Capture de sites Web en ligne

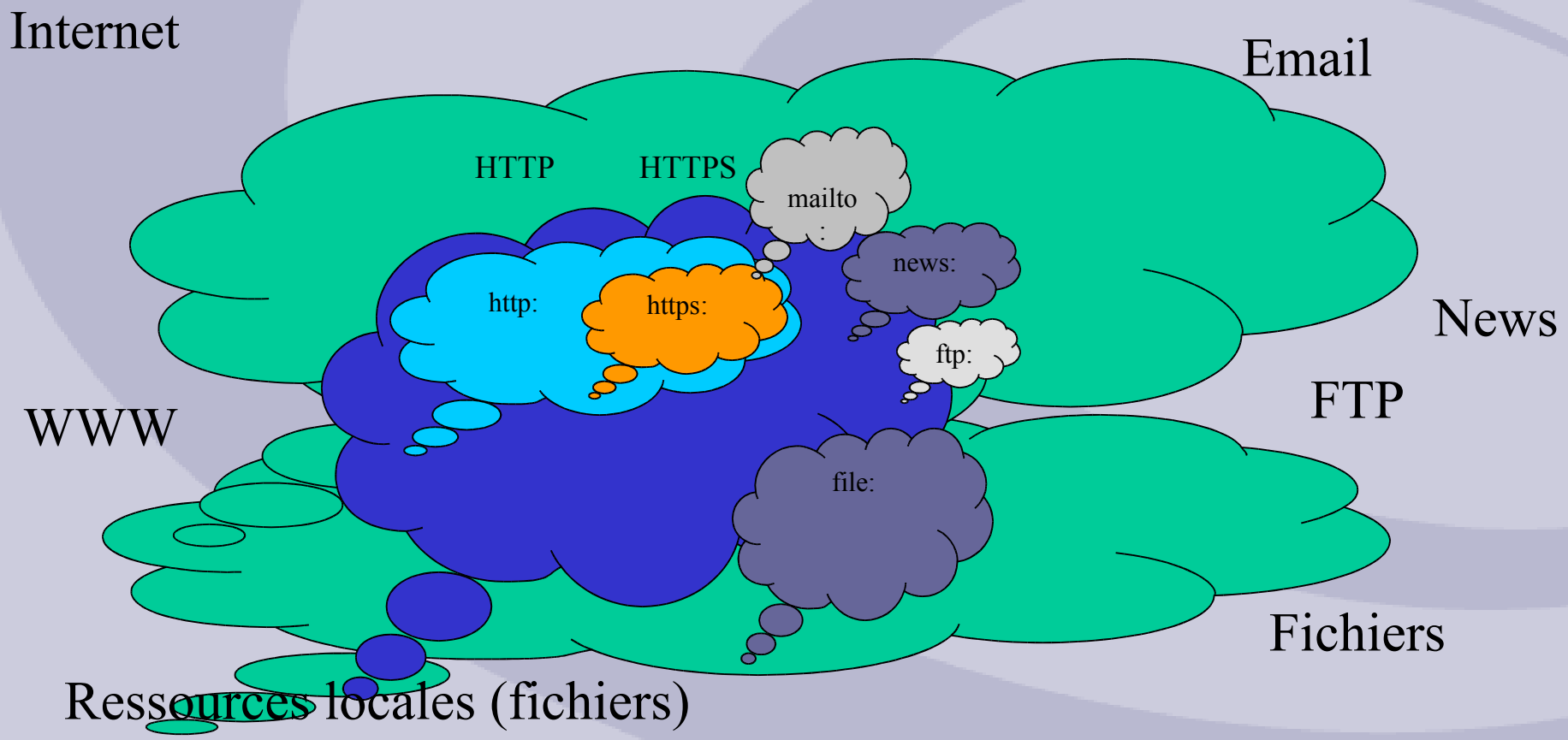
Conférence B.N.F, Avril 2004

Xavier Roche(HTTrack)

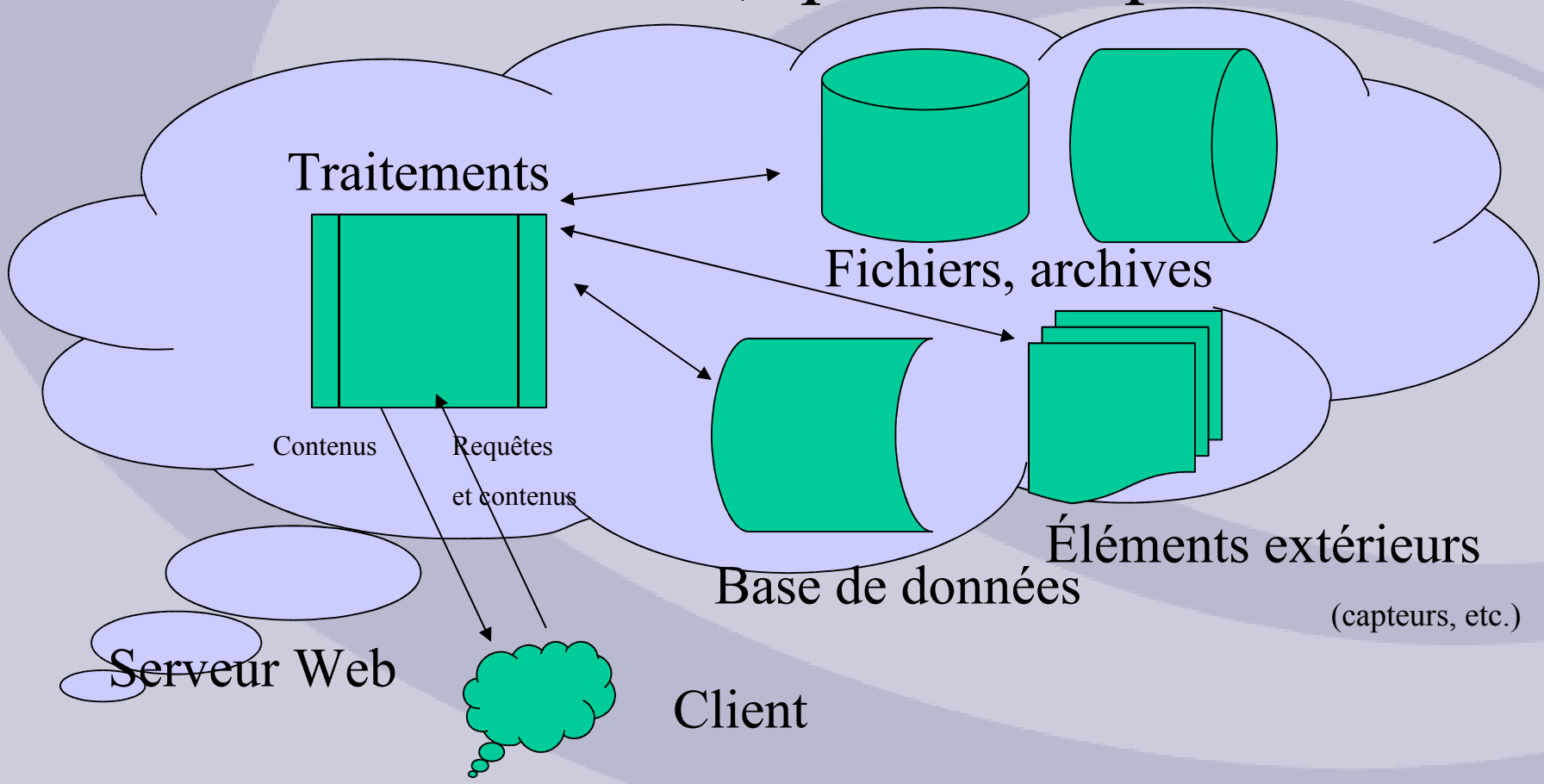
Pourquoi copier des sites web?

