Laboratory ITODYS, CNRS - UMR 7086
University Paris 7 – Denis Diderot

Using XML for Structuring the Chemical Information:

Towards a Chemical Knowledge Representation

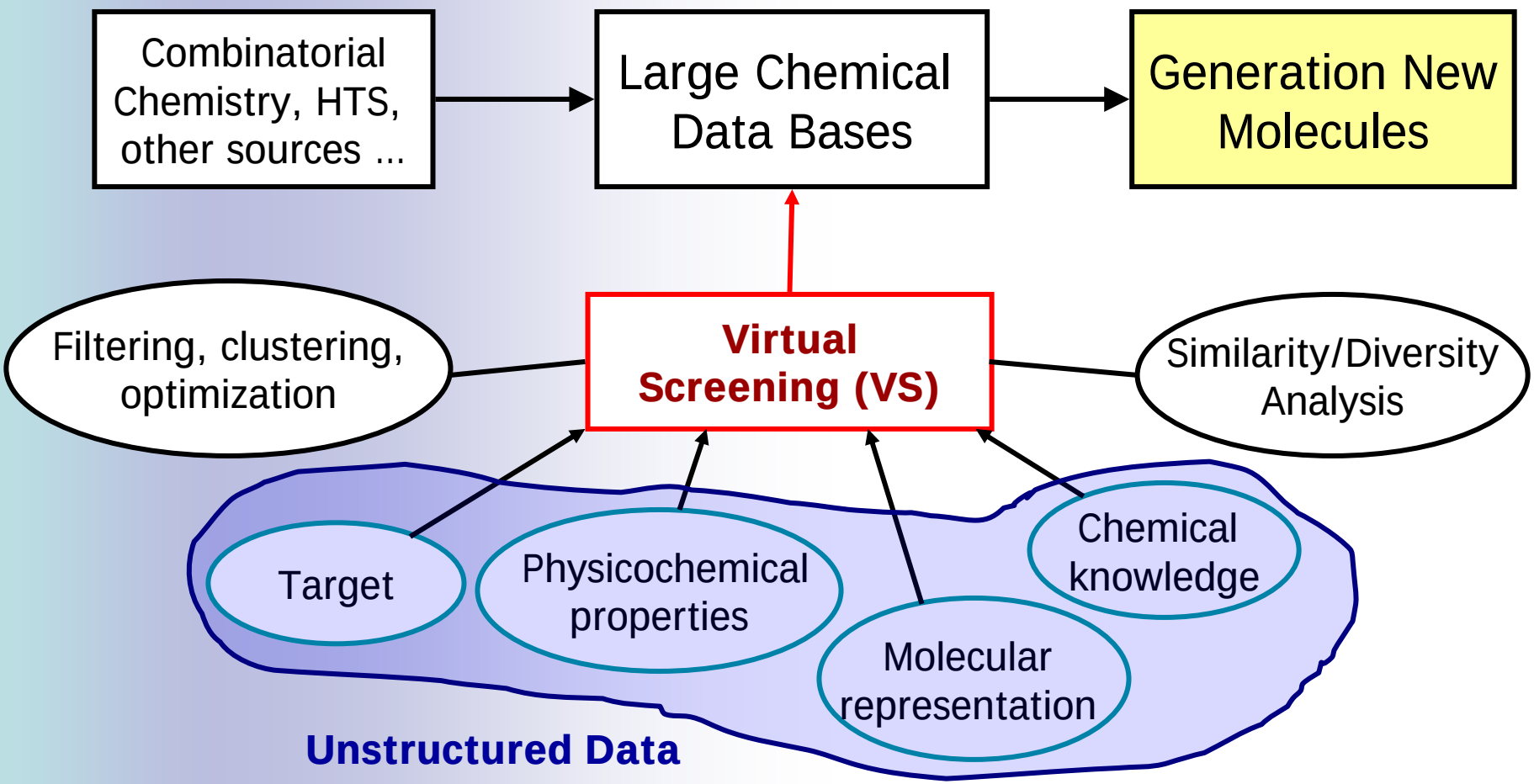
Lecture dedicated to the memory of Prof. J.E. Dubois

