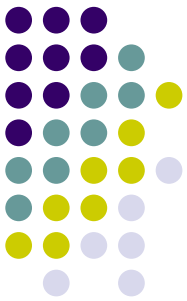
An XML Format For Proteomics Database

to Accelerate Collaboration Among Researchers



HUP-ML:

Human Proteome Markup Language

K. Kamijo, H. Mizuguchi, A. Kenmochi,
M. Sato, Y. Takaki and A. Tsugita
Proteomics Res. Center, NEC Corp.

Overview (HUP-ML)

- Introduction
- XML features
- XMLs in life science
- HUP-ML concept
- Example of HUP-ML
- Details of HUP-ML

- Demonstration of “HUP-ML Editor”

Introduction

- Download entries from public DBs as a flat-file
  - easy for a person to read
  - different formats for every DB
  - sometimes needs special access methods and special applications for each format

- Needs machine-readable formats for software tools

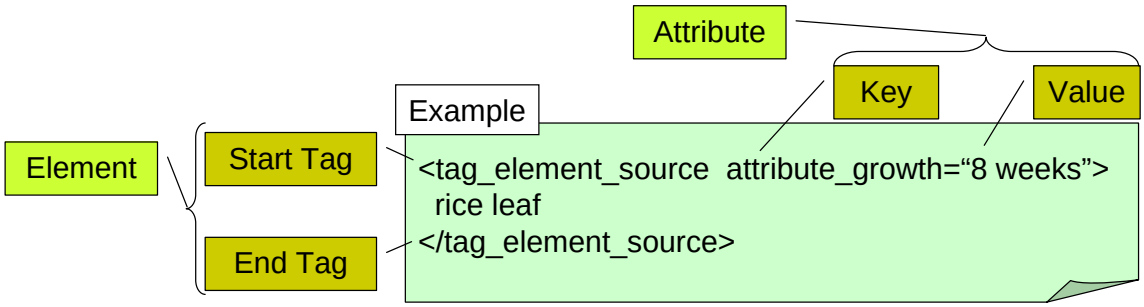
- To boost studies by exchanging data among researchers

Activates standardization

XML format

- XML (eXtensible Markup Language)

  W3C: World Wide Web Consortium (inception in 1996)



  - Highly readable for machine and person
  - Can represent information hierarchy and relationships
  - Details can be added right away
- Convenient for exchanging data
  - Easy to translate to other formats
  - Logical-check by a Document Type Definition (DTD)

Overview (HUP-ML)

- Introduction
- XML features
- XMLs in life science
- HUP-ML concept
- Example of HUP-ML
- Details of HUP-ML

- Demonstration of “HUP-ML Editor”

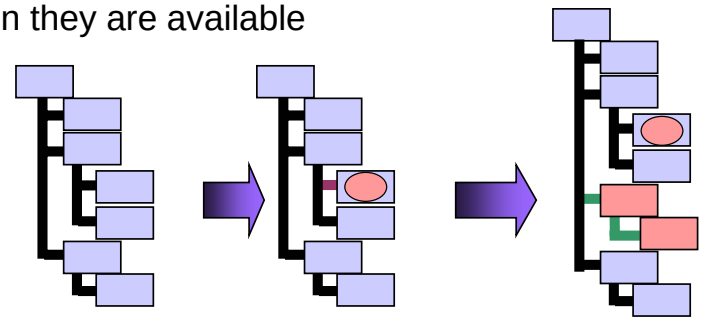
XML features

- Exchangeable
  - Inter-operability:
    - OS (Operating System): Windows, Linux, etc.
    - Software Development Languages
    - Communication Protocols
  - Inter-national / Inter-business
    - W3C has supervised the XML specifications
    - XML is well prevailed internationally
    - XML supports multilingual character code sets

W3C: World Wide Web Consortium (inception in 1996)

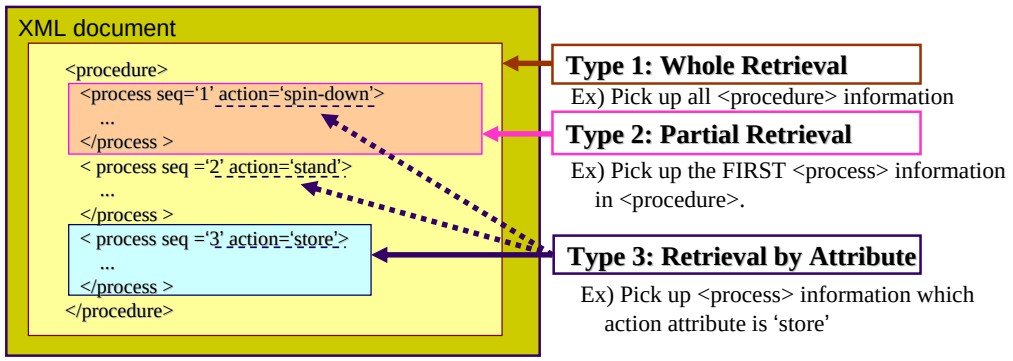
XML features

- Extensible
  - W3C supervised ONLY the language specifications
    - Everyone could create a NEW language structure based on the specifications
  - Hierarchy structure
    - A change in a node does not affect data in other nodes
  - Easy to add new information
    - when they are available



XML features

- Retrievable with a computer
  - Easy to find target information in an XML document
    - By using 'Tag' element structure
    - It is difficult to retrieve them in a flat text file and a binary file.
  - Possible to find more specific information with 'Attribute' keywords step by step



XML features

- Easy to develop XML applications
  - A variety of XML tools are available for DEVELOPERs
    - Powerful XML parser / processor
    - A developer does not need to create an XML format parser
    - Little tools for original formats
  - Possible to validate a well-formatted document
    - By using tools and schema files
    - DTD (Document Type Definition) file
    - XML Schema file, Relax NG
  - A Developer could concentrate ONLY a function of applications (NOT XML document handling)

