

Capture de sites Web en ligne

Conférence B.N.F, Avril 2004

Xavier Roche(HTTrack)

Pourquoi copier des sites web?

- Archivage pour conservation et/ou historisation
- Archivage pour raisons légales
- Miroirs de sites pour des raisons de redondance
- Copies pour une mise à disposition non connectée
- Copies par des particuliers (copie privée)
- Agents intelligents, stress de réseaux, validation de liens cassés ou des liens externes, plan du site...

HTTrack WEBSITE COPIER

Le « Web », qu'est-ce que c'est ?

Internet

Email

HTTP

HTTPS

mailto:

http:

https:

news:

ftp:

News

WWW

FTP

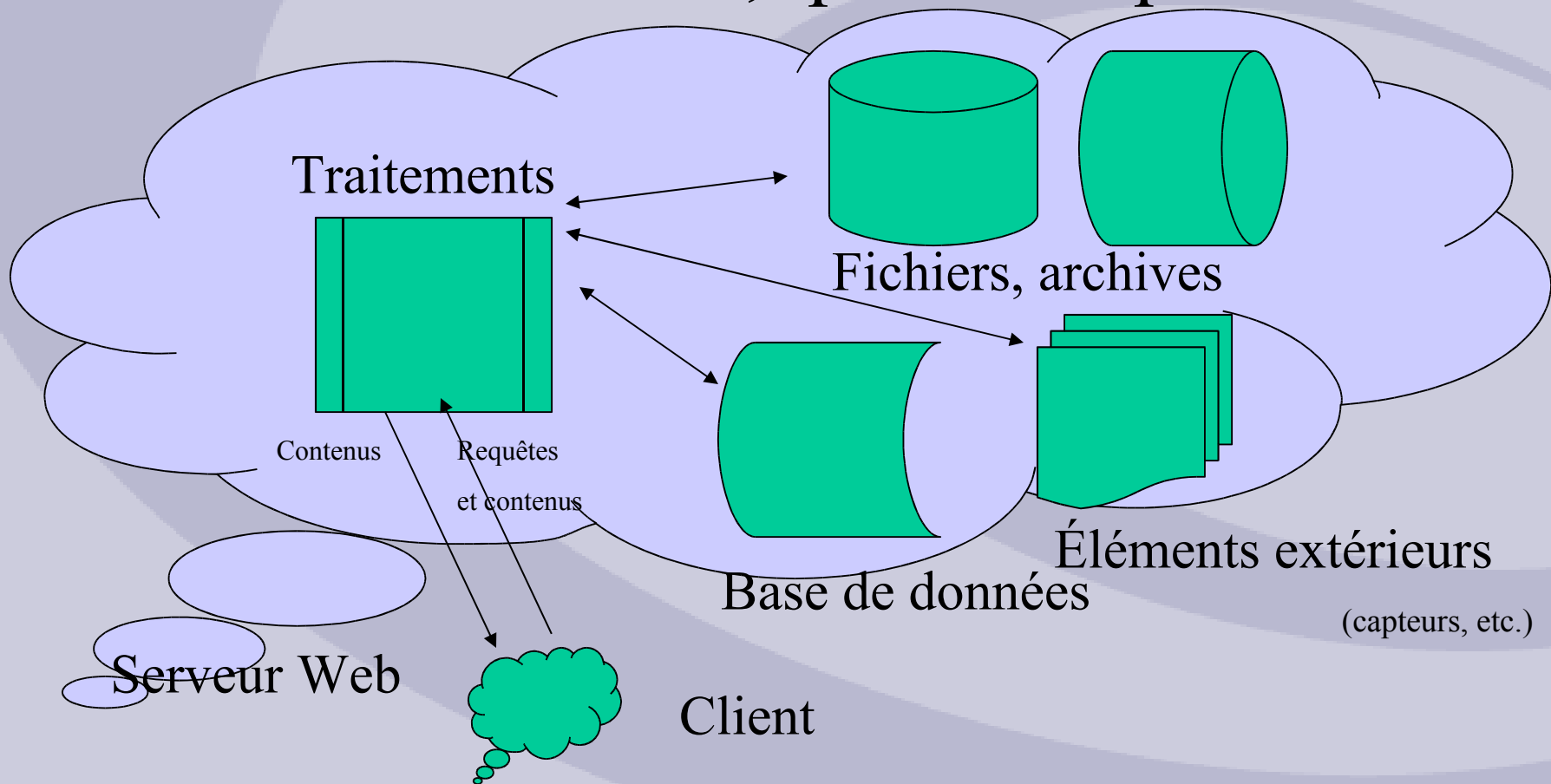
file:

Fichiers

Ressources locales (fichiers)