Ana Maldonado
FIS-2005 – 4 juillet 2005

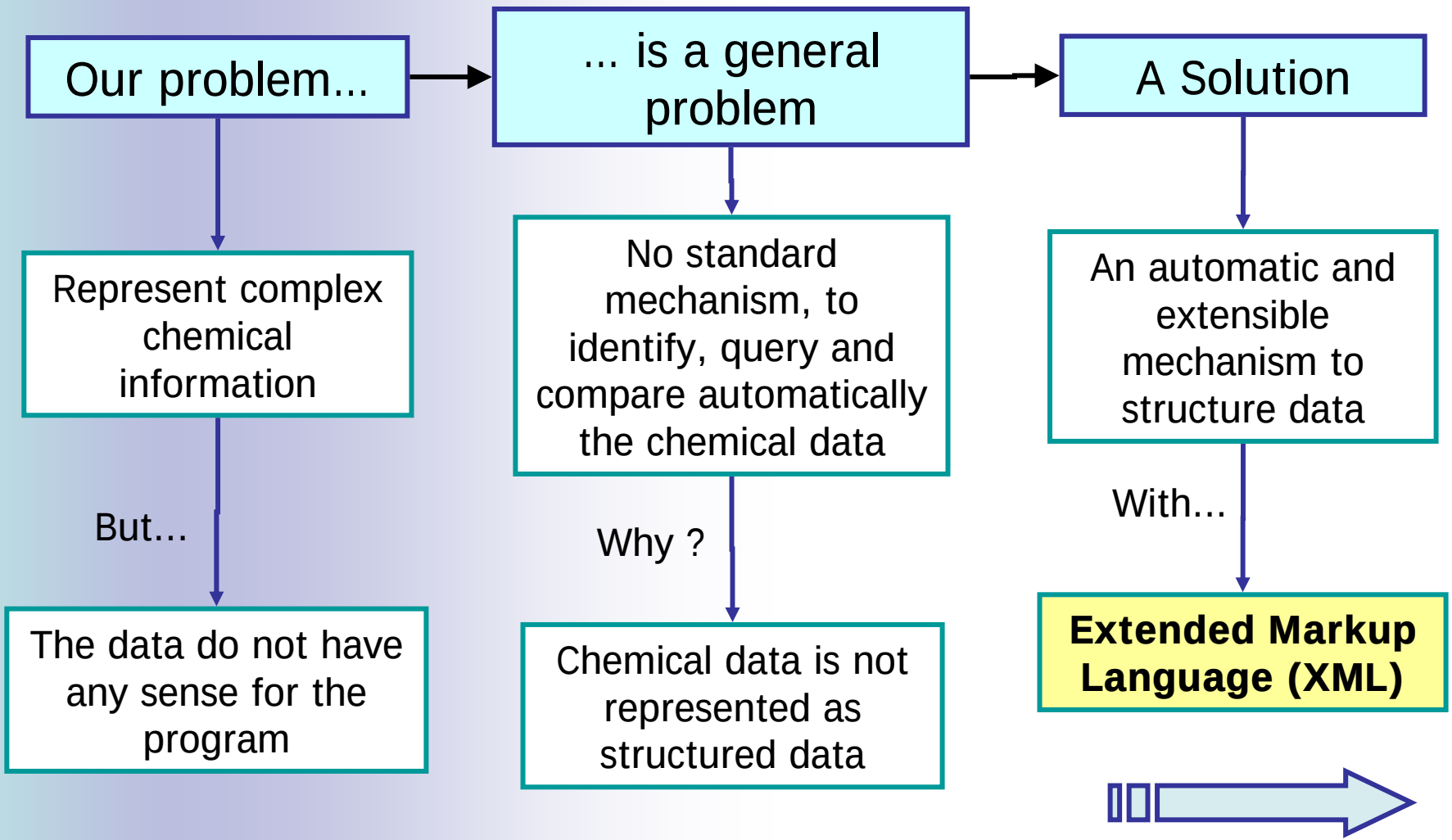
Outline

- Problematic
- XML : History and Principles
- The XML Family
- Structuring Chemical Information
- Towards a Knowledge Representation
- Conclusion
- Perspectives