- 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…

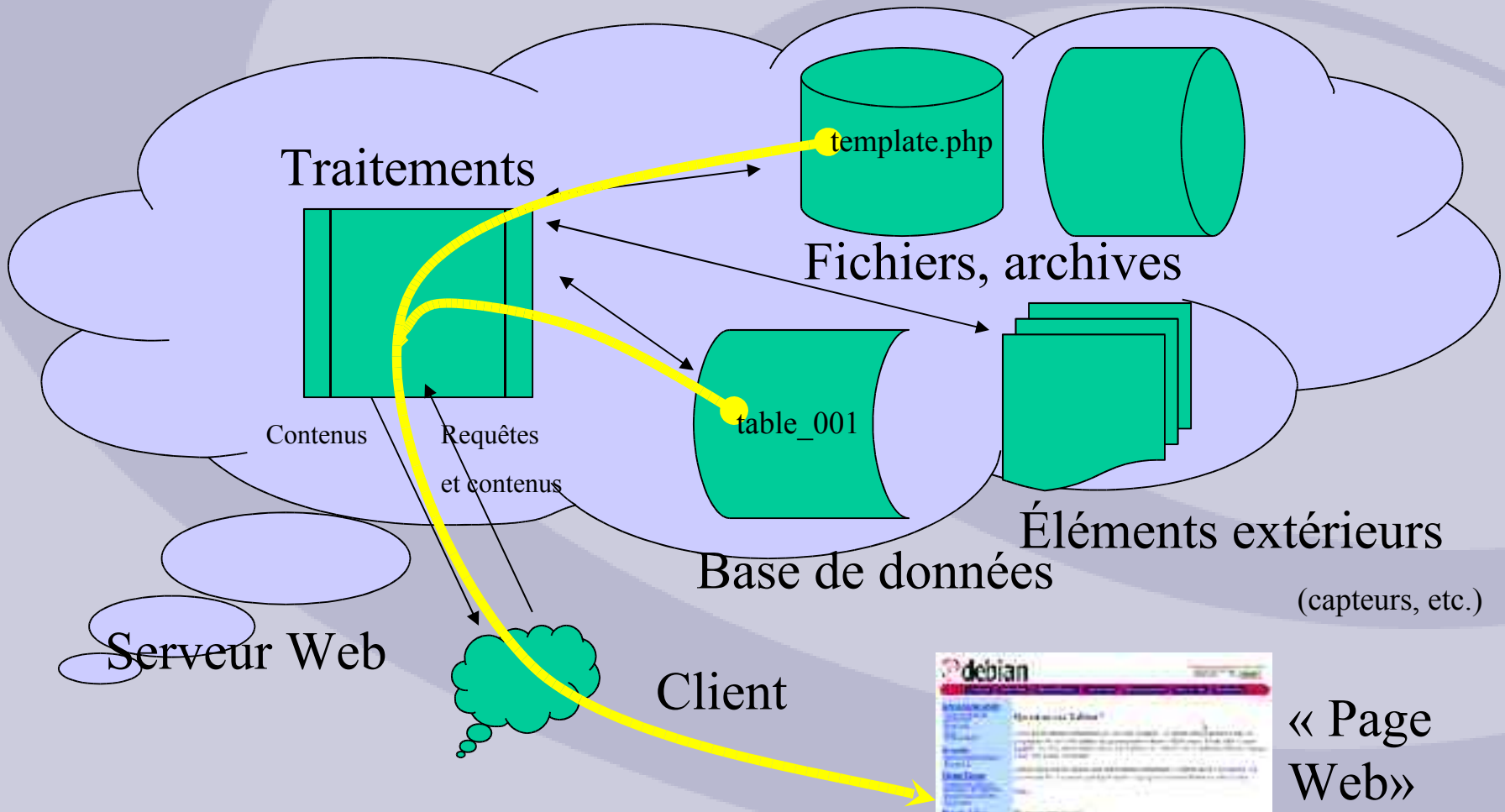
Le « Web », qu'est-ce que c'est ?



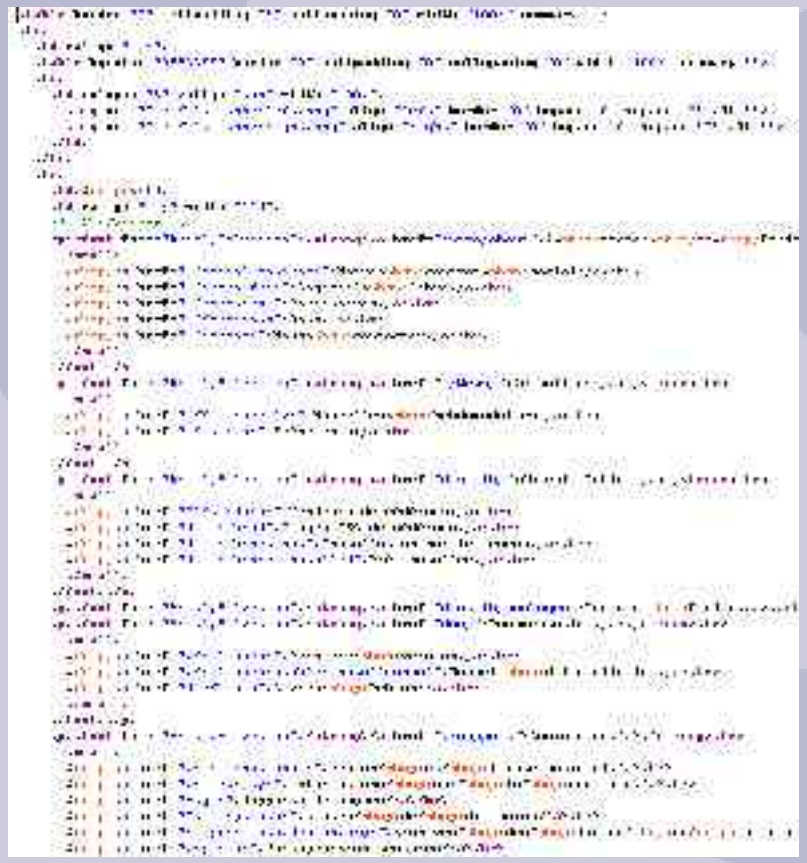
Un « serveur Web », qu'est-ce que c'est?