XML features

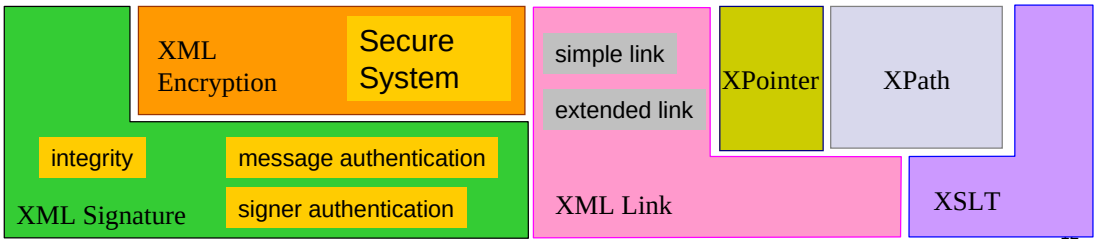
- A variety of XML tools are available for USERS
  - Standardization Tools
    - Retrieval Protocols: XPath
    - Data Transformation: XSLT (Extensible Stylesheet Language Transformations)
  - Easy to handle XML document by using such tools

XML features

- A variety of related tools
  - XML database management system
  - XML formatter, XML Editor
  - Communication Protocols (ex. SOAP)
- Major applications and tools tend to support XML document handling
  - RDB: ORACLE, SQL Server ....
  - Web browser: Microsoft Internet Explorer ...
  - Applications: Microsoft Office ...

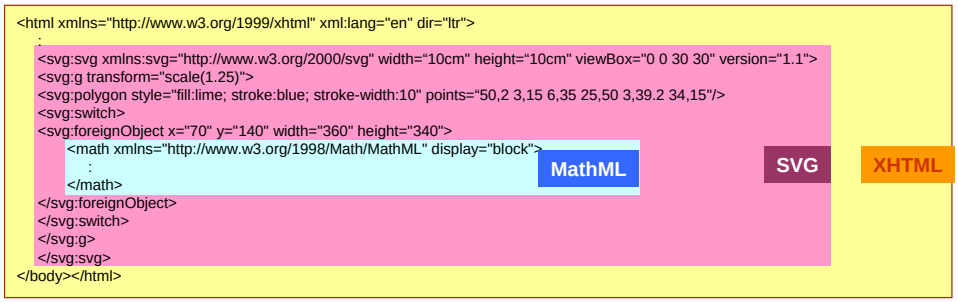
XML features

- Easy to combine XML applications with other functions
  - Data processing tools
    - Embedding link information (XLink, XPointer)
    - Electric authentication, Encryption
  - Some tools automatically bring their functions to applications



XML features

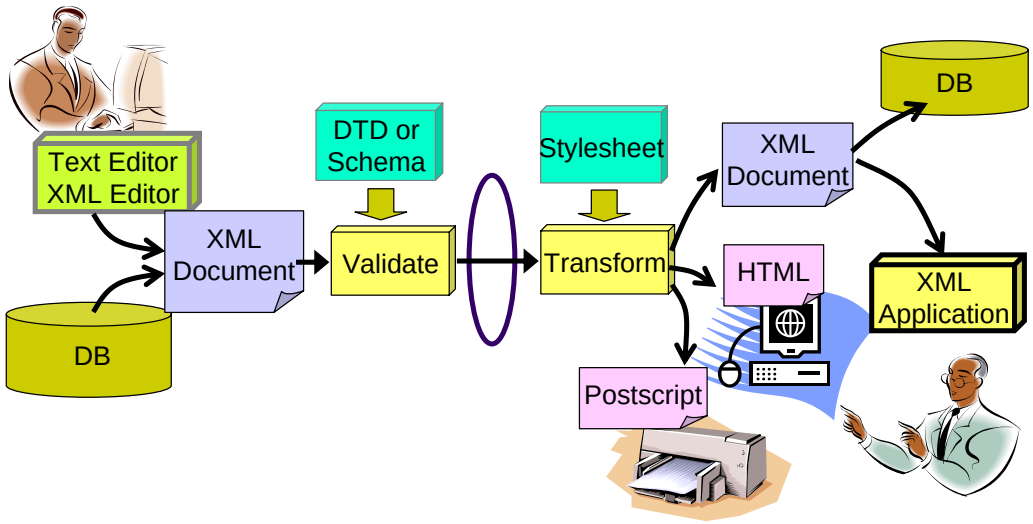
- Easy to combine with other XML format documents
  - By using ‘NameSpaces’ technique, it is possible to combine them, integrate them and divide into original XML formats
  - Example: Combination of HTML and the following XML
    - Figures (*SVG*: Scalable Vector Graphics,Web3D etc.)
    - Formula (*MathML*: Mathematical Markup Language)
    - Chemical formula (*CML*: Chemical Markup Language)



XML features

- XML is a ‘reliable’ format to use in business
  - XML gets a major position of data description languages
  - Many software venders express XML support
  - Activity of standardization organization (W3C)
  - XML will be used in long time because XML is not any vender's format

XML workflow



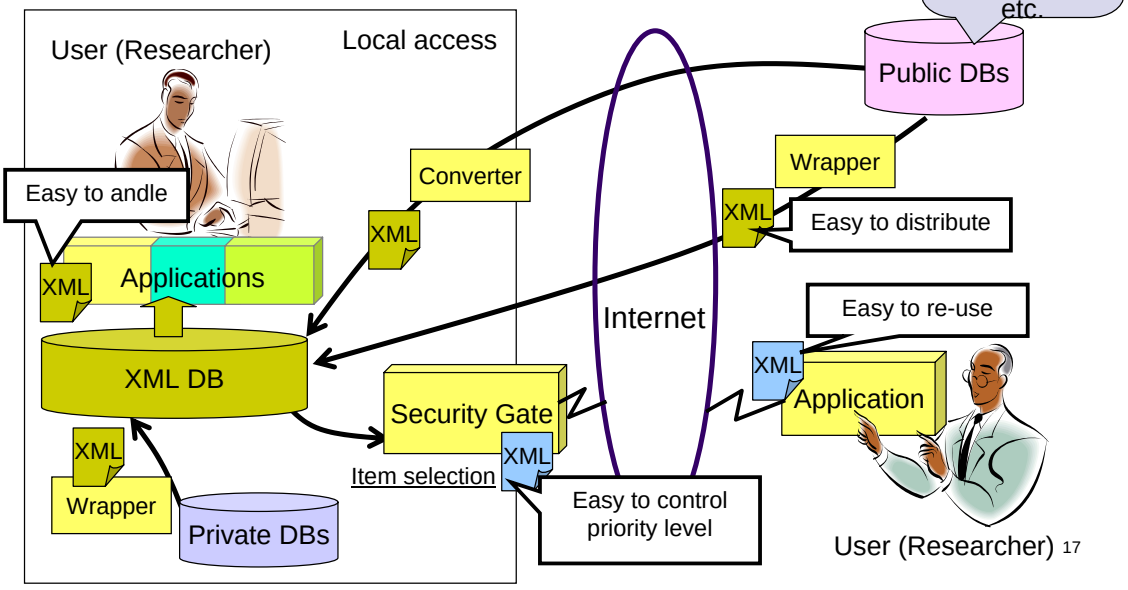
Overview (HUP-ML)