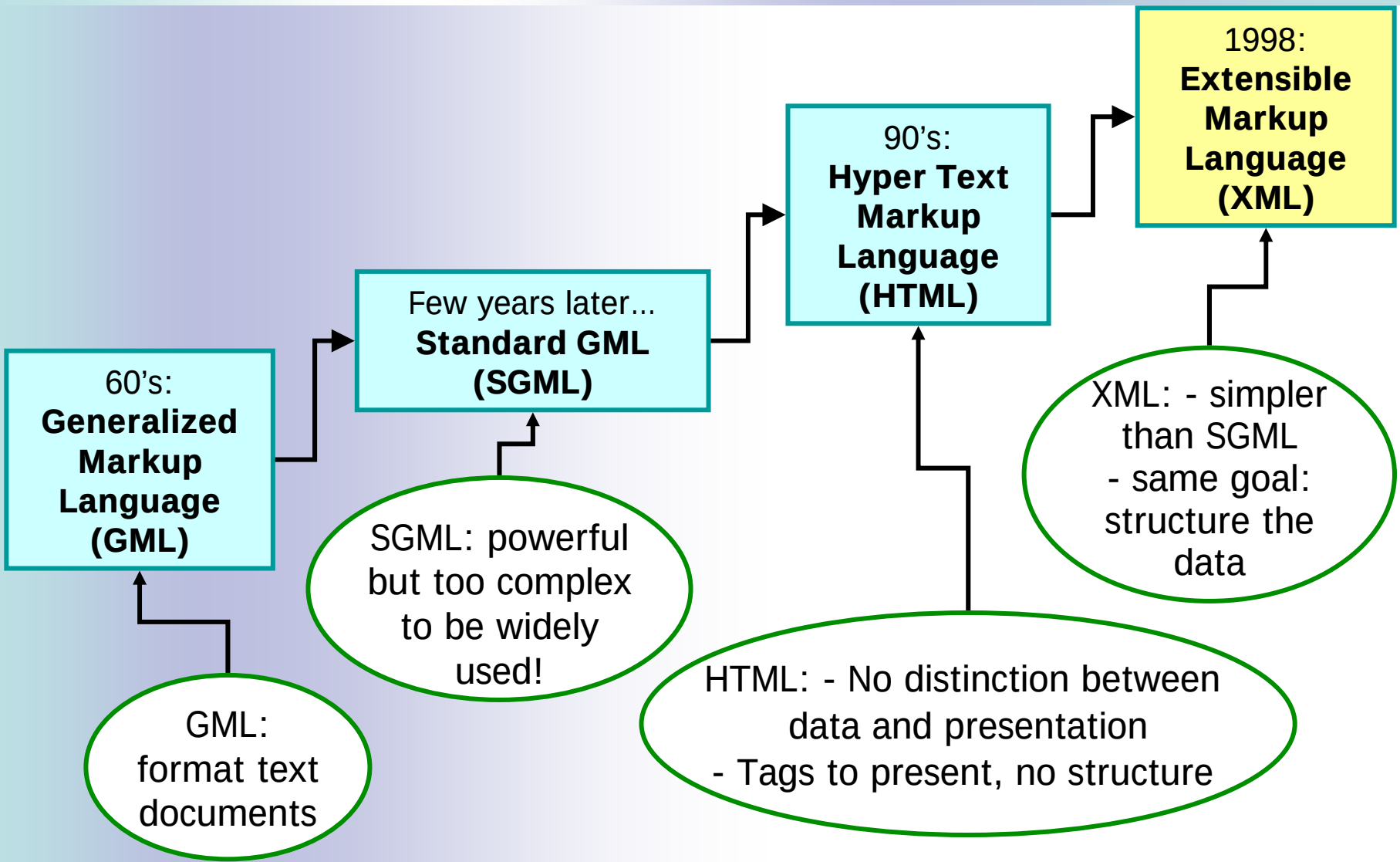
From Chemical Information to New Structures



Finding a Suitable Language



XML History


60's:
Generalized Markup Language (GML)
Few years later...
Standard GML (SGML)
90's:
Hyper Text Markup Language (HTML)
1998:
Extensible Markup Language (XML)
GML: format text documents
SGML: powerful but too complex to be widely used!
HTML: - No distinction between data and presentation
- Tags to present, no structure
XML: - simpler than SGML
- same goal: structure the data

What is XML?