Le serveur web: un « livreur » de contenus



Les documents hypertexte



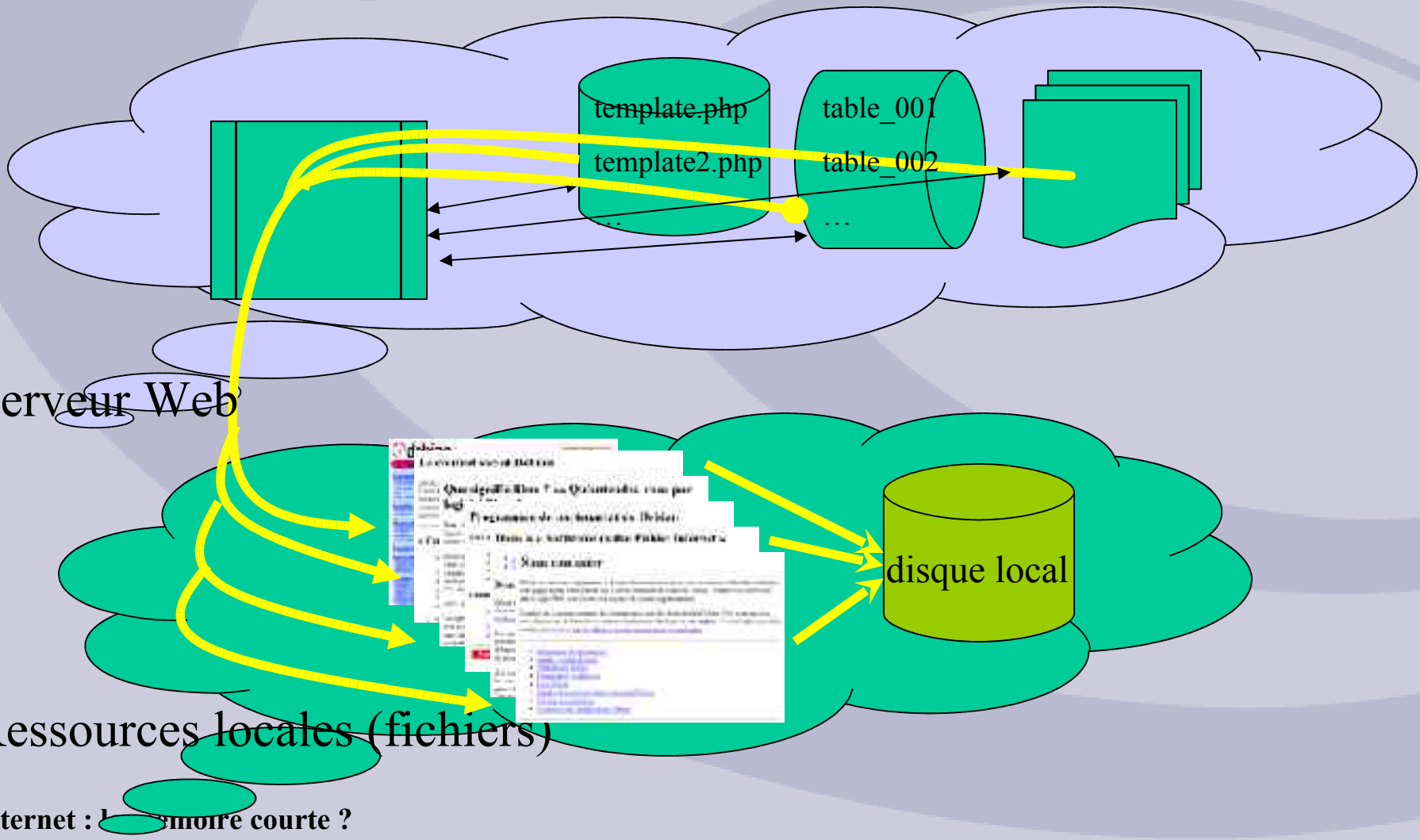
Les liens hypertexte

```
<td valign="top" width="140">
<!--VdmComment-->
<p><font face="Arial,Helvetica"><strong><a href="intro/about">À propos&
  <small>
   <a href="./ccccial_contract">Notre contrat social</a><br>
   <a href="./[illegible]">Logiciel libre</a><br>
   <a href="./partners/">Partenaires</a><br>
   <a href="./donations">Dons</a><br>
   <a href="./contact">Nous contacter</a><br>
  </small>
</font></p>
<p><font face="Arial,Helvetica"><strong><a href="./News/">Actualités</a></st
```

Copie locale d'un « site Web» ?

