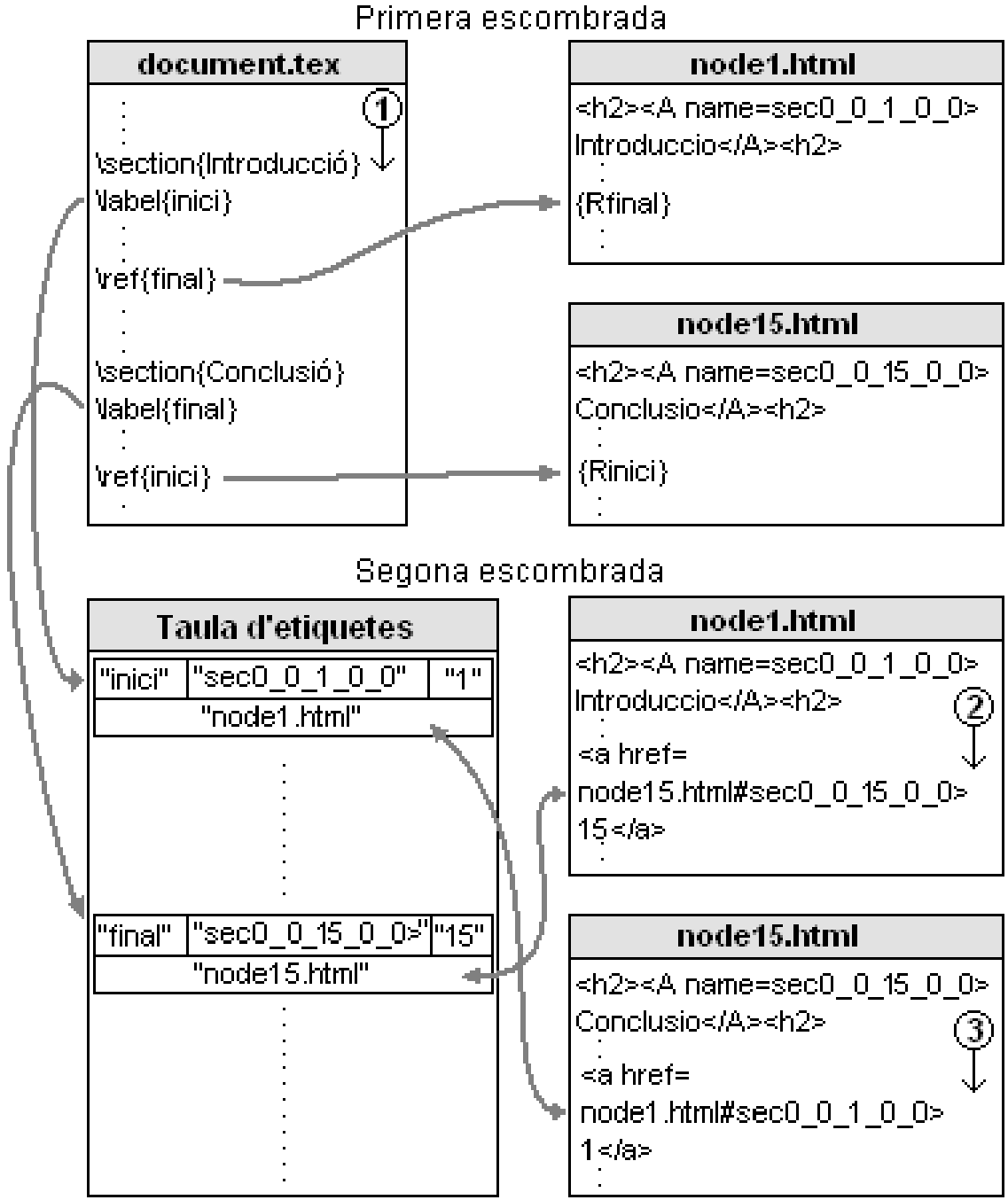
S'efectua sobre mentre es llegeix del fluxe de dades

1) Es capturen totes les etiquetes definides per la macro *label* i s'emmagatzemen en una taula juntament amb altres dades adicionals com el nom del fitxer del node on es troba. Es reemplacen les macros *ref* per una cadena de caràcters de tipus informatiu.