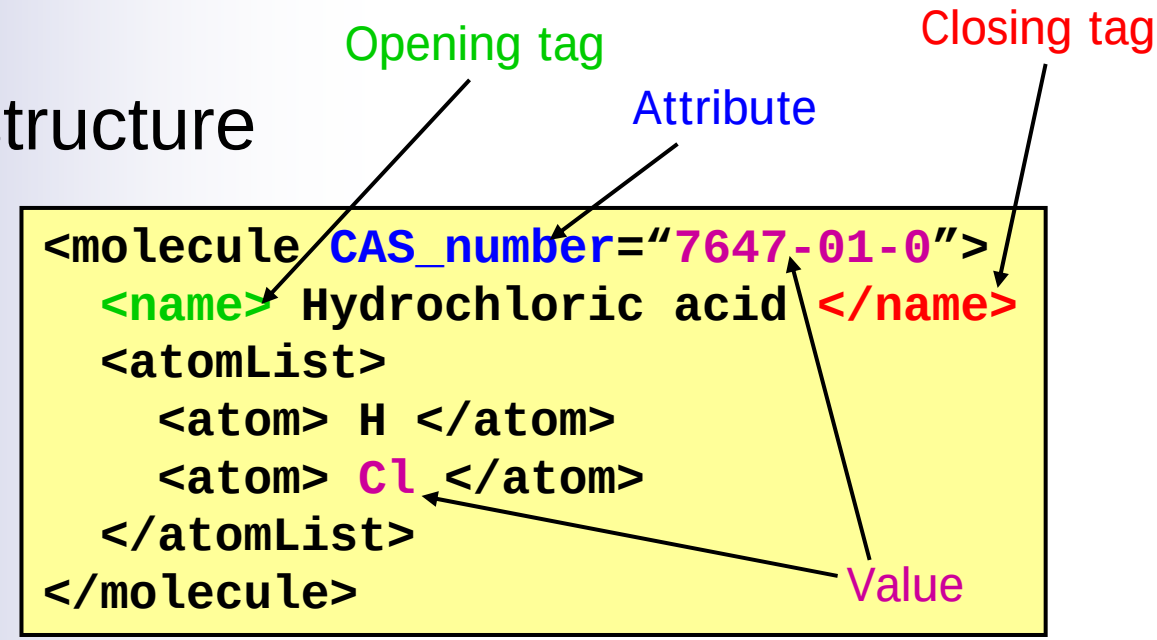
- A set of named tags
- Each opening tag has its matching closing tag in order to form a tree
- A set of attributes / values for each tag
- Rules for controlling the ordering and the nesting of the tags (DTDs)