HTTrack WEBSITE COPIER

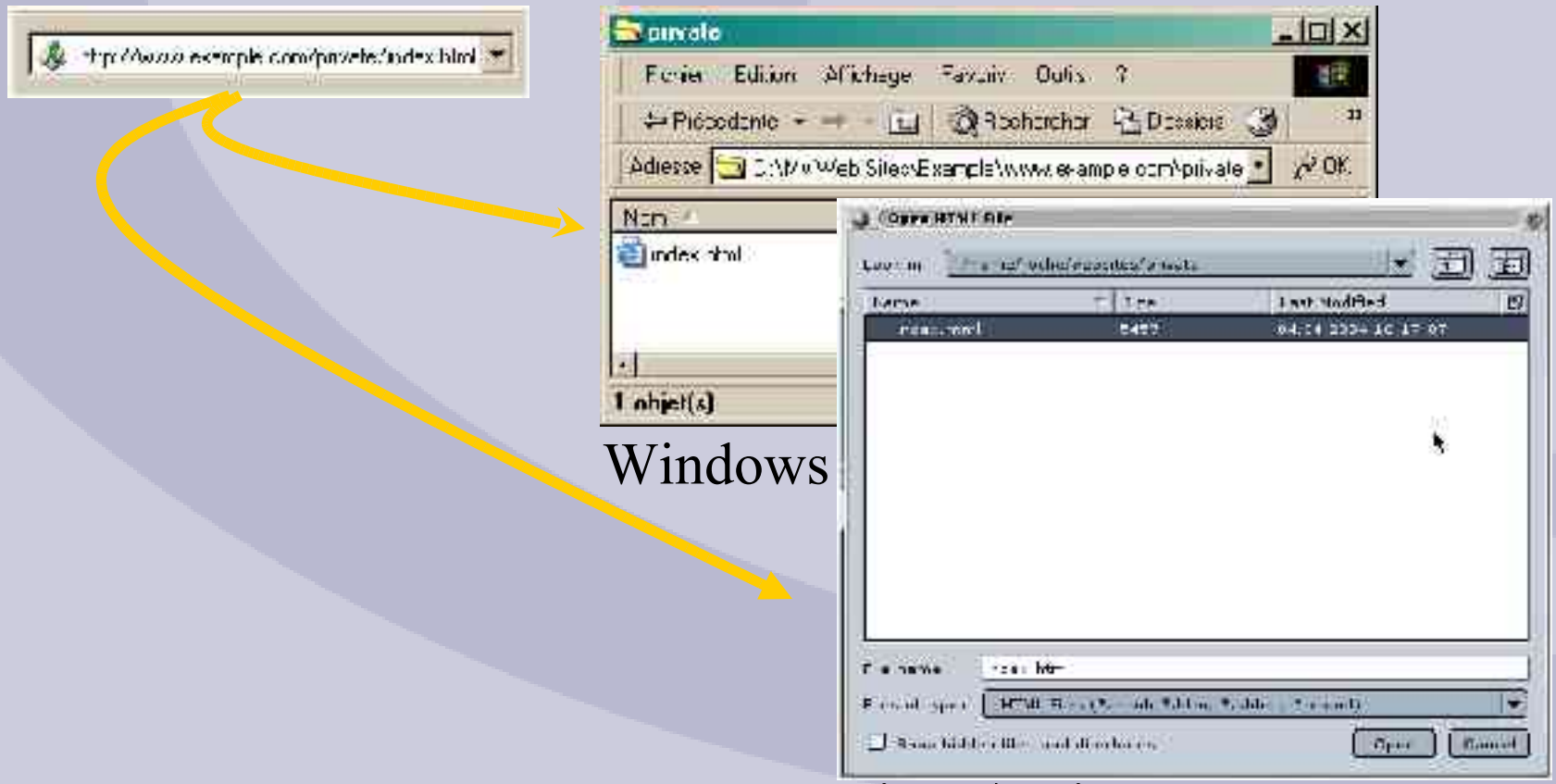
Copie locale d'un « site Web»