S'efectua sobre els nodes de l'arbre.

2) Aquest segon pas s'interpreta com una operació pendent del traductor un cop ha finalitzat la traducció base del document LaTeX.

Es recorre cada node de l'arbre i es busquen les cadenes de caràcters de tipus informatiu inserides en el primer pas. Quan es troba una cadena informativa es consulta la taula d'etiquetes i es reemplaça adientment amb un link.


Primera escombrada
document.tex
1
\section{Introducció}
\label{inici}
\ref{final}
\section{Conclusió}
\label{final}
\ref{inici}
node1.html
<h2><A name=sec0_0_1_0_0>
Introduccio</A><h2>
{Rfinal}
node15.html
<h2><A name=sec0_0_15_0_0>
Conclusio</A><h2>
{Rinici}
Segona escombrada
Taula d'etiquetes
"inici" "sec0_0_1_0_0" "1"
"node1.html"
"final" "sec0_0_15_0_0" "15"
"node15.html"
node1.html
<h2><A name=sec0_0_1_0_0>
Introduccio</A><h2>
2
<a href=
node15.html#sec0_0_15_0_0>
15
node15.html
<h2><A name=sec0_0_15_0_0>
Conclusio</A><h2>
3
<a href=
node1.html#sec0_0_1_0_0>
1

Tractament de les referències creuades

La mecànica és similar amb les macros *hyperlink*, *hypertarget*, *thebibliography*, etc...

Tractament dels fragments matemàtics

Una de les idees inicials era la traducció de fragments matemàtics (composició de formulacions matemàtiques) a mode text.

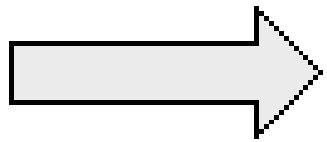
$$\frac{\pi}{2} = \frac{1}{\sqrt{\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}}\cdots}}$$

Donat que es poden composar formulacions matemàtiques molt complexes en LaTeX, aquesta opció era força difícil de portar a terme.

El mètode més fiable i eficient és la conversió dels fragments matemàtics a mode gràfic