$\Rightarrow$ **The data become structured**

XML: Structure and Syntax

- A tree model structure

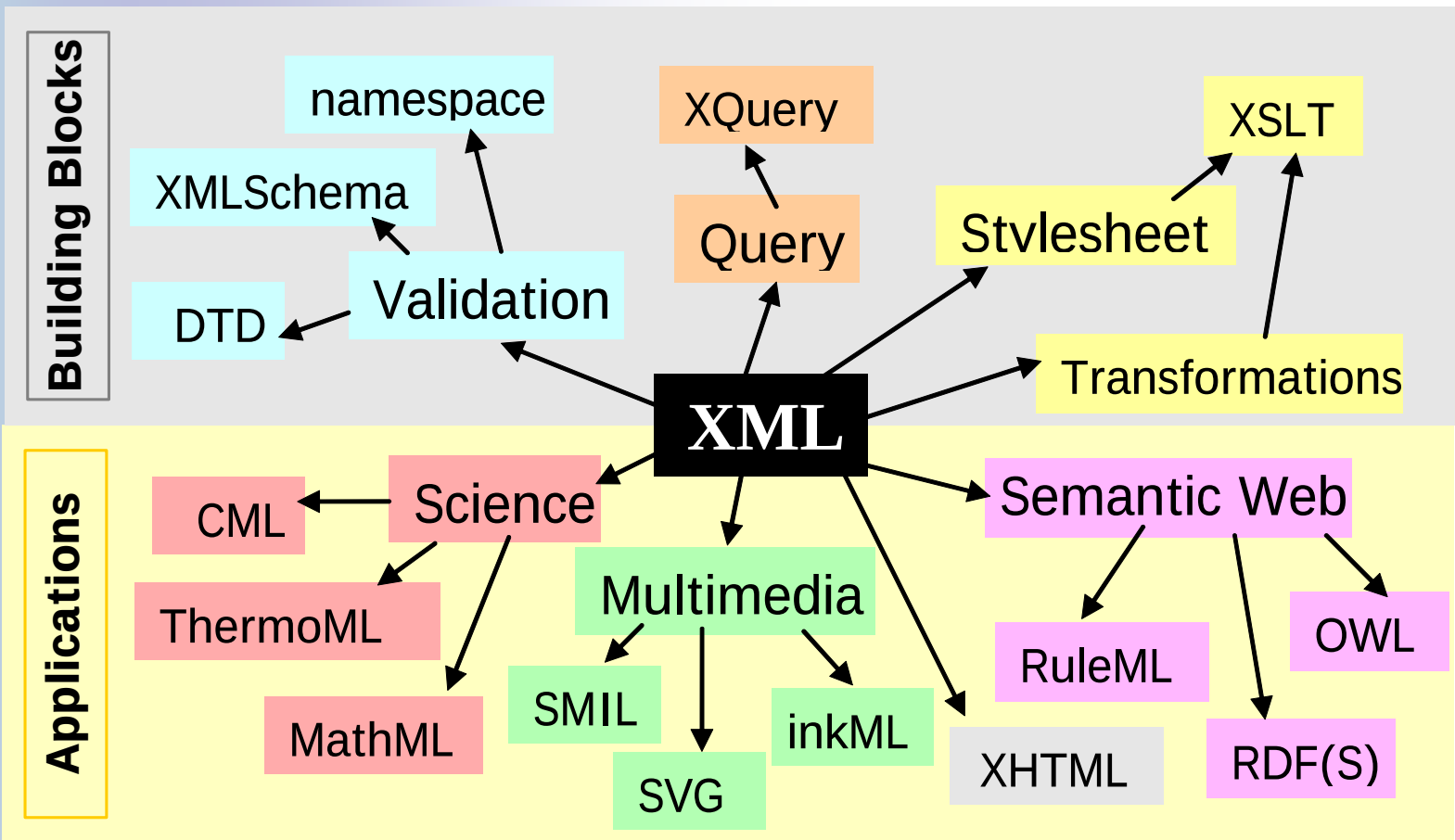
  – A grammar is defined for controlling the data



- The grammar does not define the meaning
  - The tags are linguistics entities that are meaningful for human: names are chosen with agreement

