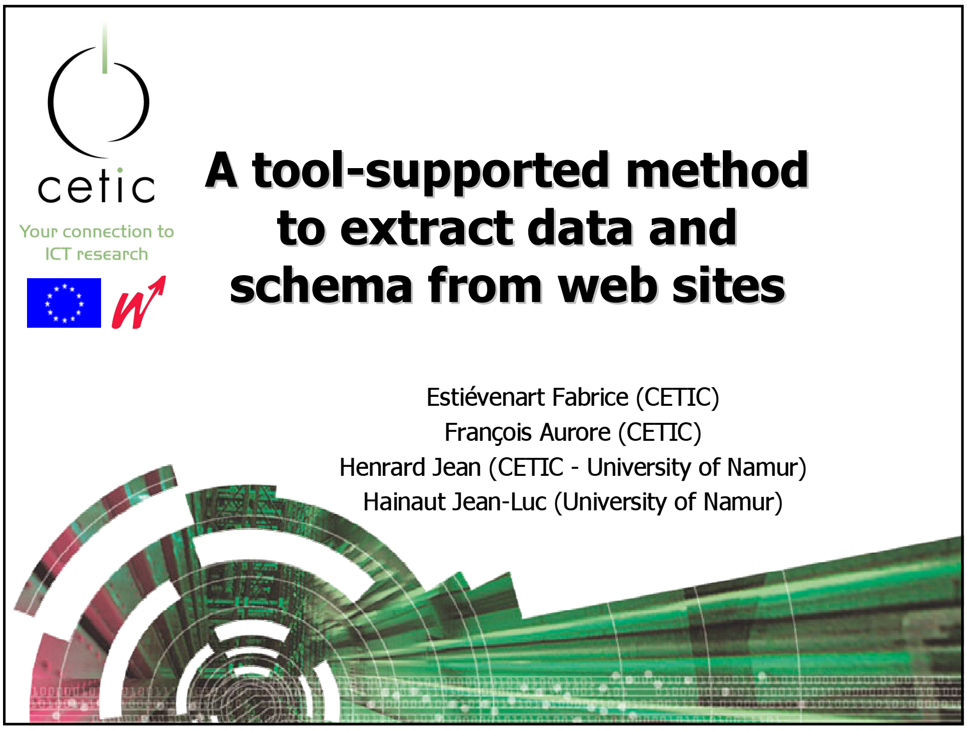
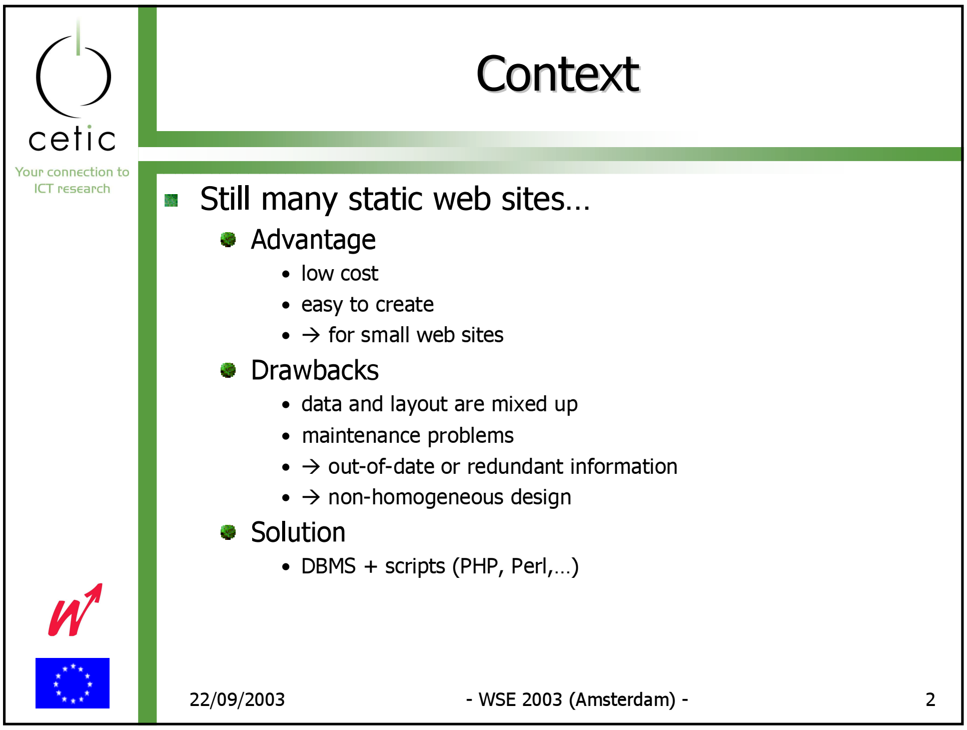