- Introduction
- XML features
- XMLs in life science
- HUP-ML concept
- Example of HUP-ML
- Details of HUP-ML

- Demonstration of “HUP-ML Editor”

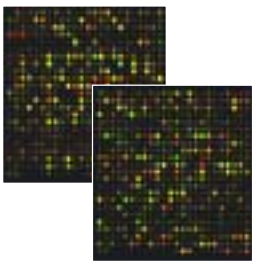
XML in Bioinformatics

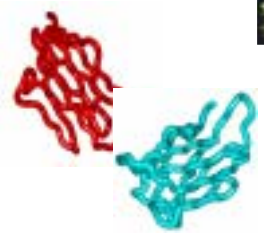
"The Extensible Markup Language (XML) is the universal format for structured documents and data on the Web." -- W3C XML Web site, 2000-07-06.



Conventional XMLs in Life Science

- Genomics
  - BSML, GAME, DAS, DDBJ-XML,…
- Gene Expression
  - MAGE-ML(GEML, MAML), GeneXML, …
- Proteomics
  - PSDML (PIR's XML)
  - BioML: Functional Proteomics
  - ProML: Structural Proteomics
  - (BSML)

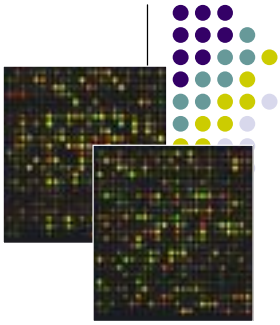




Conventional XMLs in Life Science

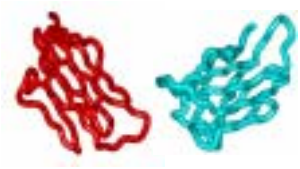
- Genomics
  - BSML
    - Bioinformatic Sequence Markup Language
    - For exchanging genomic sequence and annotation
    - By LabBook developers under a 1997 grant from National Human Genome Research Institute in the US
    - One Main XML format used by the Interoperable Informatics Infrastructure Consortium (I3C)
  - GAME
    - Genome Annotation Markup Language
    - To represent information about specific regions of sequence
    - Supporter: EBI XEMBL Project

Conventional XMLs in Life Science



- Gene Expression
  - MAGE-ML
    - Micro Array and Gene Expression Markup Language
    - To describe and communicate information on microarray-based experiments
    - Discuss in OMG LSR Meeting in 2001
    - Supporter: EMBL-EBI, Rosetta Inpharmatics, etc.
    - Contains microarray designs, manufacturing info., experiment setup and execution information, as well as gene expression data and data analysis results
    - Purpose: To provide a framework where the researchers can exchange microarray data and compare the analysis results even if the maker of the array chip is different

Conventional XMLs in Life Science



- Proteomics
  - PSDML
    - Protein Sequence Database Markup Language
    - An open-standard markup language used to store protein information in the PIR database
  - BioML
    - Biopolymer Markup Language
    - Developed by ProteoMetrics company
    - designed to be used for the general annotation of biopolymer sequence information
    - specifying all experiment information on molecular entities
      - proteins, genes and other biopolymers

Advantage of Using A Markup Language For Life Science

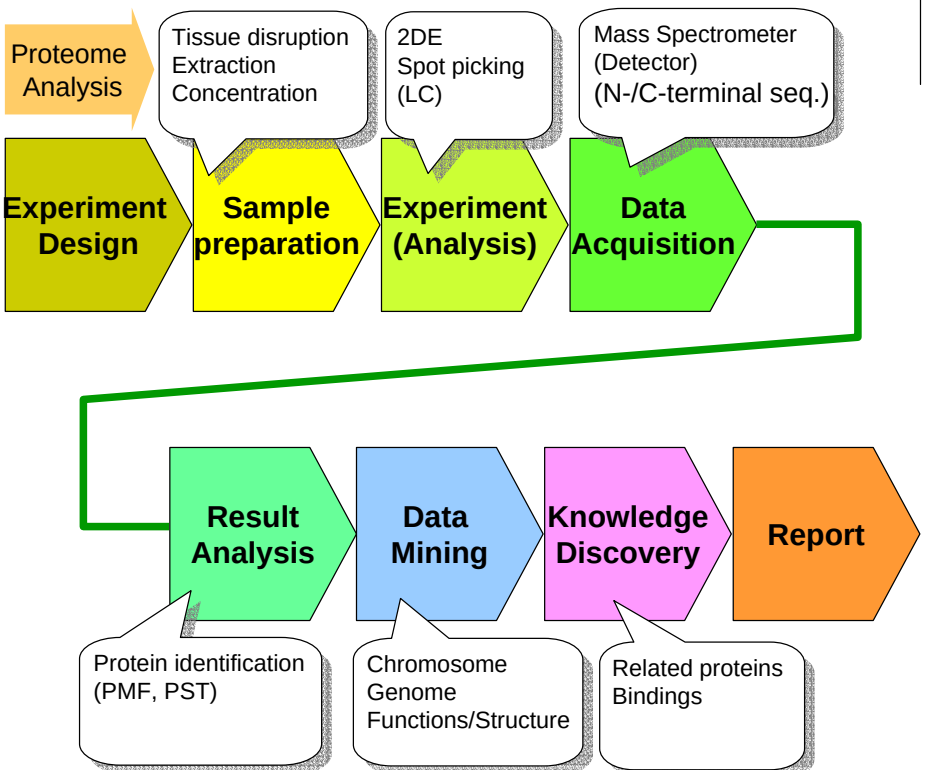
- An XML-based tree structure
  - Easily maps into the biopolymer information
    - hierarchical and nested at different levels of complexity
  - Bioinformatics data have numerous relationships
- Extensible (Flexibility)
  - New data emerge regularly
  - Data are updated frequently