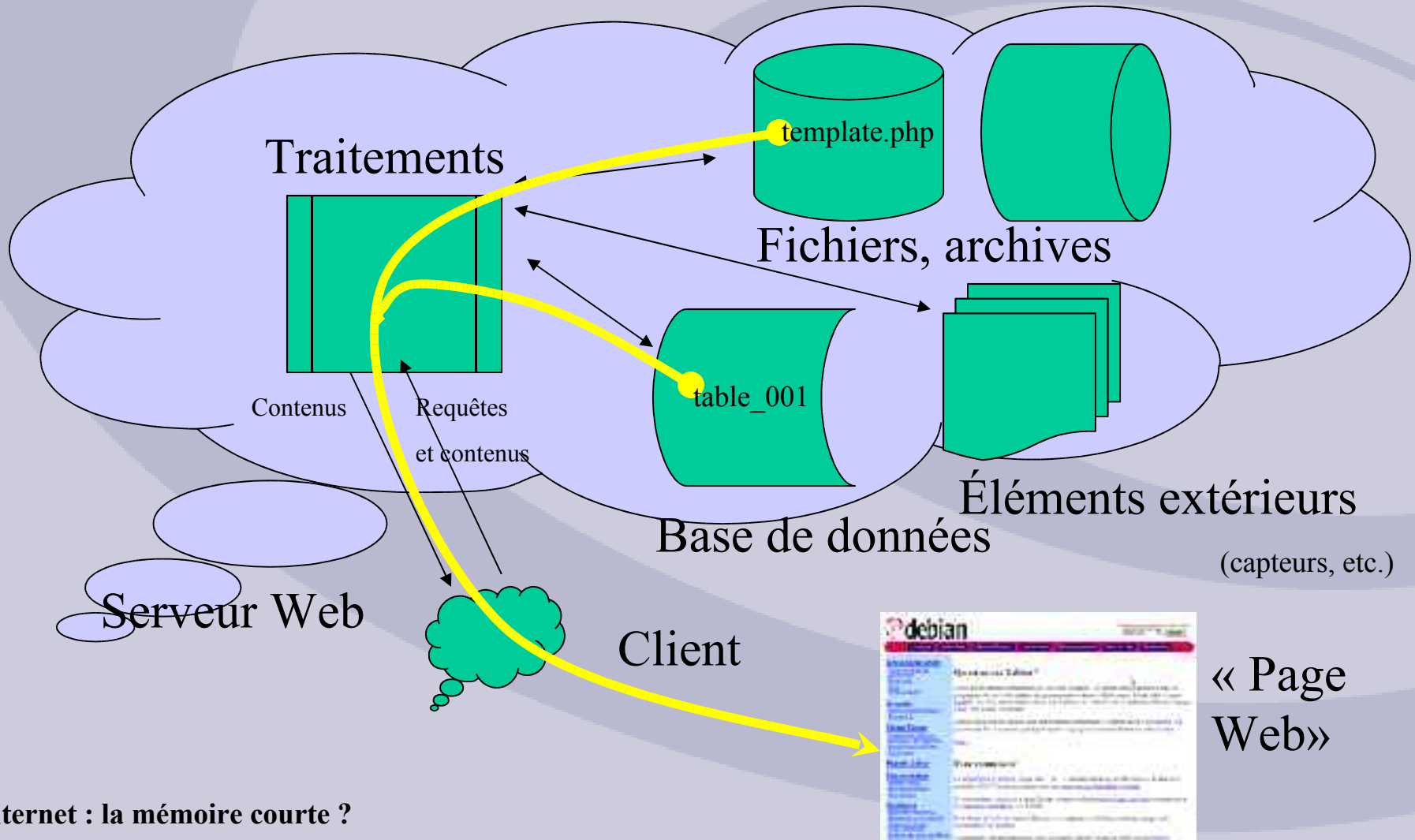
HTTrack WEBSITE COPIER

Un « serveur Web », qu'est-ce que c'est?



HTTrack WEBSITE COPIER

Le serveur web: un « livreur » de contenus



HTTrack WEBSITE COPIER

Les documents hypertexte

```
[Ligne 1] <html>
[Ligne 2] <head>
[Ligne 3] <title>
[Ligne 4] </title>
[Ligne 5] </head>
[Ligne 6] <body>
[Ligne 7] <h1>
[Ligne 8] </h1>
[Ligne 9] </body>
[Ligne 10] </html>
```

The screenshot shows the Debian website with a navigation bar at the top containing links for 'À propos', 'Actualité', 'Où est Debian?', 'Assistance', 'Développement', 'Plan du site', and 'Recherche'. The main content area is divided into several sections: 'À propos de Debian', 'Où est Debian?', 'Débian Debian', 'Documentation', and 'Assistance'. A yellow arrow points from the bottom of the code block on the left towards the 'Assistance' section on the right.