Le « nommage » local des fichiers en ligne copiés

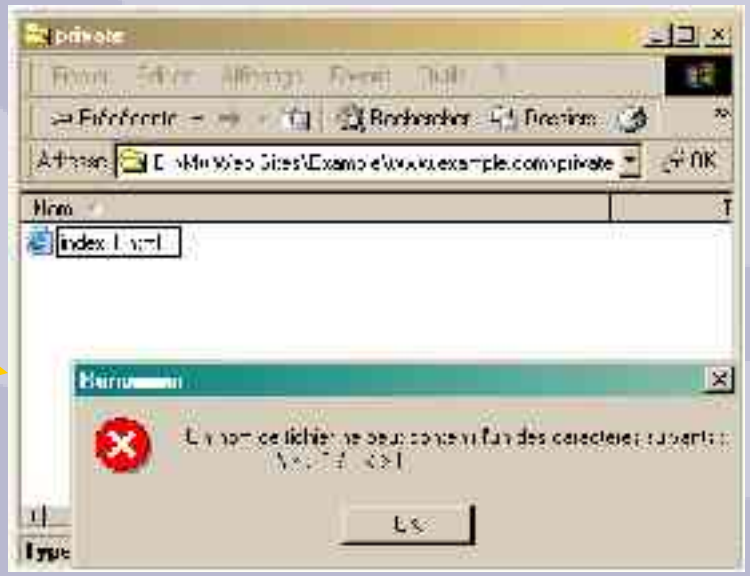
Nommage des fichiers copiés

- **`Exemple: fichier html`**



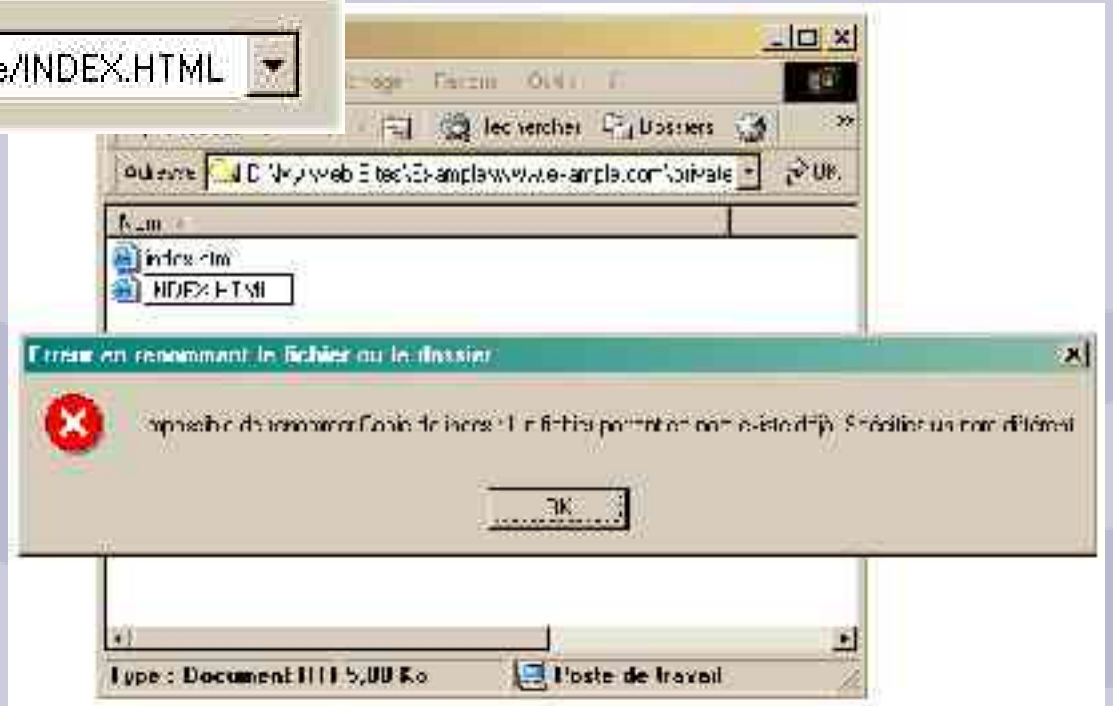
Nommage : restrictions

- **Nommage des fichiers comportant des « caractères spéciaux »**



Nommage : duplications

- **Duplication de noms**



Nommage : solutions

- **Résoudre les collisions**

Modification des liens hypertexte

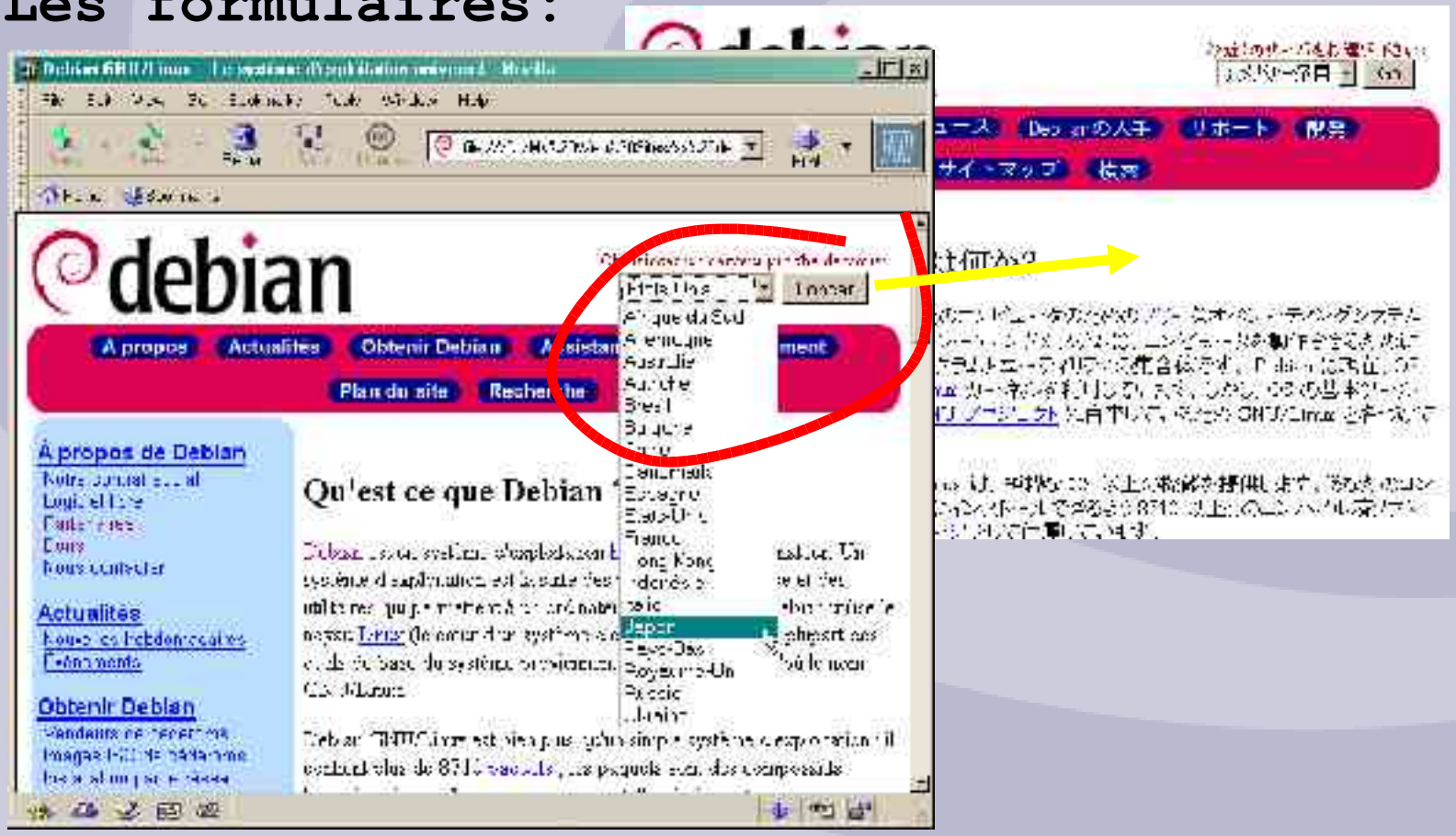
Les problèmes apparaissent!

Les problèmes apparaissent!

- **Les liens:**
- `<a href= 'page 2.html'>`
- `<a href= "page%202.html">`
- `<a href= page%202.html>`
- `<a href= "page 2 .html">`
- `<a href= "http:page 2.html">`
- `<a href= "//www.example.com/page 2.html">`
- `<a href= "page 2.html">`
- `<a href  <a href= "page2.html">>`

Les problèmes apparaissent!