Overview (HUP-ML)

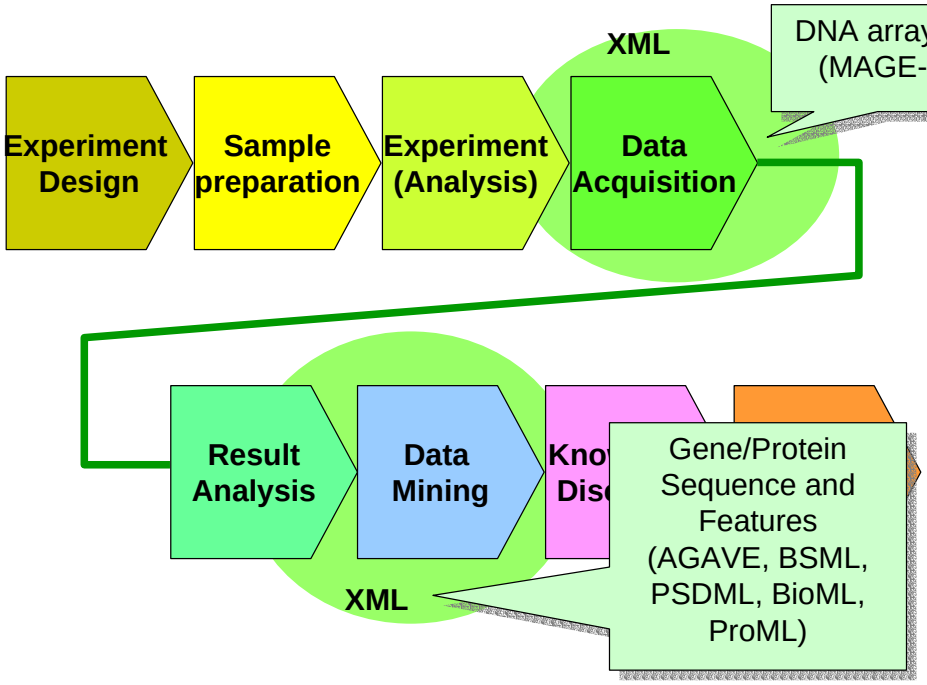
- Introduction
- XML features
- XMLs in life science
- HUP-ML concept
- Example of HUP-ML
- Details of HUP-ML

- Demonstration of “HUP-ML Editor”

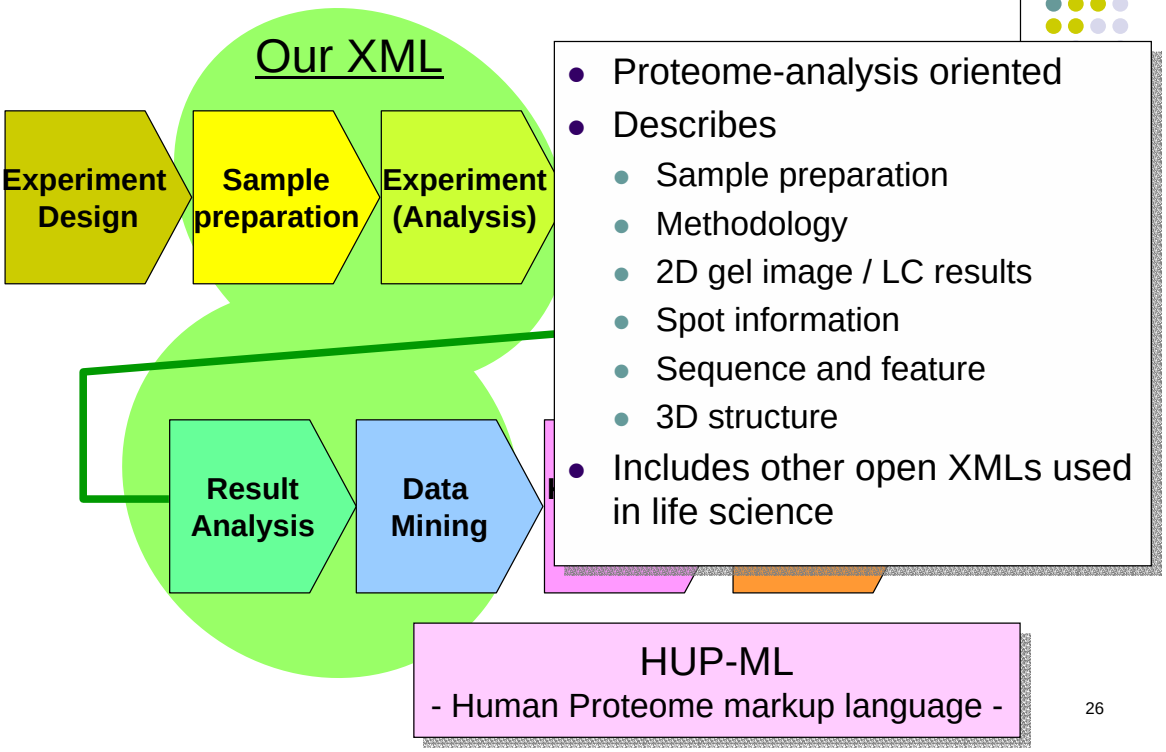
Analysis flow in Life Science



Conventional XMLs in Life Science



Our XML-based data model



HUP-ML

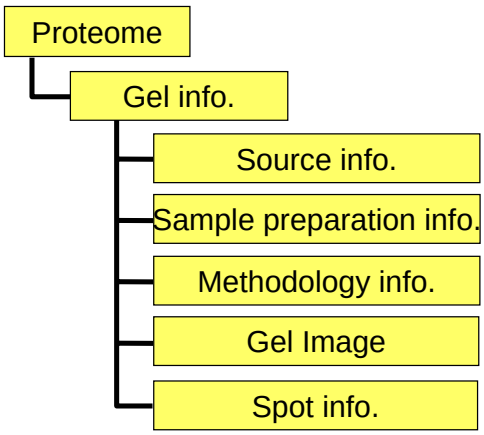
- Current version (version 0.43, beta)
  - source information
  - sample preparation information
  - Target method: 2DE
    - Gel information
    - Spot information etc.