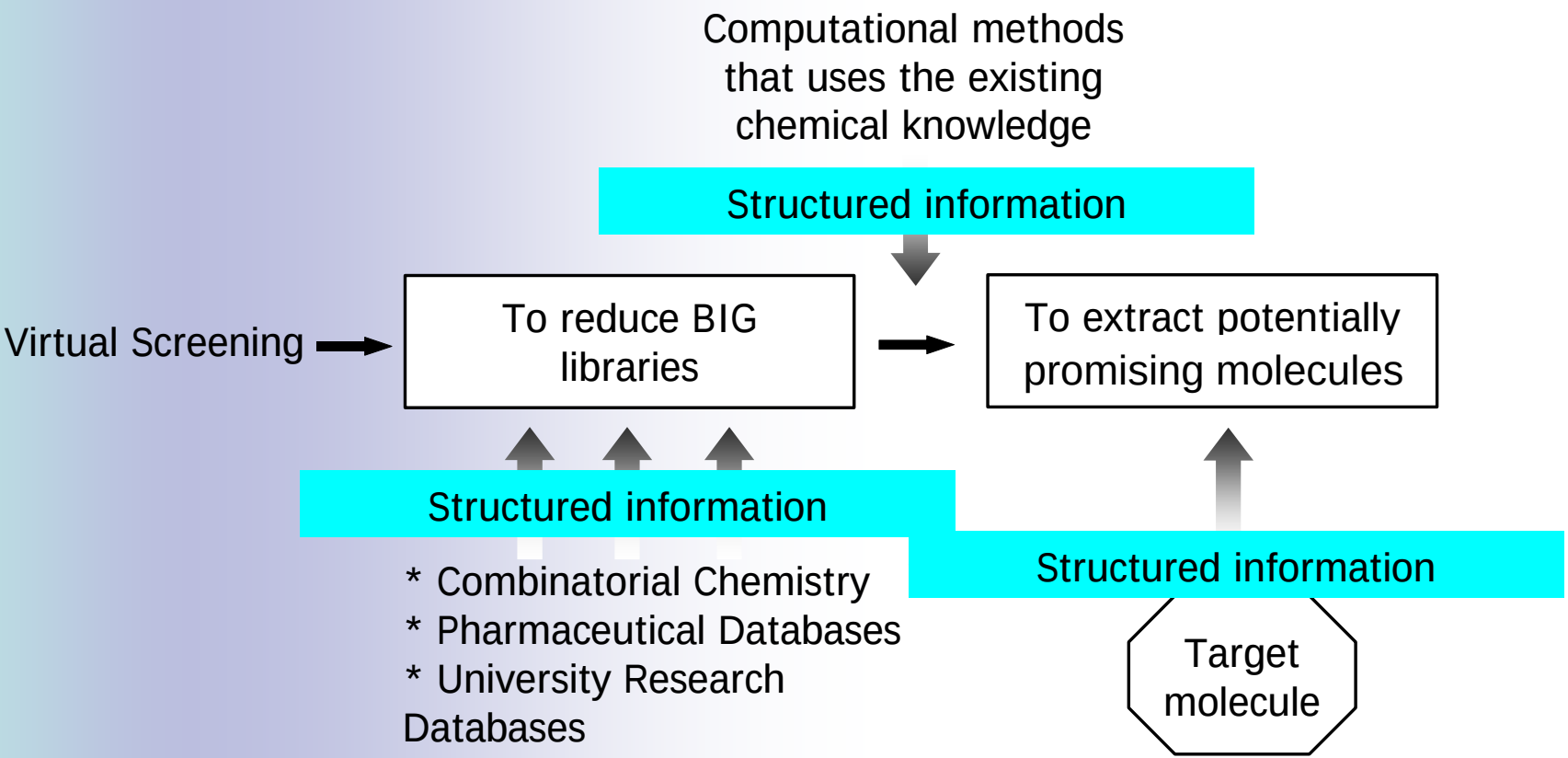
The XML Family

- XML: a meta-language for defining other languages

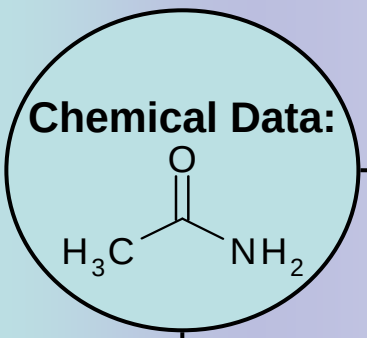


Structuring Chemical Information

- Our context: Virtual Screening (VS)



Structuring chemical information



1)Presented Data

```
  4  3  0  0  0  0  0  0  0  0999 V2000
    2.9167   -0.2459    0.0000 C   0  0  0...
    2.9167   -1.0791    0.0000 O   0  0  0...
    2.2042    0.1630    0.0000 C   0  0  0...
    1.4875   -0.2467    0.0000 N   0  0  0...
  1  2  1  0  0  0  0
  2  3  2  0  0  0  0
  3  4  1  0  0  0  0
M  END
```