- **Les formulaires:**

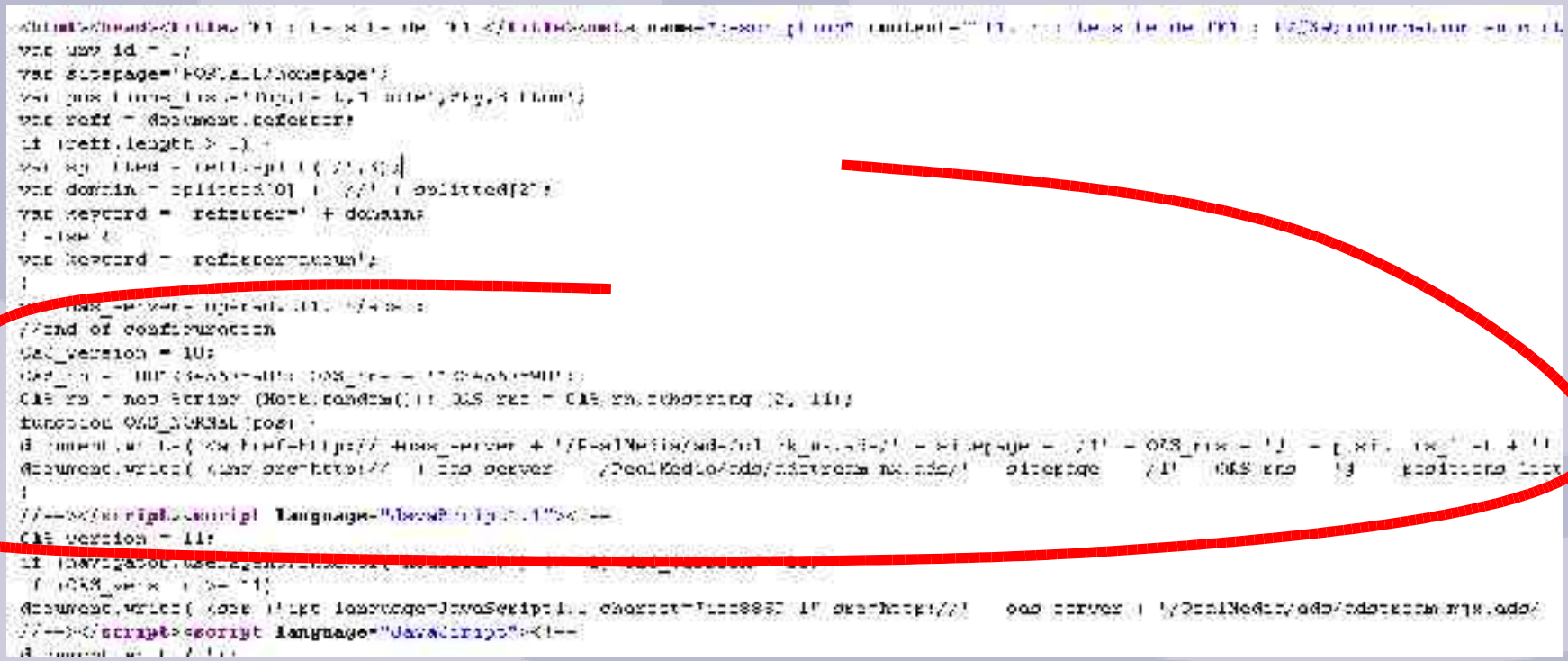


Les problèmes apparaissent!

- **Les formulaires (suite) :**

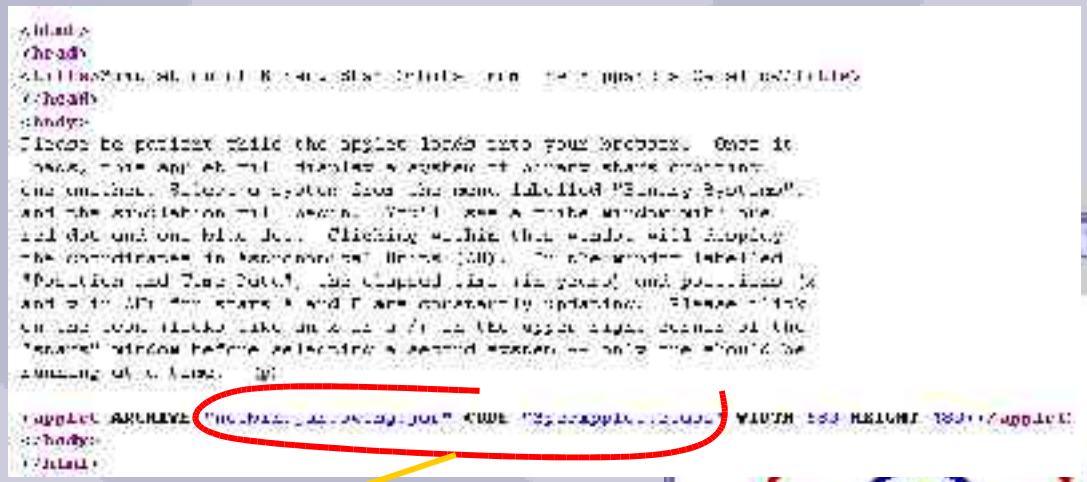
Les problèmes apparaissent!

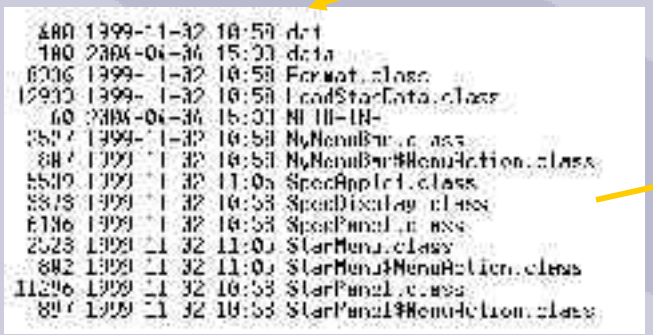
- **Javascript :**

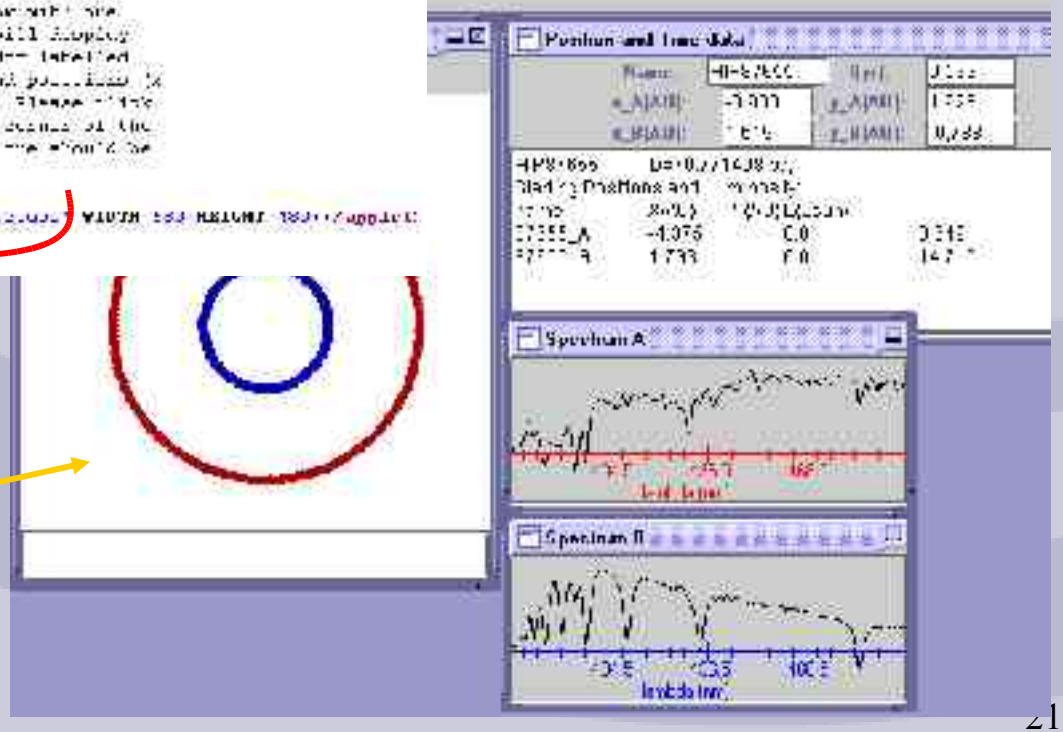


Les problèmes apparaissent!