A tool-supported method to extract data and schema from web sites

Estiévenart Fabrice (CETIC)
François Aurore (CETIC)
Henrard Jean (CETIC - University of Namur)
Hainaut Jean-Luc (University of Namur)

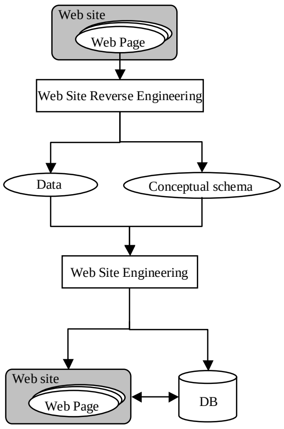


Context

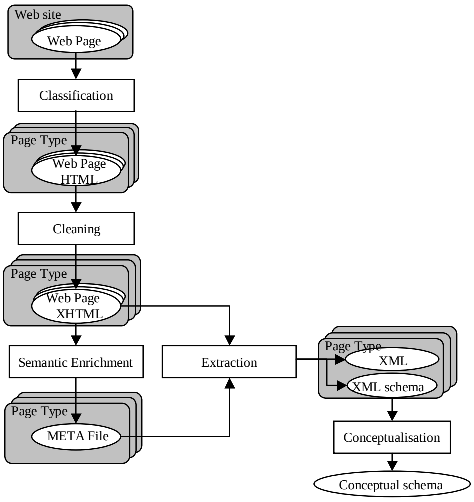
- Still many static web sites…
  - Advantage
    - low cost
    - easy to create
    - → for small web sites
  - Drawbacks
    - data and layout are mixed up
    - maintenance problems
    - → out-of-date or redundant information
    - → non-homogeneous design
  - Solution
    - DBMS + scripts (PHP, Perl,…)

Goals

- To provide methods and tools for web sites reengineering :

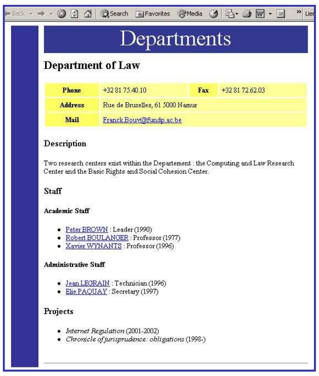


Method : overview



Method : step 1

- Pages classification
  - « Page type » = a set of pages relative to the same concept, that display very similar information in a very similar layout



Departments

Department of Law

Phone	+32 81 75.40.10	**Fax**	+32 81 72.62.03
Address	Rue de Bruxelles, 61 5000 Namur		
Mail	Franck.Bouv@fundp.ac.be		

Description

Two research centers exist within the Departement : the Computing and Law Research Center and the Basic Rights and Social Cohesion Center.