HTTrack WEBSITE COPIER

Les liens hypertexte

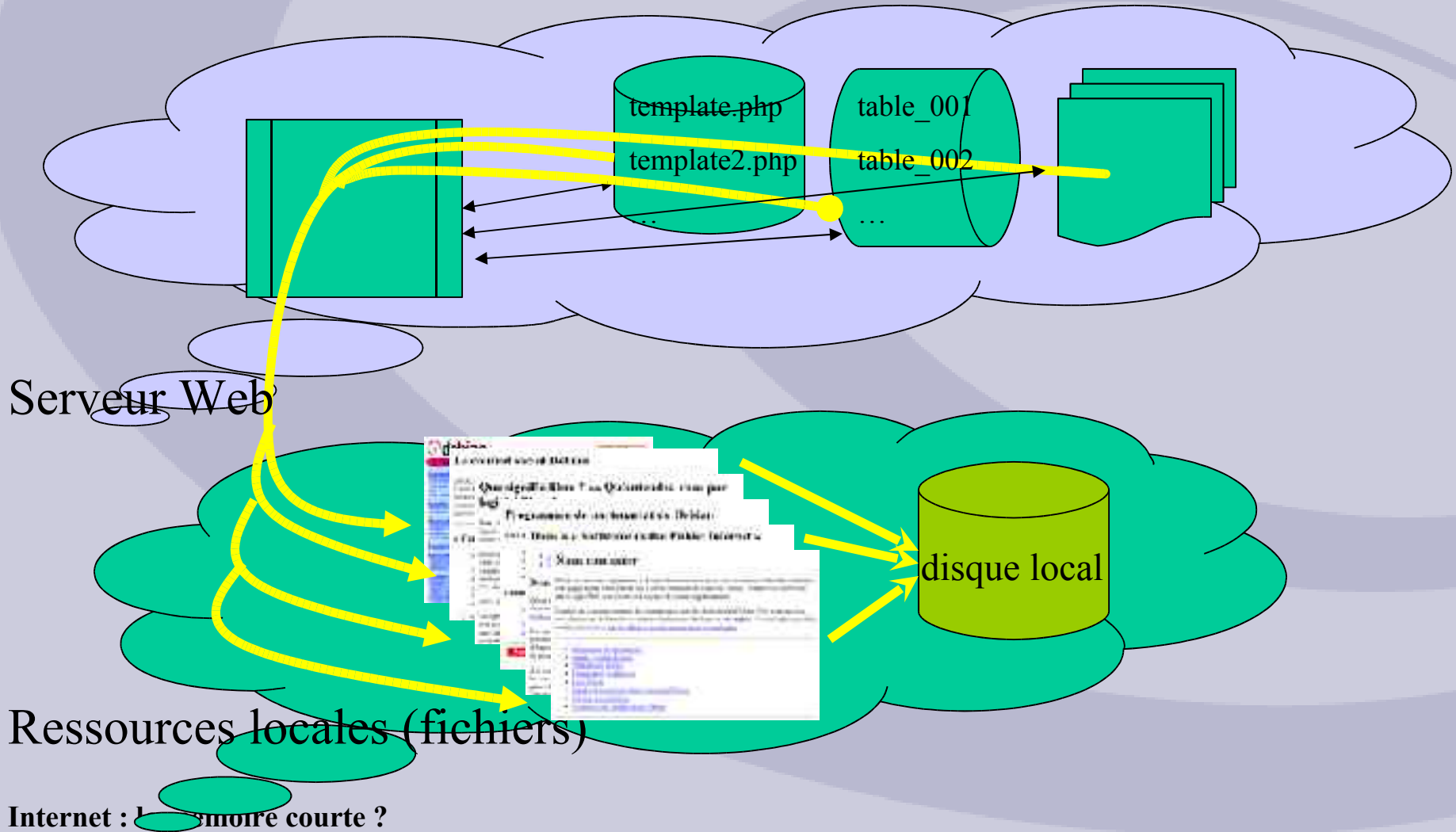
```
<td valign="top" width="140">  
<!-- VainComment -->  
<p><font face="Arial,Helvetica"><strong><a href="intro/about">&nbsp;&nbsp;&nbsp;propos</a>  
<small>  
&nbsp;&nbsp;&nbsp;<a href="./social_contract">Notre&nbsp;&nbsp;&nbsp;contrat&nbsp;&nbsp;&nbsp;social</a><br>  
&nbsp;&nbsp;&nbsp;<a href="./intro/free">Logiciel&nbsp;&nbsp;&nbsp;libre</a><br>  
&nbsp;&nbsp;&nbsp;<a href="./partners/">Partenaires</a><br>  
&nbsp;&nbsp;&nbsp;<a href="./donations">Dons</a><br>  
&nbsp;&nbsp;&nbsp;<a href="./contact">Nous&nbsp;&nbsp;&nbsp;contacter</a><br>  
</small>  
</font></p>  
<p><font face="Arial,Helvetica"><strong><a href="./News/">Actualités</a></font></p>
```



Copie locale d'un « site Web » ?