Simpson,R.J.,Tsugita,A.,Celis,J.E,Garrels,J.I.& Mewes,H.W., "Workshop on two-dimensional gel protein database," Electrophoresis 13, 1055-1061(1992)

- Discussion points for future version (tomorrow)
  - MS information description
  - Other method description (ex. LC)
  - etc.

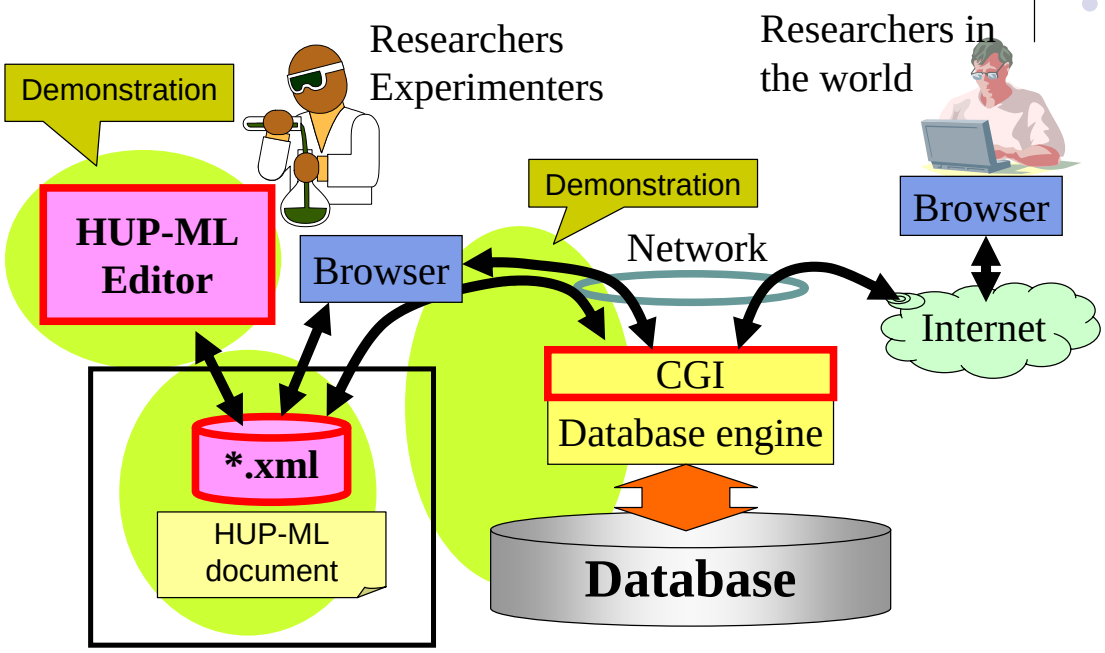
HUP-ML (current version, 0.43)

- Information Structure:




```
<proteome>
  <gel   id="1">
    <source_info>
    <gel_img >
    <sample_preparation>
    <gel_conditions>
    <marker>
    <detection>
    <gel_image>
    <spot   id="1">
      ....
    <spot   id="2">
    ....
  <gel   id="2">
```

HUP-ML document exchange scheme



Overview (HUP-ML)

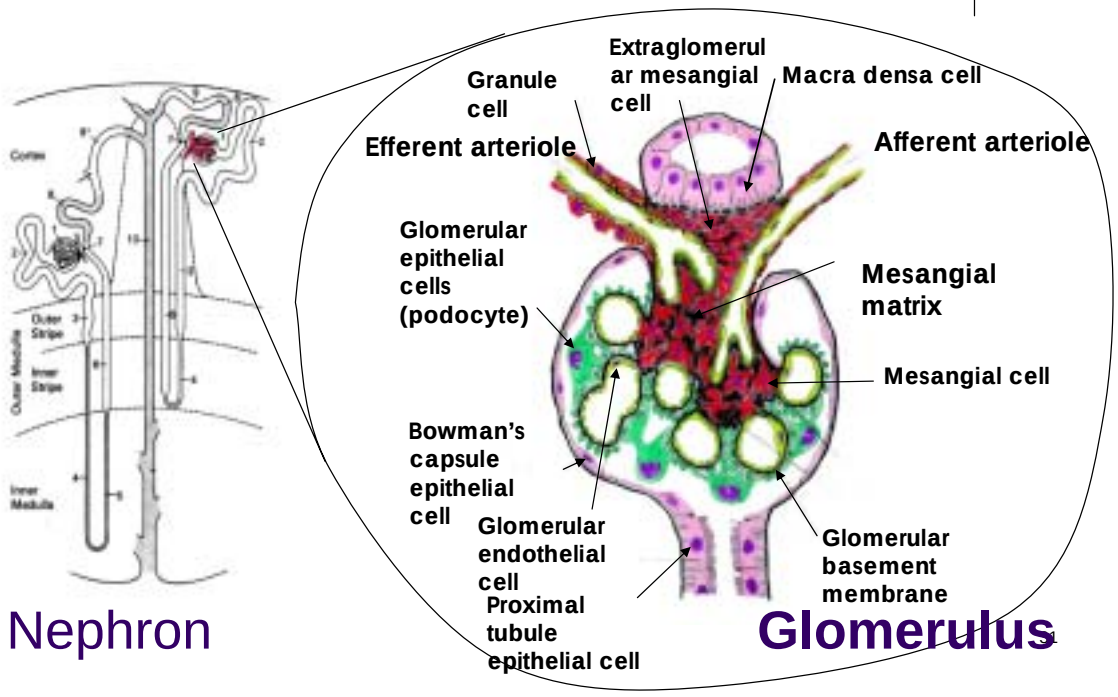
- Introduction
- XML features
- XMLs in life science
- HUP-ML concept
- Example of HUP-ML
- Details of HUP-ML

- Demonstration of “HUP-ML Editor”

Example:

By A. Tsugita et al.(2002)

Human Kidney Glomerulus Proteome



Sample of HUP-ML (1)