Staff

Academic Staff

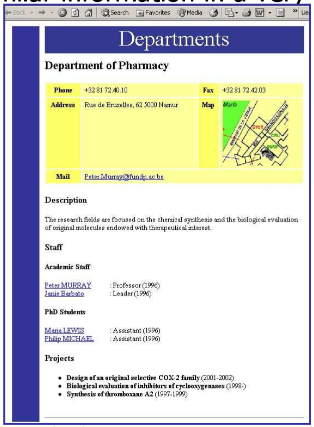
- Peter BROWN : Leader (1990)
- Robert BOULANGER : Professor (1977)
- Xavier WYNANTS : Professor (1996)

Administrative Staff

- Jean LEGRAIN : Technician (1996)
- Elie PAQUAY : Secretary (1997)

Projects

- *Internet Regulation* (2001-2002)
- *Chronicle of jurisprudence: obligations* (1998-)



Departments

Department of Pharmacy

Phone	+32 81 72.40.10	**Fax**	+32 81 72.42.03
Address	Rue de Bruxelles, 62 5000 Namur	**Map**	
Mail	Peter.Murray@fundp.ac.be		

Description

The research fields are focused on the chemical synthesis and the biological evaluation of original molecules endowed with therapeutical interest.

Staff

Academic Staff

Peter MURRAY	: Professor (1996)
Janie Barbato	: Leader (1996)

PhD Students

Maria LEWIS	: Assistant (1996)
Philip MICHAEL	: Assistant (1996)

Projects

- **Design of an original selective COX-2 family** (2001-2002)
- **Biological evaluation of inhibitors of cyclooxygenases** (1998-)
- **Synthesis of thromboxane A2** (1997-1999)

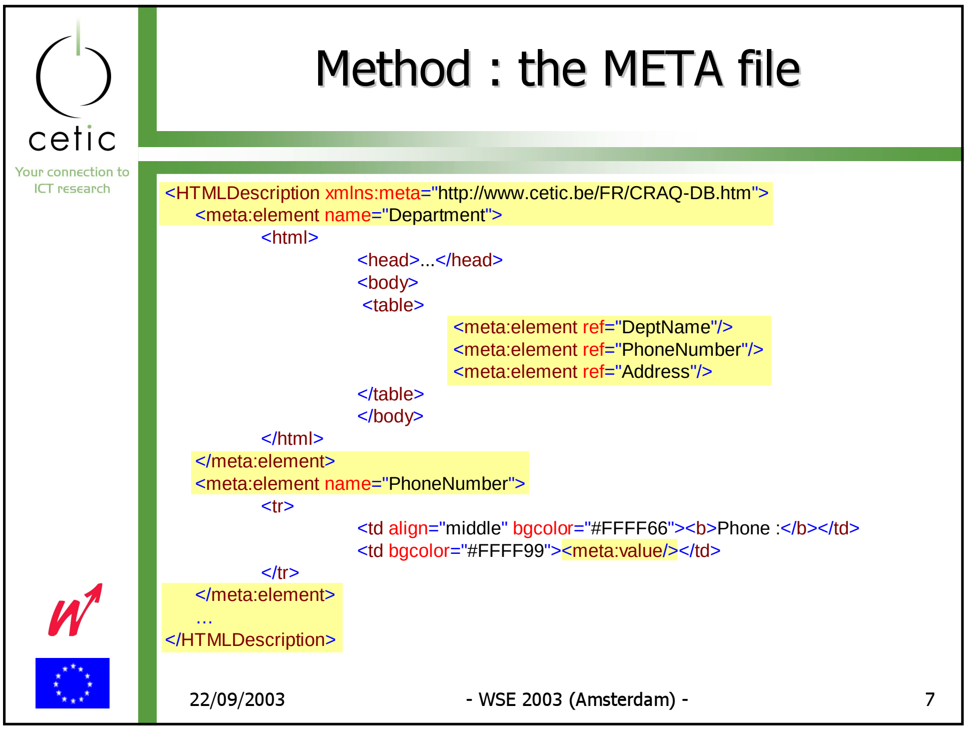
Method : steps 2 and 3.1

- HTML cleaning
- Semantic enrichment
  - For each page type
    - Concepts identification and description on a sample page
      - « Concept » = a part of the HTML tree describing the layout, the structure and possibly the value of a certain reality
      - Ex : the concept « Phone Number »

```
<tr>
   <td align="middle" bgcolor="#FFFF66">
      <b>Phone :</b>
   </td >
   <td bgcolor="#FFFF99">+32 71 72.23.49</td>
</tr>
```