- **Java :**



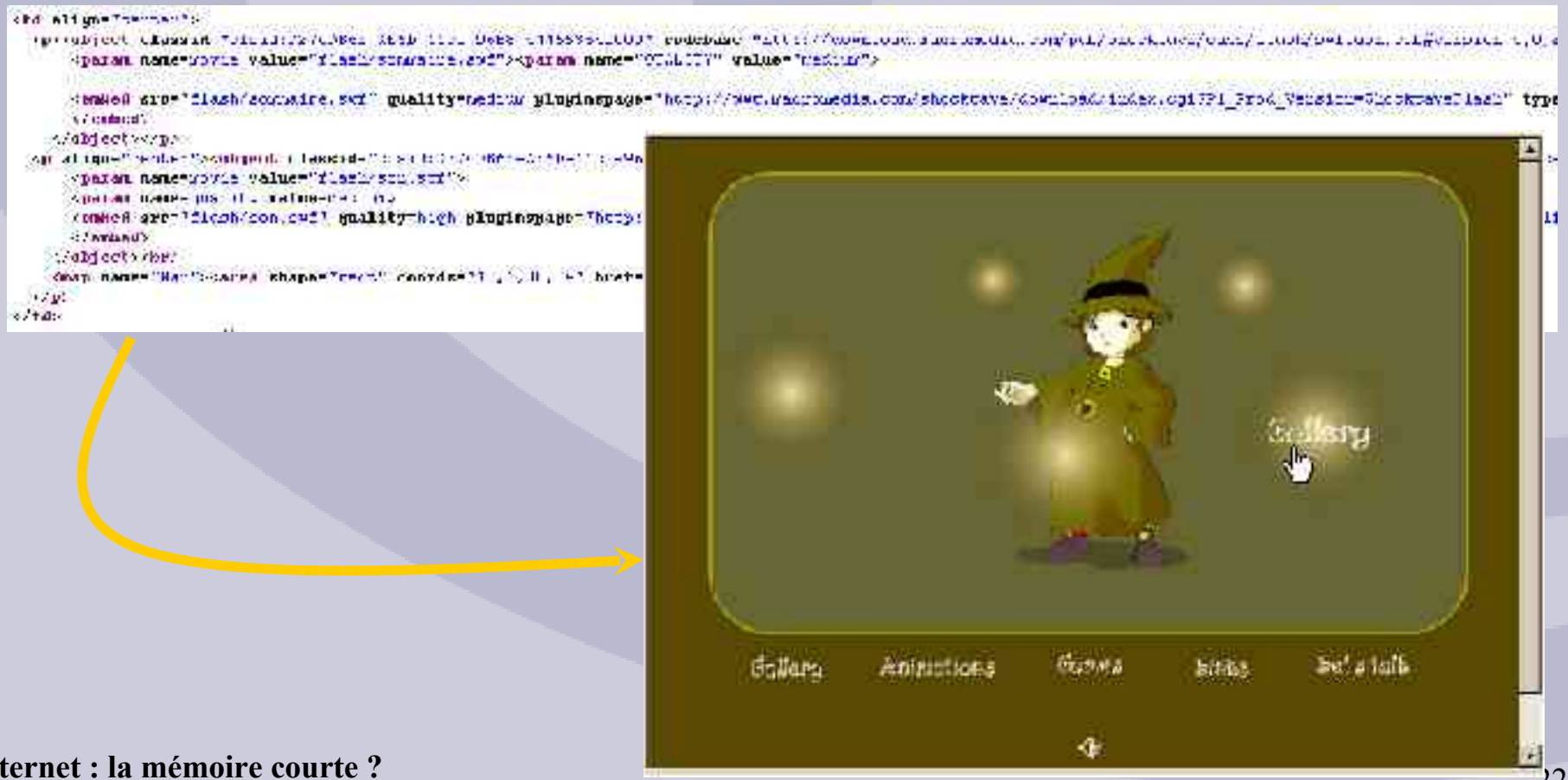




HTTrack WEBSITE COPIER

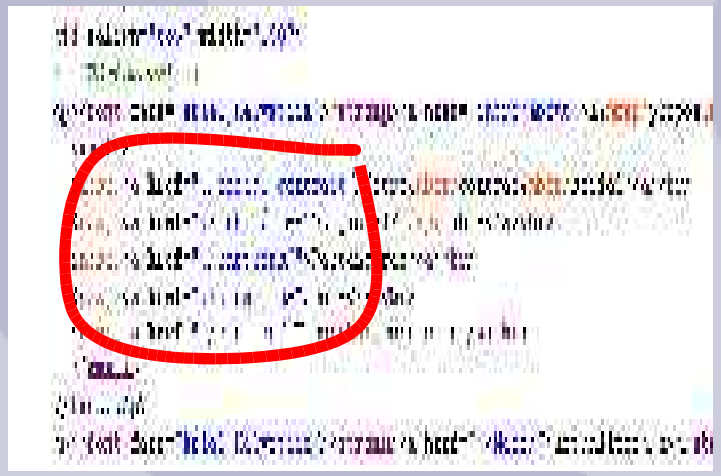
Les problèmes apparaissent!