Source information

```
- <source_info source_info_ID="HKG-1"
    creDate="2002-07-20T12:00:00"
    modDate="2002-08-10T17:20:00">
  <source>Homo sapiens</source>
  <common_name>Human</common_name>
  <strain />
  <cultiva />
  <cell_line />
  <tissue>Kidney Glomerulus</tissue>
  <plasmid />
  <growth_phase unit="year">48</growth_phase>
  <induction />
  <host />
  <description>Normal</description>
</source_info>
```

Sample of HUP-ML (2)

Sample preparation

```
- <sample_preparation>
  <tissue-disruption>Standard sieving technique using four stainless sieves. The glomeruli on the 150 micro m sieves were collected ice cold phosphate-buffered saline (PBS).</tissue-disruption>
- <extraction>
  - <procedure>
      <process seq="1" action="spin-down"
                sample="collection" />
      <process seq="2" action="homogenize"
                sample="precipitate" >
                <add_solution solution_ID="sol-A"/>
      </process>
      <process seq="3" action="stand"
                time="60" time_unit="min"
                temp="37" temp_unit="degree in C"
/>
      <process seq="4" action="centrifuge"
                sample="suspension"
                time="20" time_unit="min">
               <times_g>12000</times_g>
       </process>
```

```
      <process seq="5" action="store"
                sample="supernatant"
                temp="-80" temp_unit="degree in C"
  </procedure>
  <comment_e
  </extraction>

- <solution solution_ID="sol-A" label="2-DE lysis solution">
  <item_solution con="9.8" unit="M" name="Urea" />
  <item_solution con="2" unit="% w/v" name="NP-40" />
  <item_solution con="2" unit="% v/v"
name="Pharmalyte(pH3-10)" />
  <item_solution con="10" unit="mM" name="DDT" />
  <item_solution con="0.5" unit="micro g/mL" name="E-64"
/>
  <item_solution con="0.5" unit="mM" name="PMSF" />
  <item_solution con="40" unit="micro g/mL" name="TLCK"
/>
  <item_solution con="1" unit="micro g/mL"
name="aprotinin" />
  <item_solution
name="chymos
  <item_solution
  <item_solution
  <comment_solution />
  </solution>
```

Procedure :
(action, target, condition) lists

Solution list :
solution item information

Sample of HUP-ML (3)

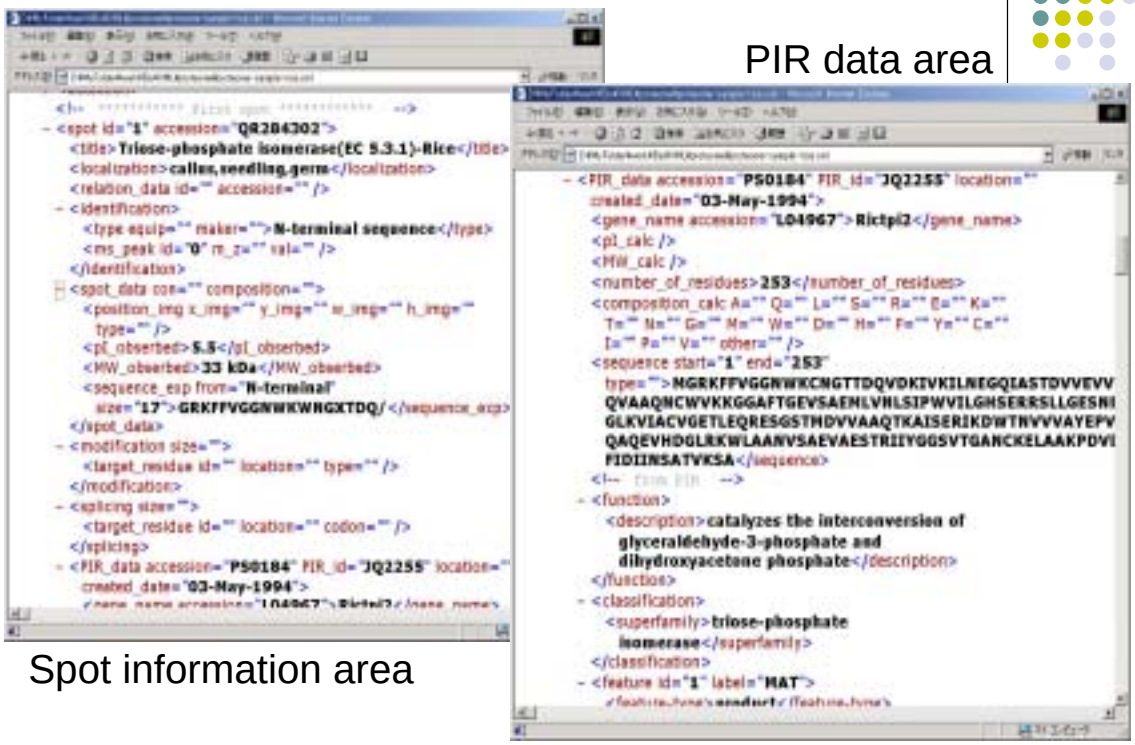
Gel condition

```
- <gel_conditions gel_conditions_ID="" creDate="2002-07-26T10:00:00"
    modDate="2002-08-10T17:20:00">
- <first_dim>
  - <gel_info>
      <gel_name maker="">linear dry strip</gel_name>
      <gel_pH low="3" high="10" />
      <gel_size length="24" unit="cm" />
    </gel_info>
  - <protein_solution solution_size="400" solution_unit="micro L"
       protein_amount="100" protein_unit="micro g" guiding_dye="PBP">
      <description>including standard proteins</description>
    </protein_solution>
    <rehydrate temp="20" temp_unit="degree in C" time="12" unit="hour" />
  - <running>
      <apply step="1" current="50" current_unit="micro A"
        voltage="500" voltage_unit="V" temp="20" temp_unit="degree in C"
        time="1" unit="hour" />
      <apply step="2" current="50" current_unit="micro A"
        voltage="1000" voltage_unit="V" temp="20" temp_unit="degree in C"
        time="1" unit="hour" />
      <apply step="3" current="50" current_unit="micro A
        voltage="8000" voltage_unit="V" temp="20" temp_
        time="10" unit="hour" />
    </running>
  <IEF pH_low="3" pH_high="10" load_direction="cathode to anode" />
```