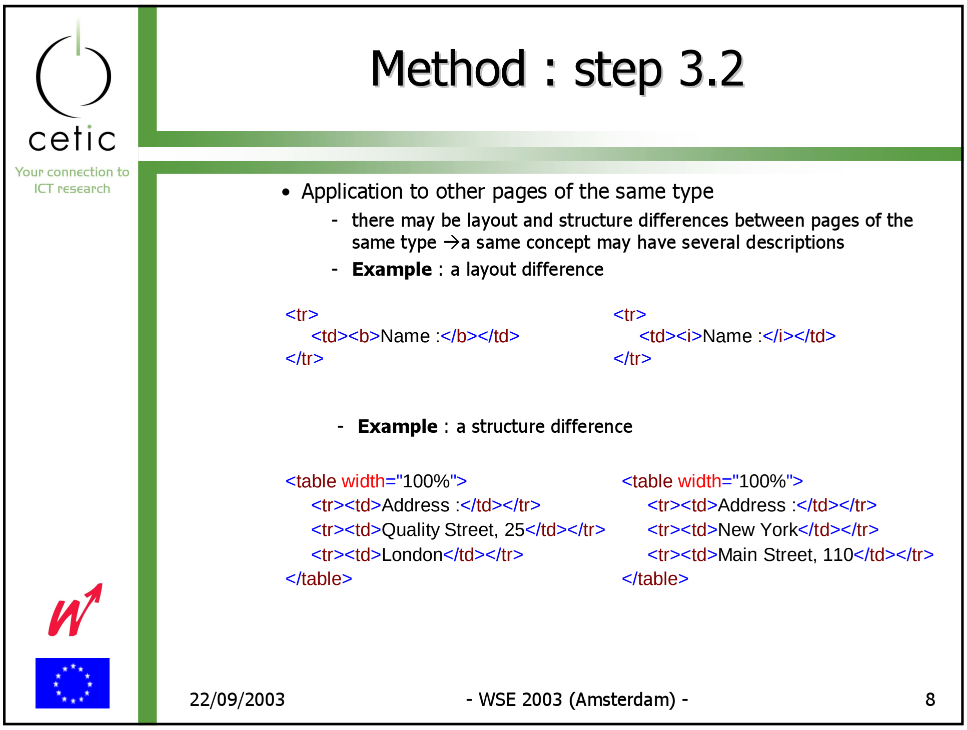
      - Ex : the concept « Address » composed of « Street » and « City »

```
<table width="100%">
   <tr><td><b>Address :<b></td></tr>
   <tr><td>Quality Street, 25</td></tr>
   <tr><td>London</td></tr>
</table>
```



Method : the META file

```
<HTMLDescription xmlns:meta="http://www.cetic.be/FR/CRAQ-DB.htm">
     <meta:element name="Department">
                <html>
                                <head>...</head>
                                <body>
                                 <table>
                                                <meta:element ref="DeptName"/>
                                                <meta:element ref="PhoneNumber"/>
                                                <meta:element ref="Address"/>
                                </table>
                                </body>
                </html>
     </meta:element>
     <meta:element name="PhoneNumber">
                <tr>
                                <td align="middle" bgcolor="#FFFF66"><b>Phone :</b></td>
                                <td bgcolor="#FFFF99"><meta:value/></td>
                </tr>
     </meta:element>
     ...
</HTMLDescription>
```



Method : step 3.2

- Application to other pages of the same type
  - there may be layout and structure differences between pages of the same type →a same concept may have several descriptions
  - **Example** : a layout difference

```
<tr>                                         <tr>
    <td><b>Name :</b></td>                       <td><i>Name :</i></td>
</tr>                                        </tr>
```