Parsing dependent of presentation

Particular Processes

Data Structure **to be defined**

2)Structured Data

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
- <molecule name="Acetamide" id="mol34">
  - <atomArray>
    - <atom id="a1">
        <string builtin="elementType">C</string>
        <float builtin="x2">2.9167</float>
        <float builtin="y2">-0.2459</float>
      </atom>
  ...
    </atomArray>
  - <bondArray>
    - <bond id="b1">
        <string builtin="atomRef">a1</string>
        <string builtin="atomRef">a2</string>
        <string builtin="order">1</string>
      </bond>
  ...
    </bondArray>
</molecule>
```

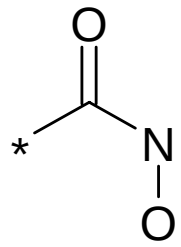
Parsing independent of presentation

Generic Processes

Fixed Data Structure (e.g. libXML)

XML Implementation

```
<Index>
    <File name="AGCM-29c.mol">
        <Keys>
            <Key name="FID" value="29c"/>
            <Key name="FAtomSum" value="5"/>
            <Key name="FRing" value="0"/>
            <Key name="FGF" value="N-hydroxy-formamide"/>
        </Keys>
        <Properties>
            <Property name = "HBondAD" value = "1"/>
            <Property name = "PotPCharged" value = "0"/>
            <Property name = "PotNCharged" value = "1"/>
            <Property name = "HydPhi" value = "1"/>
            <Property name = "Aromat" value = "0"/>
            <Property name = "Polar" value = "1"/>
            <Property name = "HydPho" value = "0"/>
        </Properties>
    </File>
....
</Index>
```



AGCM-29c.mol

```
-ISIS-  05200314222D

  5  4  0  0  0  0  0  0  0  0999 V2000
   -1.6083    1.0958    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -1.6083    0.2667    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -2.3208    1.5047    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.3208    2.3366    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -3.0333    1.0958    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  3  4  2  0  0  0  0
  3  5  1  0  0  0  0
  3  1  1  0  0  0  0
M  END
```

XML Implementation

- Need for a DTD (Document Type Definition)

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<!-- Data Structure for mol files   -->
<index>
  <molecule name="water-1a.mol">
    <formula>
      <ConnecTable>table-1a.skc</ConnecTable>
      <AtomSum>3</AtomSum>
    </formula>
    <property>
      <HBondDonnor>0</HBondDonnor>
      <PotPCharged>1</PotPCharged>
    </property>
    <key>
      <ID>1a</ID>
      <Aromatic>0</Aromatic>
    </key>
  </molecule>
</index>
```

```
<!-- DTD for data structure -->
<!ELEMENT index (molecule+)>
<!ELEMENT molecule (formula,property,key)>
<!ATTLIST molecule name CDATA #REQUIRED>
<!ELEMENT formula (ConnecTable,AtomSum)>
<!ELEMENT ConnecTable (#PCDATA)>
<!ELEMENT AtomSum (#PCDATA)>
<!ELEMENT Property (HBondDonnor,PotPCharged)>
<!ELEMENT HBondDonnor(#PCDATA)>
<!ELEMENT PotPCharged (#PCDATA)>
<!ELEMENT key (ID,Aromatic)>
<!ELEMENT ID (#PCDATA)>
<!ELEMENT Aromatic (#PCDATA)>
```

XML as a Partial Solution

(+)	• Definition of languages adapted for each scientific domain • Structure of the information open and flexible • Queries/screening of DB simplified and automatic • “Intelligent” searches are possible • Numerous tools can process it (C family, JAVA, Perl, etc.)
(–)	• Verbose language • Tags have no meaning for machines

Towards a Knowledge Formalization

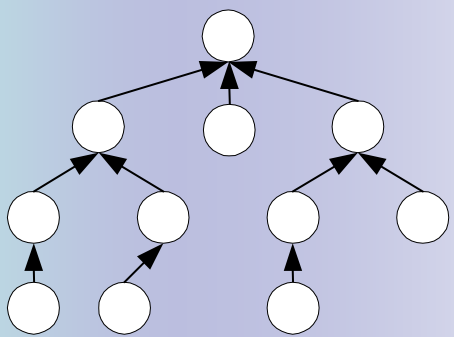
- Why?

- For making the meaning of the data machine understandable

- How?