HTTrack WEBSITE COPIER

Copie locale d'un « site Web »

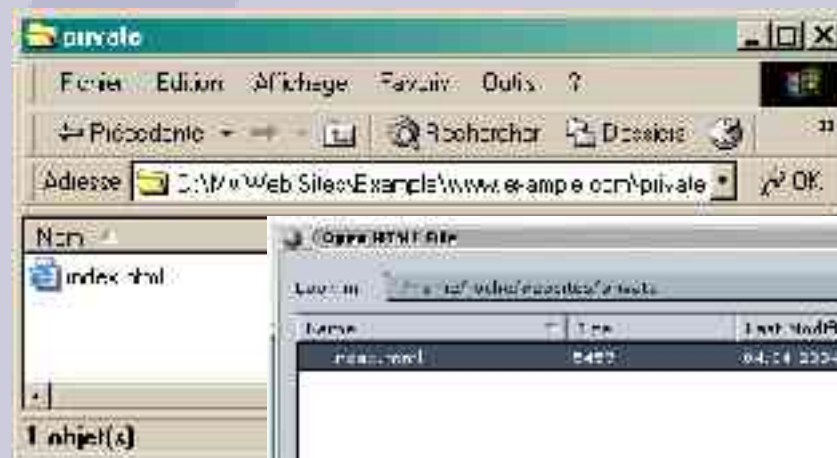


Le « nommage » local des
fichiers en ligne copiés

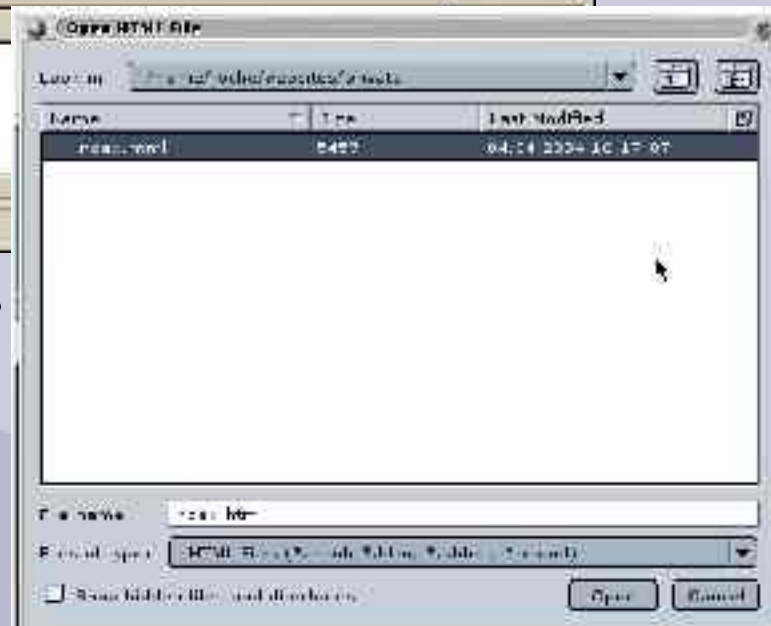
HTTrack WEBSITE COPIER

Nommage des fichiers copiés

- Exemple: fichier html



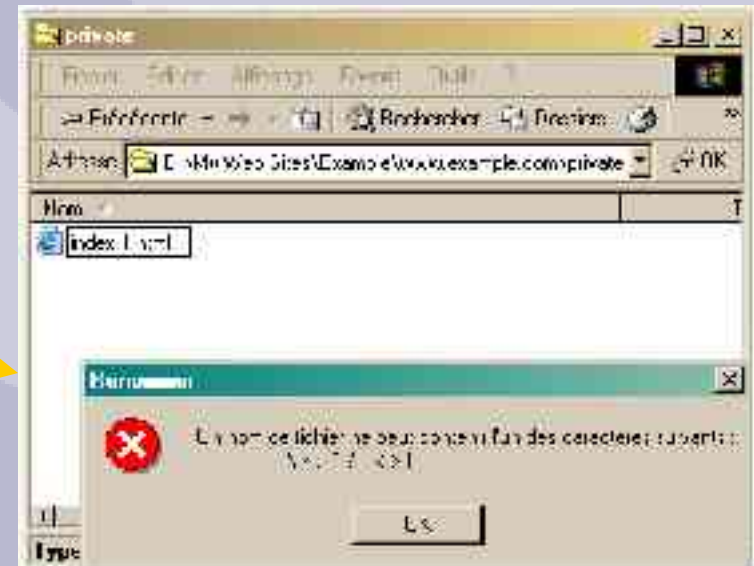
Windows



Linux/Unix

Nommage : restrictions

- Nommage des fichiers comportant des « caractères spéciaux »



Nommage : duplications

- Duplication de noms

The diagram illustrates a naming conflict during website copying. It shows two browser address bars: the top one contains `http://www.example.com/private/index.html` and the bottom one contains `http://www.example.com/private/INDEX.HTML`. A yellow plus sign is between them, and a yellow arrow points from the bottom address bar to a Windows error dialog box. The dialog box has a red 'X' icon and the text: "Erreur en renommant le fichier ou le dossier" and "Impossible de renommer l'avis de lecture. Le fichier portant ce nom n'existe pas. Choisissez un nom différent." Below the dialog box, a file explorer window shows a folder named "private" containing two files: "index.html" and "INDEX.HTML". The status bar at the bottom of the file explorer shows "Type : Document HTML 5,00 Ko" and "l'poste de travail".