Tractament dels fragments matemàtics

```
$$\frac{n}{2}
\int_a^b x\,dx$$
```



$$\frac{n}{2} \int_a^b x\,dx$$

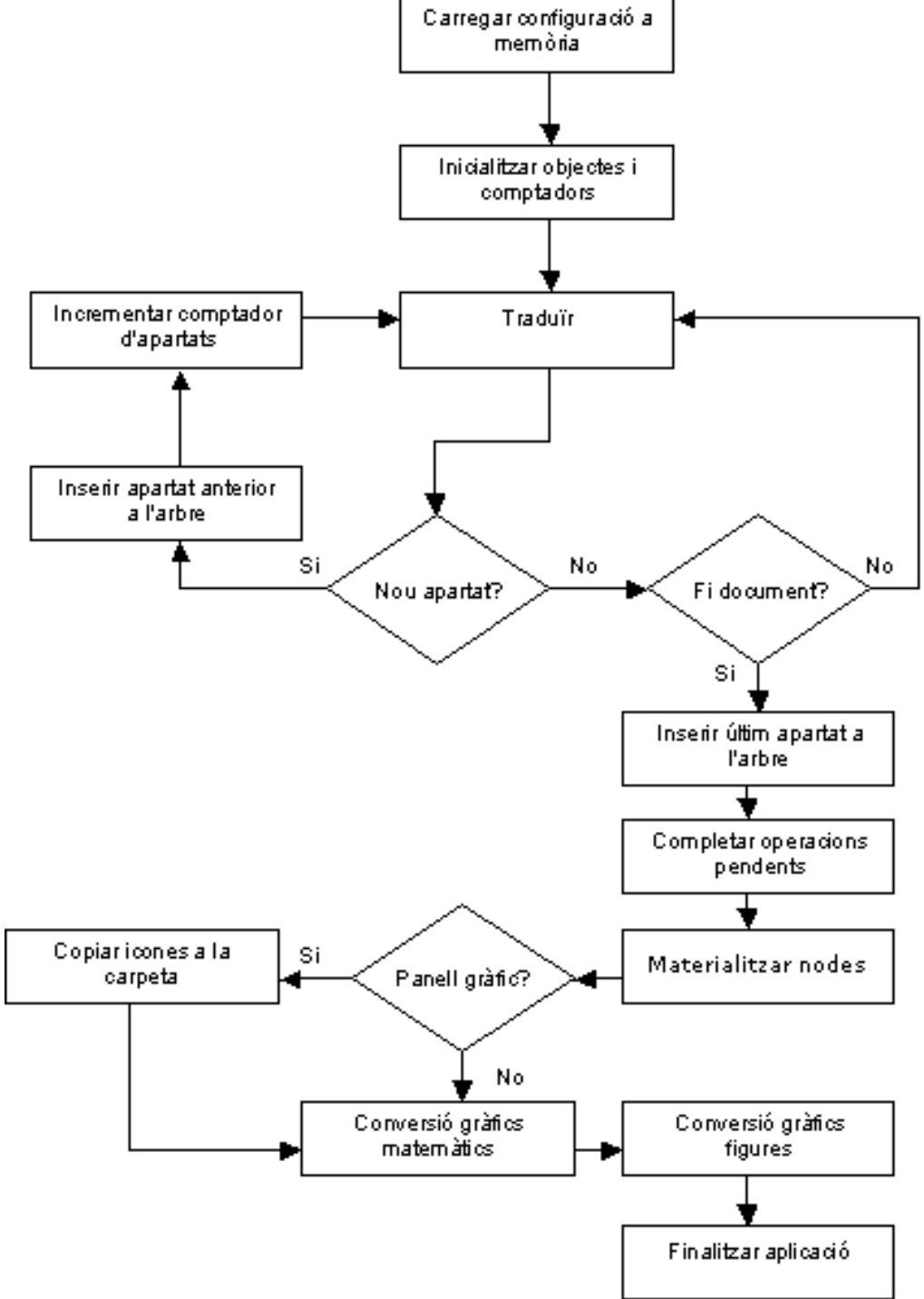
Quan el traductor detecta un fragment de text inclòs entre marques $ o $$ emmagatzema el fragment en una llista i assigna en aquell punt un enllaç a un fitxer gràfic que contindrà gràficament la fórmula matemàtica.

Les conversions es realitzen per mitjà de l'script *latex2gif* i *ps2gif*, scripts implementats per a aquesta finalitat.

latex2gif i *ps2gif*

```
#!/bin/sh
latex -interaction=batchmode $1.tex > /dev/null
dvips $1.dvi -o $1.ps -q > /dev/null
./ps2gif $1.ps $2.gif 140
rm $1.tex
rm $1.aux
rm $1.dvi
rm $1.log
rm $1.ps
```

```
#!/bin/sh
#Requereix tenir Ghostscript i les utilitats pbmplus instal·lades
echo "["$2"]»&2
gs -sDEVICE=ppmraw -sOutputFile=- -sNOPAUSE -r$3 -q $1 -c
showpage -c quit | pnmcrop -quiet | pnmmargin -white 10 | ppmquant
256 -quiet | ppmtogif -quiet >$2
```


Carregar configuració a memòria
Inicialitzar objectes i comptadors
Traduïr
Incrementar comptador d'apartats
Inserir apartat anterior a l'arbre
Nou apartat?
Si
No
Fi document?
No
Si
Inserir últim apartat a l'arbre
Completar operacions pendents
Materialitzar nodes
Panell gràfic?
Si
Copiar icones a la carpeta
No
Conversió gràfics matemàtics
Conversió gràfics figures
Finalitzar aplicació