- Building ontologies (kind of thesaurus)

- Encoded in logical languages (XML syntax or not)



Ontology

Logical Languages

$\forall x$ Molecule(x) $\Rightarrow$
Cyclic(x) $\sqcup$ Aliphatic(x)

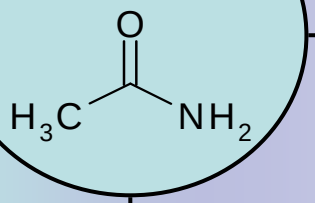
$\forall x$ Bond(x) $\Rightarrow$ Single(x) $\sqcup$
Double(x) $\sqcup$ Triple(x)

...

Knowledge Representation

Treating Chemical Data and Knowledge

Chemical Data:



2)Structured Data

```
<?xml version="1.0" encoding="iso-8859-1"
standalone="no" ?>
- <molecule name=" Acetamide ">
  - <atomArray>
    - <atom id="a1">
        <string builtin="elementType">C</string>
        <float builtin="x2">2.9167</float>
        <float builtin="y2">-0.2459</float>
      </atom>
  ...
    </atomArray>
  </molecule>
```

Generic Processes

Parsing independent of presentation

Fixed Data Structure

3)Structured Data & Knowledge

```
<?xml version="1.0" encoding="iso-8859-1"
standalone="no" ?>
- <molecule name=" Acetamide ">
  - <atomArray>
    - <atom id="a1">
        <string builtin="elementType">C</string>
        <float builtin="x2">2.9167</float>
        <float builtin="y2">-0.2459</float>
      </atom>
  ...
    </atomArray>
  </molecule>
```

Ontology → Logical Languages → Knowledge Representation

Parsing independent of presentation

Generic Processes

Possibly automatized

Fixed Data Structure (ontology+XML)

Conclusion

- Representation of the structured chemical information (VS software):
  - Conventional file storage (mol files)
  - Markup languages (XML files)
- XML is a convenient way to retrieve all kinds of data
- Limitations of XML
  - Verbose language
  - Needs for a formal chemical ontology

Perspectives

- Evolve to a 100% XML/CML chemical structured data system
  - Interesting for displaying and managing the molecular information
  - Keeping a correct granularity
  - The use of CML standard tags will allow the exchange of our data with other systems
- Make use of the semantics of the chemical data
  - Assume that a chemical ontology is defined by the community

Thanks for your attention!

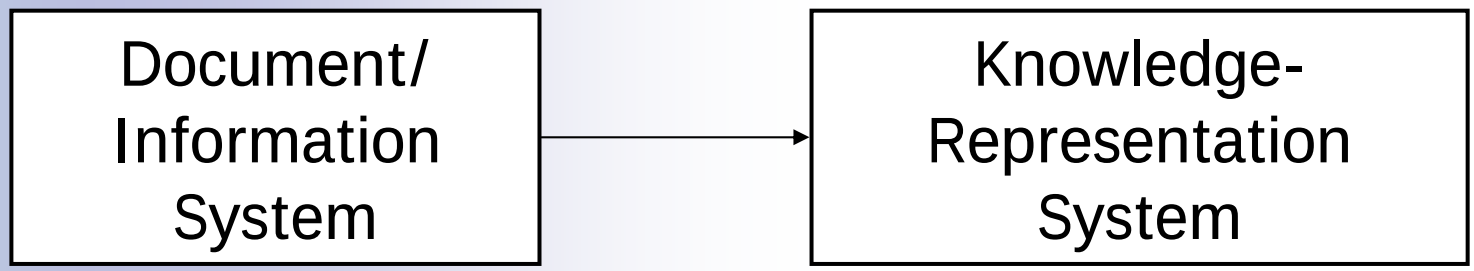
Laboratoire ITODYS, CNRS - UMR 7086
Université Paris 7 – Denis Diderot

Bibliography

- Extended Markup Language (XML) 1.0, W3C Recommendation, 04 February 2004. http://www.w3.org/TR/REC-xml
- Chemical Markup Language (CML). http://www.xml-cml.org
- The World Wide Web Consortium (W3C). http://www.w3c.org
- Document Type Definitions. http://www.xmlfiles.com/dtd/

Ontologies

- Allows the transformation:



- XML can be used at different levels
  - as a *model* for structuring the data
  - as a *possible syntax* to represent ontologies (but there is others, like logical syntax)

$\Rightarrow$ **usually source of confusion**