  - **Example** : a structure difference

```
<table width="100%">                         <table width="100%">
    <tr><td>Address :</td></tr>                  <tr><td>Address :</td></tr>
    <tr><td>Quality Street, 25</td></tr>         <tr><td>New York</td></tr>
    <tr><td>London</td></tr>                     <tr><td>Main Street, 110</td></tr>
</table>                                     </table>
```

cetic

Your connection to ICT research

Method : step 4

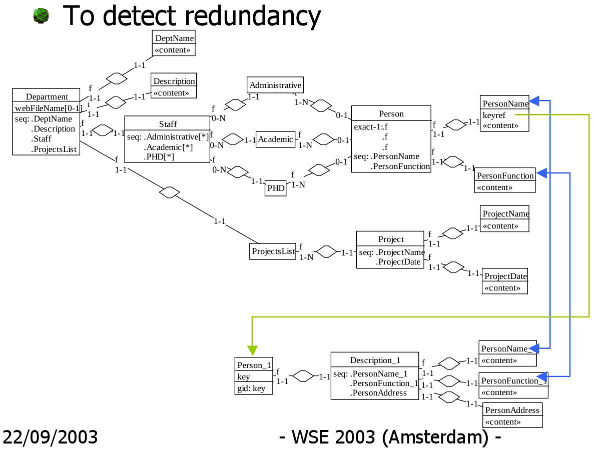
- Data and schema extraction – for each page type
  - Data extraction
    - Web pages + META file → XML document
  - Data structure extraction
    - META file → XML schema

cetic

Your connection to ICT research

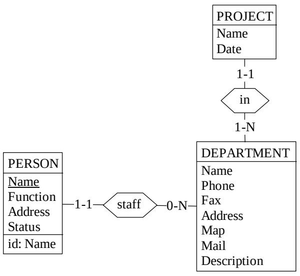
Method : step 5.1

- Schema integration
  - To discover relationships between concepts
  - To detect redundancy