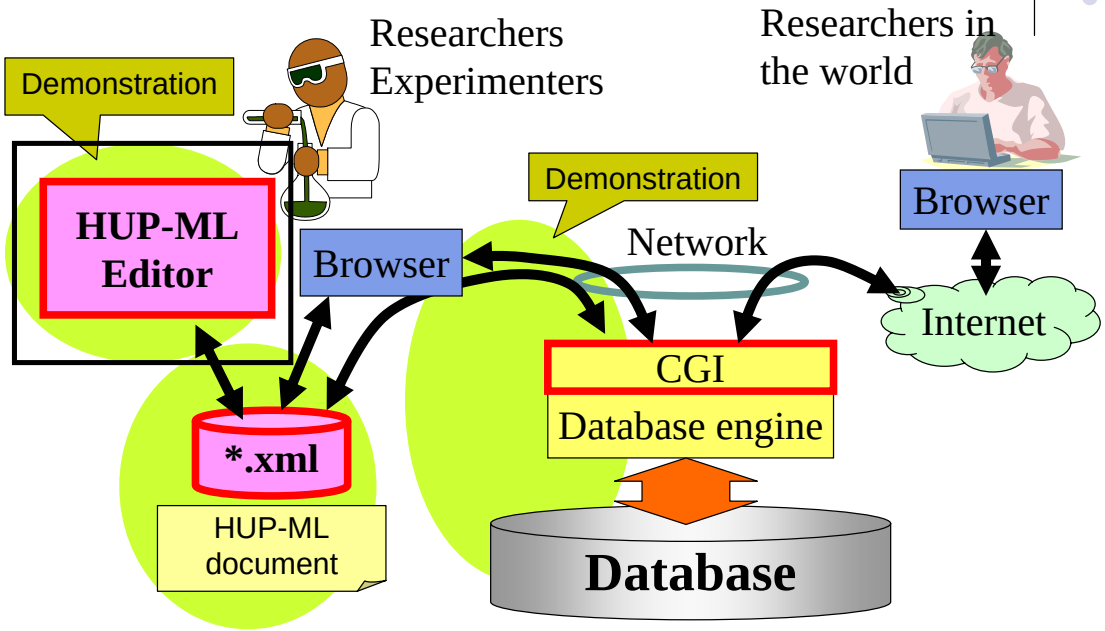
Gel Information :
Size, pH, ...

Running :
(action, condition) lists

Sample of HUP-ML (4)



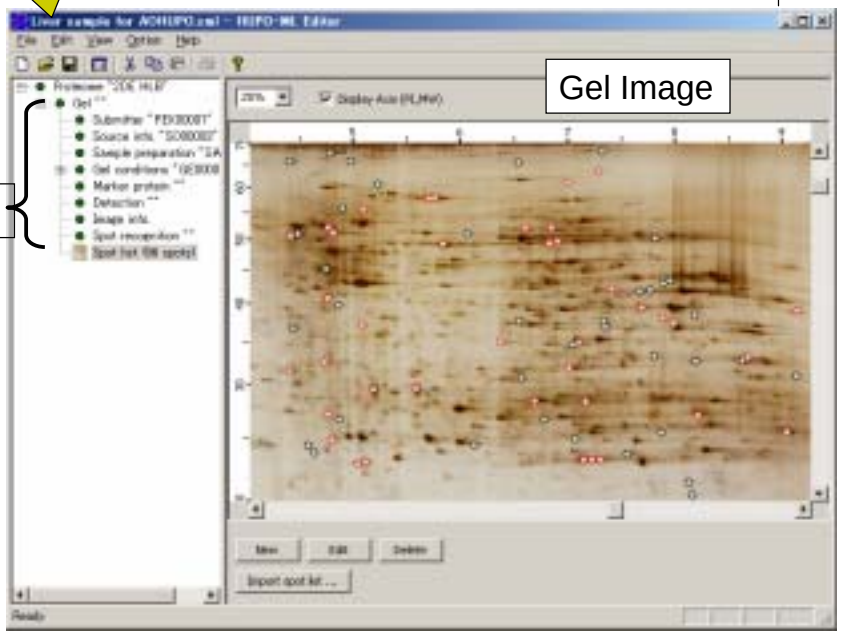
HUP-ML document exchange scheme



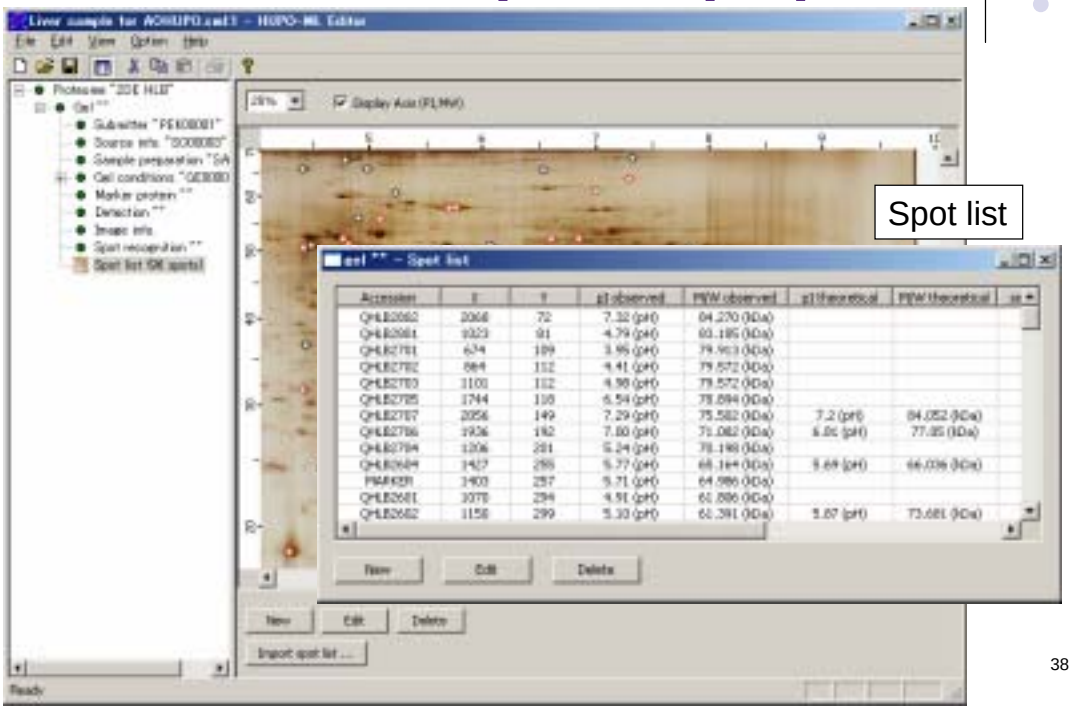
HUP-ML Editor for Proteomics Information