Nommage : solutions

- Résoudre les collisions



Modification des liens hypertexte

```
<td valign="top" width="140">
  <!--Comment-->
  <p><font face="Arial,Helvetica"><strong><a href="intro/about">À propos</a>
  <small>
    <sub><a href="/social_contract">Notre</sub><sub>cont</sub><sub><a href="/social">social</a></a><br>
    <sub><a href="/intro/free">Logiciel</sub><sub><a href="/libre">libre</a></a><br>
    <sub><a href="/partners/">Partenaires</a></a><br>
    <sub><a href="/donations">Dons</a></a><br>
    <sub><a href="/contact">Nous</sub><sub><a href="/contact">contact</a></a><br>
  </small>
</font></p>
<p><font face="Arial,Helvetica"><strong><a href="/News/">actualités</a></p>
```



```
<td valign="top" width="140">
  <!--Comment-->
  <p><font face="Arial,Helvetica"><strong><a href="http://www.debian.org/intro/about">À propos</a>
  <small>
    <sub><a href="social_contract.html">Notre</sub><sub>cont</sub><sub><a href="social">social</a></a><br>
    <sub><a href="intro/free.html">Logiciel</sub><sub><a href="/libre">libre</a></a><br>
    <sub><a href="http://www.debian.org/partners/">Partenaires</a></a><br>
    <sub><a href="http://www.debian.org/donations">Dons</a></a><br>
    <sub><a href="contact.html">Nous</sub><sub><a href="/contact">contact</a></a><br>
  </small>
</font></p>
<p><font face="Arial,Helvetica"><strong><a href="http://www.debian.org/News/">actualités</a></p>
```

Les problèmes apparaissent!

Les problèmes apparaissent!

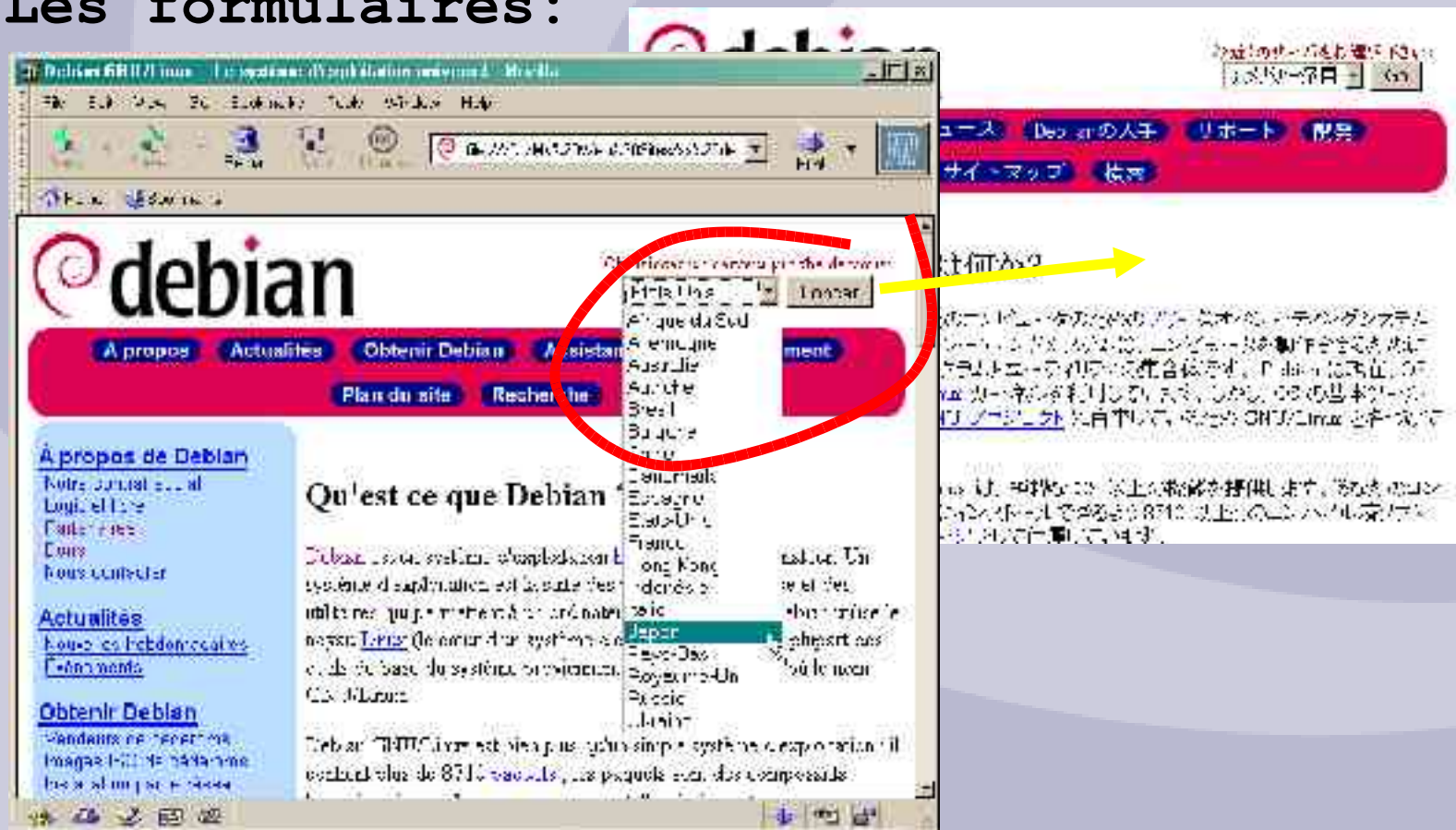
- Les liens :