- **Flash :**

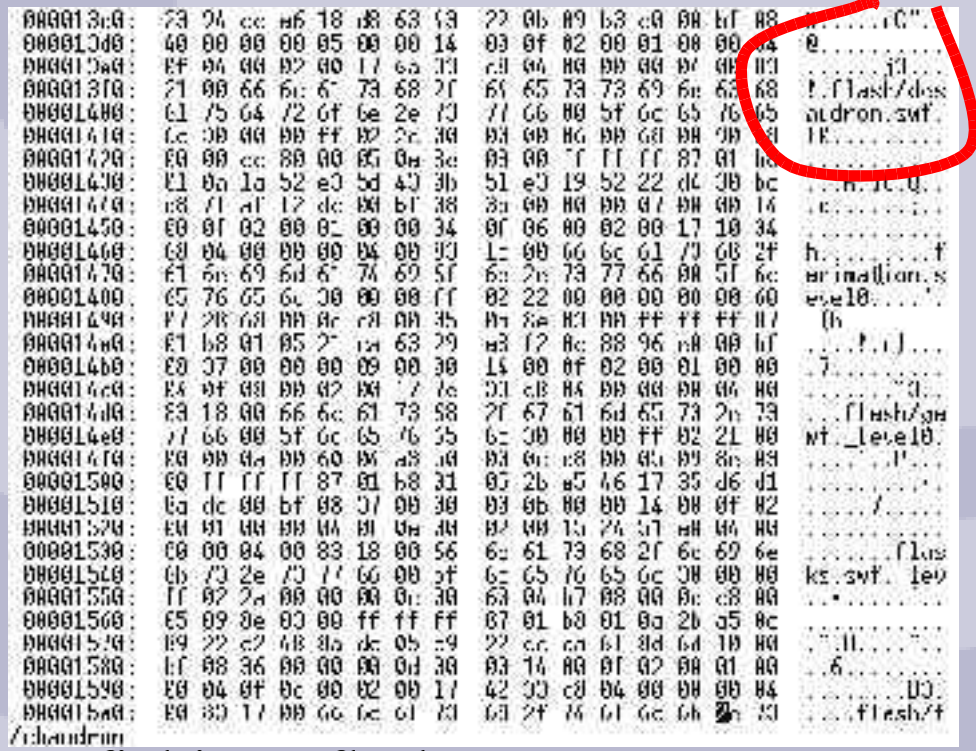


Les problèmes apparaissent!

- **Fichiers hypertextes vs binaires (Java/Flash) :**



un fichier html



un fichier « flash »

Les problèmes apparaissent!

- **« Horodatage » intégré aux liens hypertexte**

http://www.example.com/page2.html?t=1999311223595999

- **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 ?

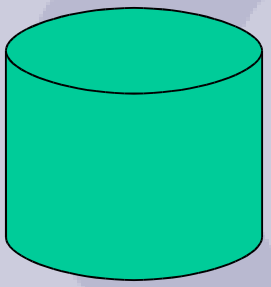
Mise à jour ?

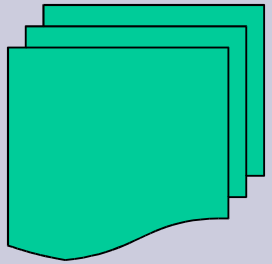
Document capturé le 15/01/2004 à 17h32

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)



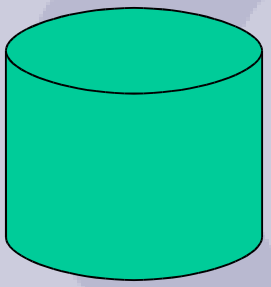


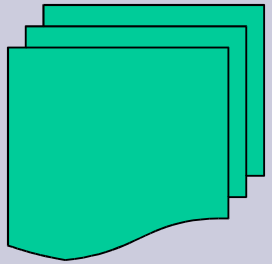
document mis à jour depuis le 15/01/2004 à 17h32?

interrogation du système de fichiers

oui, nouveau document disponible

Mise à jour « incrémentale » (2)





le document « 098f6bcd4621 d373cade4e832 627b4f6 » est il périmé ?

interrogation de la base de donnée

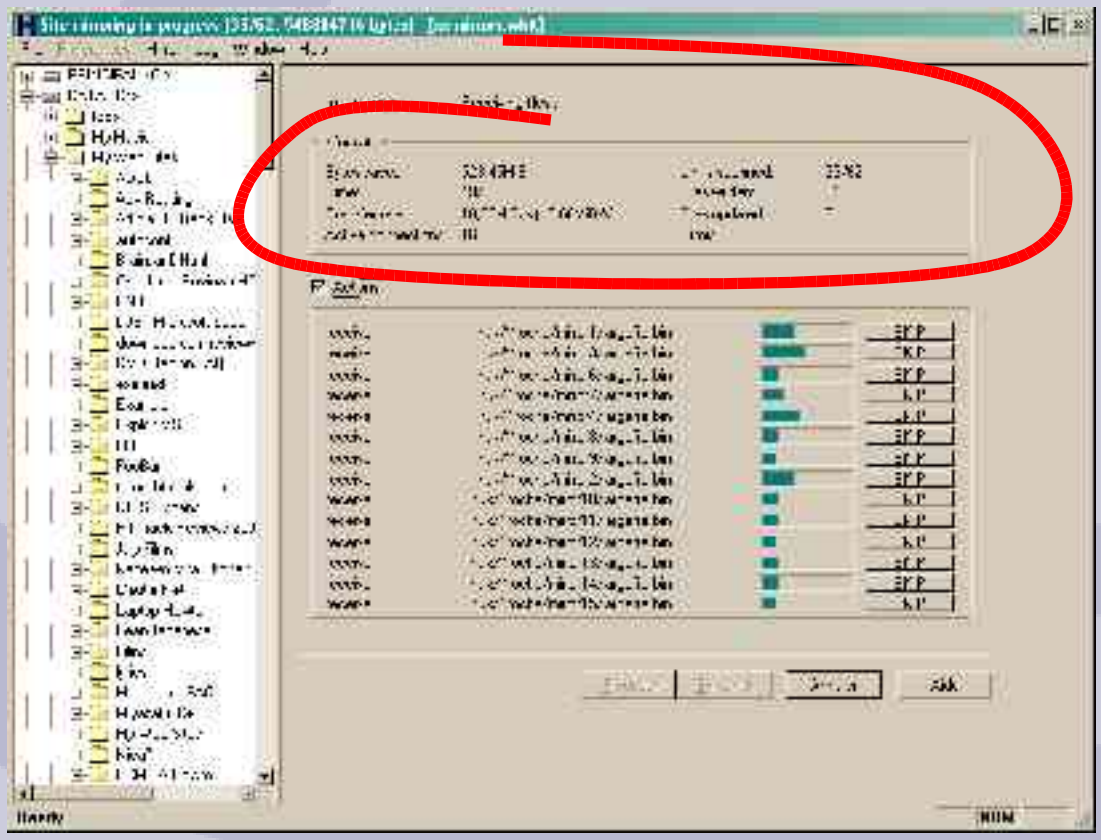
oui, je vous transmet le document « ad023482920 5b9033196ba81 8f7a872b»

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

HTTrack WEBSITE COPIER

Les précautions à prendre : surcharge du site

- Limiter la bande passante et le nombre de connexions simultanées!



Les précautions à prendre : aspects légaux ?

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

Navigateur?

Robot?

Proxy-cache?

Conclusion

- …