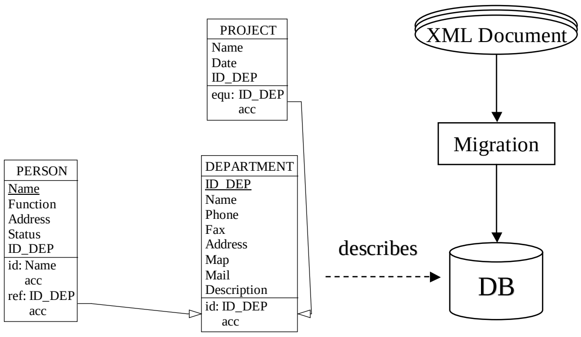
Method : step 5.2

- Schema conceptualisation



Web Site Engineering

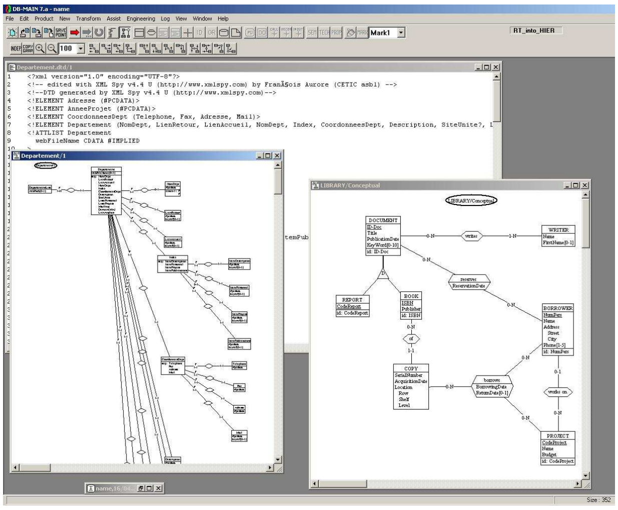
- Database engineering + Data migration



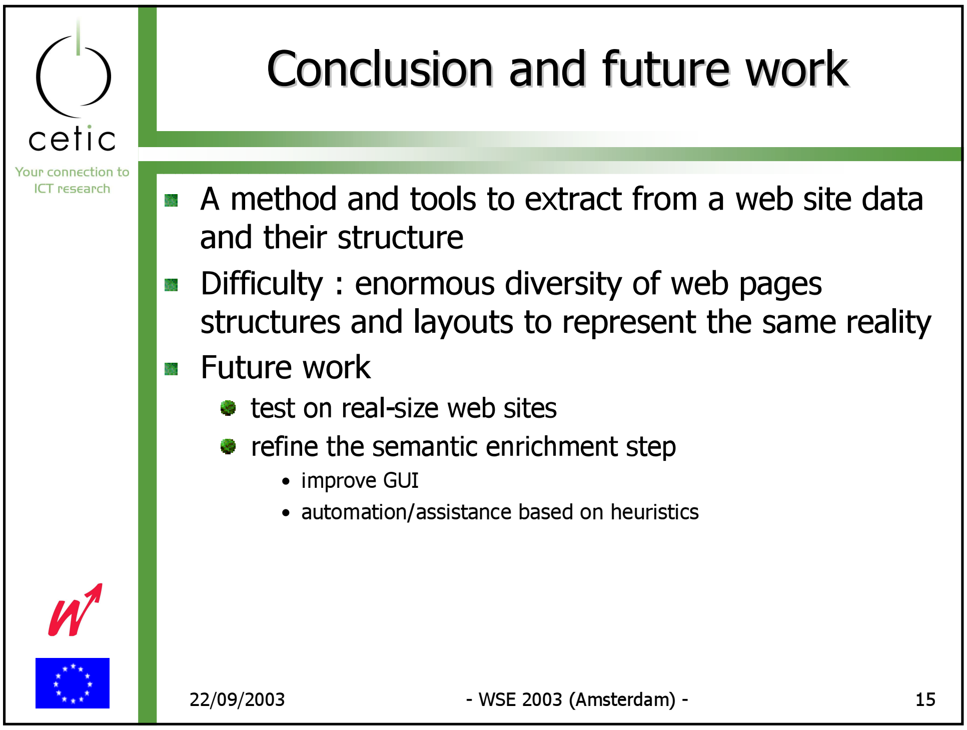
Tools

- HTML cleaning
  - Tidy
- Semantic enrichment
  - XML editor or the semantic browser (based on Mozilla) to edit/generate the META file
- Data and schema extraction
  - XML parsers (Java DOM)
  - *pageType.extractSchema(METAfile)* → *XMLSchema*
  - *pageType.extractData(HTML*, METAfile)* → *XML*
- Schemas integration/conceptualisation and database engineering
  - CASE tool DB-Main

DB-Main (http://www.db-main.be)



```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XML Spy v4.4 U (http://www.xmlspy.com) by FranÃ§ois Aurore (CETIC asbl) -->
<!--DTD generated by XML Spy v4.4 U (http://www.xmlspy.com)-->
<!ELEMENT Adresse (#PCDATA)>
<!ELEMENT AnneeProjet (#PCDATA)>
<!ELEMENT CoordonneesDept (Telephone, Fax, Adresse, Mail)>
<!ELEMENT Departement (NomDept, LienRetour, LienAccueil, NomDept, Index, CoordonneesDept, Description, SiteUnite?, L
<!ATTLIST Departement
  webFileName CDATA #IMPLIED
```



Conclusion and future work

- A method and tools to extract from a web site data and their structure
- Difficulty : enormous diversity of web pages structures and layouts to represent the same reality
- Future work
  - test on real-size web sites
  - refine the semantic enrichment step
    - improve GUI
    - automation/assistance based on heuristics