- ``
- ``
- ``
- `<a href= "page 2`
`.html">`
- ``
- ``
- ``
- `<a href >`

HTTrack WEBSITE COPIER

Les problèmes apparaissent!

- Les formulaires:



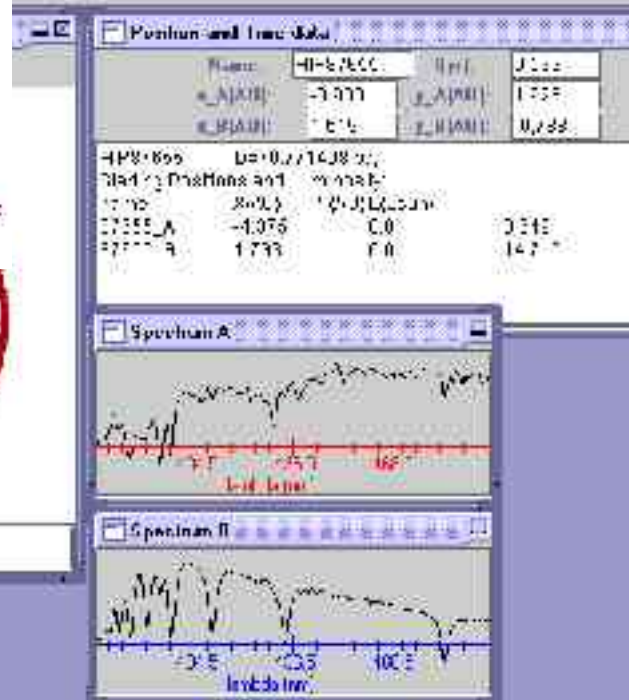
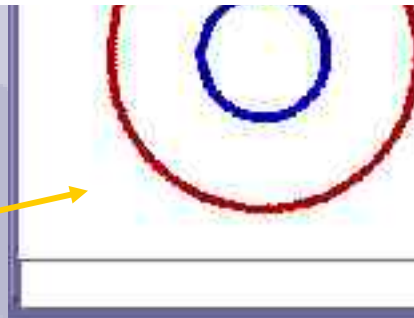
HTTrack WEBSITE COPIER

Les problèmes apparaissent!

- Java :

```
ghtml:
<head>
<title>Sun, 01 Jun 1999 10:53:00 GMT</title>
</head>
<body>
Please be patient while the pages load into your browser. Once it
loads, this applet will display a number of arrays whose elements
are numbers. Below the system has been labeled "Primary System",
and the arithmetic will occur. You'll see a circle under each one
red dot and one blue dot. Clicking within this circle will display
the coordinates in hexadecimal form (00). The number is labeled
"Position and Time Data", the elapsed time in years, and position in
x and y in 10^10 units, and it will be constantly updating. Please click
on the load blocks like this:
<img alt="load" data-bbox="310 550 330 570"/>
The upper right corner of the
"main" window before entering a second window will be the
display of a table.
</body>
</html>
<applet ARCHIVE="http://www.javasoft.com" CODE="SpecApplet.class" WIDTH=300 HEIGHT=180></applet>
</body>
</html>
```

```
180 1999-1-32 10:53 dot
180 1999-01-32 10:53 data
6006 1999-1-32 10:53 Format.class
12933 1999-1-32 10:53 LoadStarData.class
60 1999-01-32 10:53 HI III-IN-
5577 1999-1-32 10:53 MyMenuBar.class
5877 1999-1-32 10:53 MyMenuBar#MenuAction.class
5539 1999-1-32 11:05 SpecApplet.class
5878 1999-1-32 10:53 SpecDialog.class
6186 1999-1-32 10:53 SpecPanel.class
2528 1999-1-32 11:05 StarMenu.class
682 1999-1-32 11:05 StarMenu#MenuAction.class
11296 1999-1-32 10:53 StarPanel.class
897 1999-1-32 10:53 StarPanel#MenuAction.class
```



Les problèmes apparaissent!

- « Horodatage » intégré aux liens hypertexte

`http://www.example.com/page2.html?t=19993112235959999`

- Liens multiples vers un seul document

`http://www.example.com/forum/article.php?id=1234`

`http://www.example.com/forum/article.php?id=1233&next`

`http://www.example.com/forum/article.php?id=5678&previous`

`http://www.example.com/forum/article.php?id=6548&previous10`

`http://www.example.com/forum/article.php?id=879&next10`

...

- Etc etc etc

Aperçu de quelques autres problèmes

- Taille limite des fichiers
- Gestion des erreurs, des liens cassés
- Sites protégés par mot de passe
- Sites utilisant des « cookies » / des sessions
- Fichiers locaux « Intranet » (file://)
- Sites sécurisés (HTTPS)
- Sites ftp
- Sites accessibles via Ipv6 uniquement (recherche, universités)

Mise à jour ?

Mise à jour ?



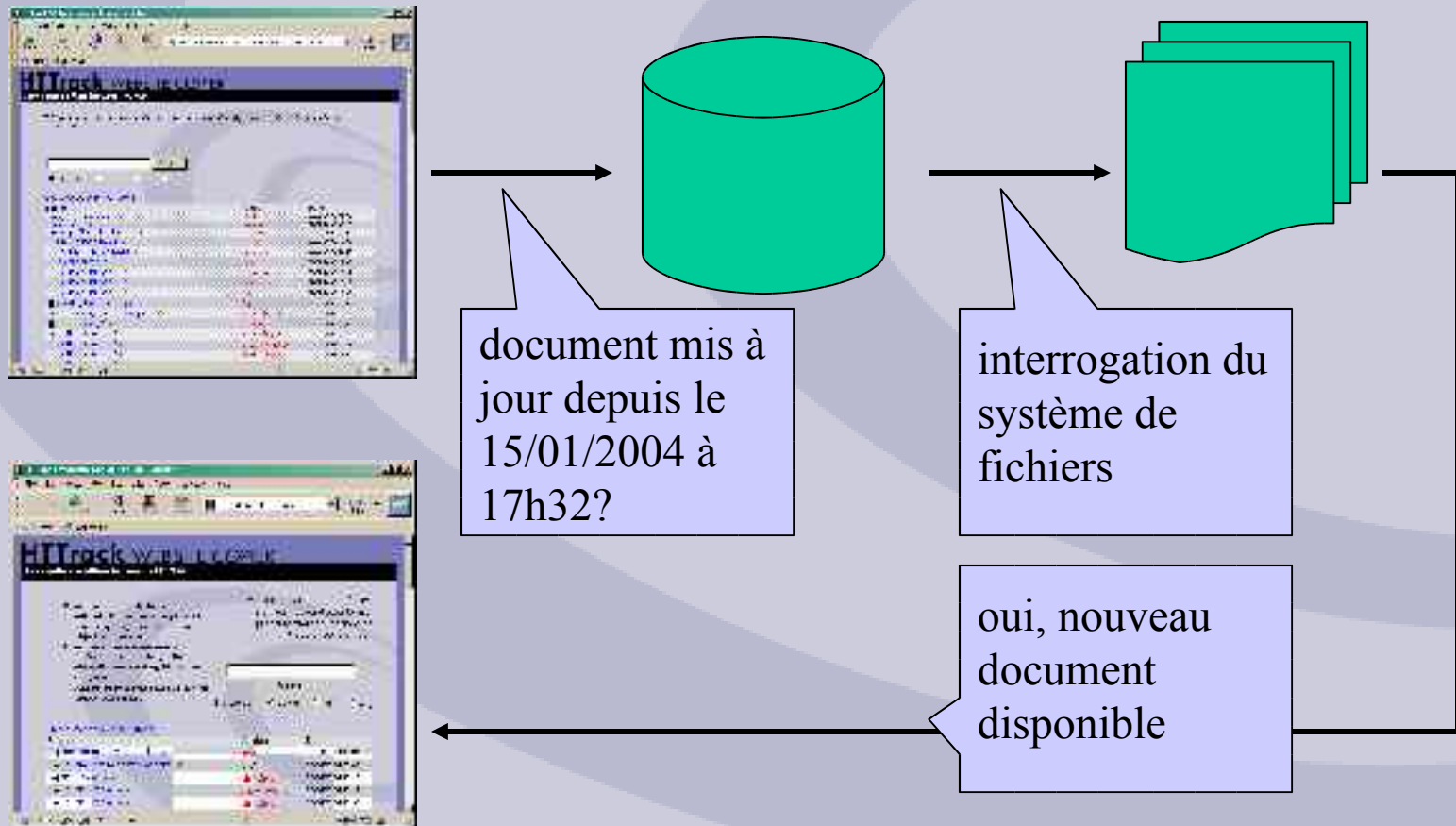
Document capturé le
15/01/2004 à 17h32



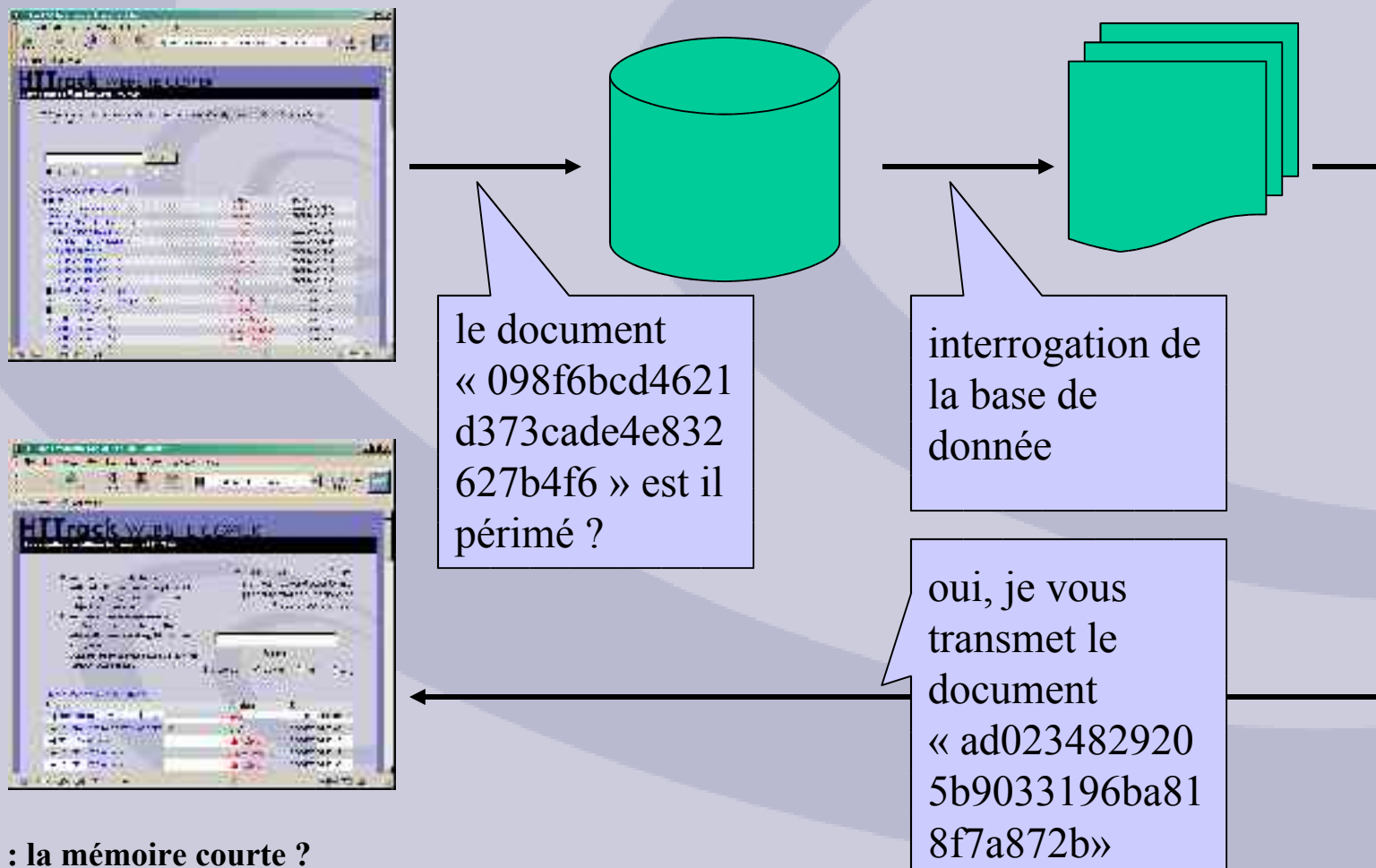
Une version plus
récente est-elle
disponible
aujourd'hui ?

- Économie de bande passante
- Économie de temps
- Économie d'espace de stockage

Mise à jour « incrémentale » (1)



Mise à jour « incrémentale » (2)



Les précautions à prendre lors de la capture d'un site

HTTrack WEBSITE COPIER

Les précautions à prendre : aspects légaux ?

- Copie privée / publique ?
- Protection du site ? (loi n°95-597 du 1er juillet 1992 , art 1 353-3 du CPI)
- Statut d'un aspirateur de sites Web ?



Navigateur?



Robot?



Proxy-cache?

Conclusion

- ...