Our XML Document

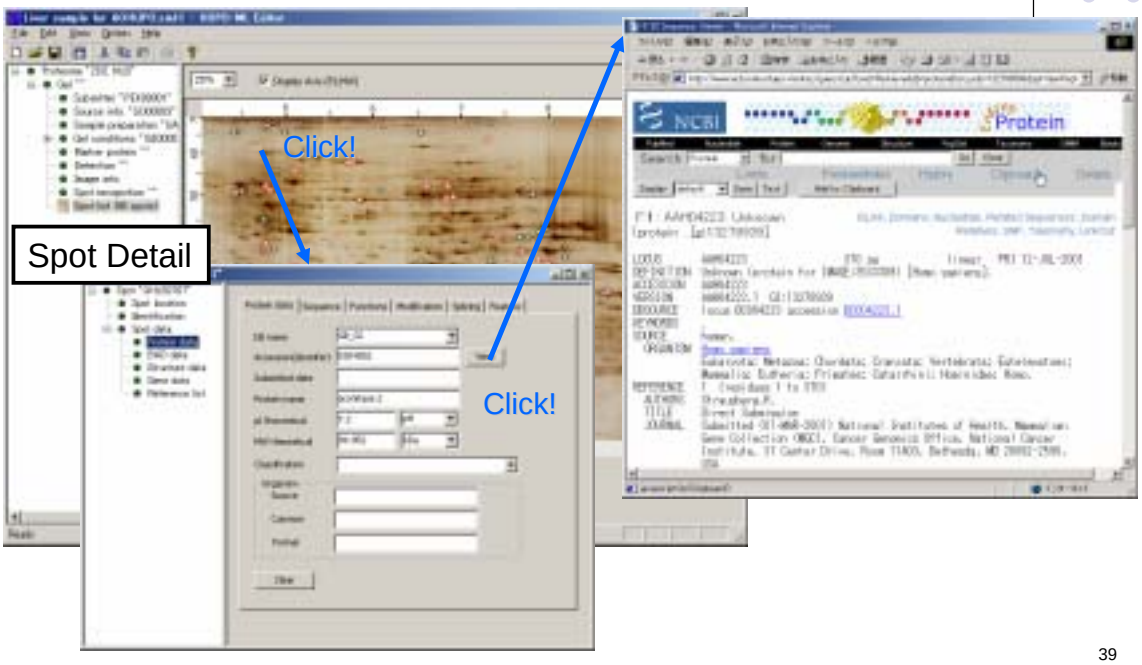
Gel Info.



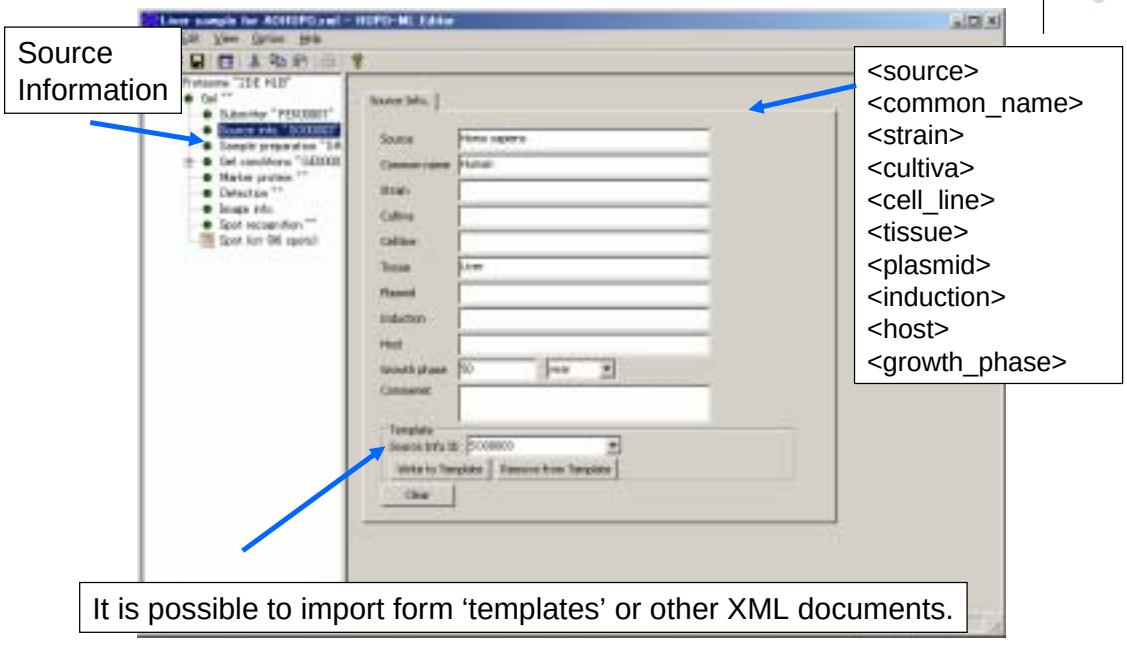
HUP-ML Editor (Example)



HUP-ML Editor (Browsing)



HUP-ML Editor
(Source Information)



Features of our data model

Our proteomics XML:

- describes sample preparations
  - Improves reliability of analysis results
- can distribute experimental information
  - share know-how
  - improves skills
- handle both gel-image and analysis results

- describes analysis information
  - image recognition

Overview (HUP-ML)

- Introduction
- XML features
- XMLs in life science
- HUP-ML concept
- Example of HUP-ML
- Details of HUP-ML

- Demonstration of “HUP-ML Editor”

Description Items in HUP-ML

- Related materials
  - An example of HUP-ML document
  - Current version of DTD file
  - Item list in DTD
  - CD-ROM (HUP-ML Editor version 0.43 beta)

Overview (HUP-ML)

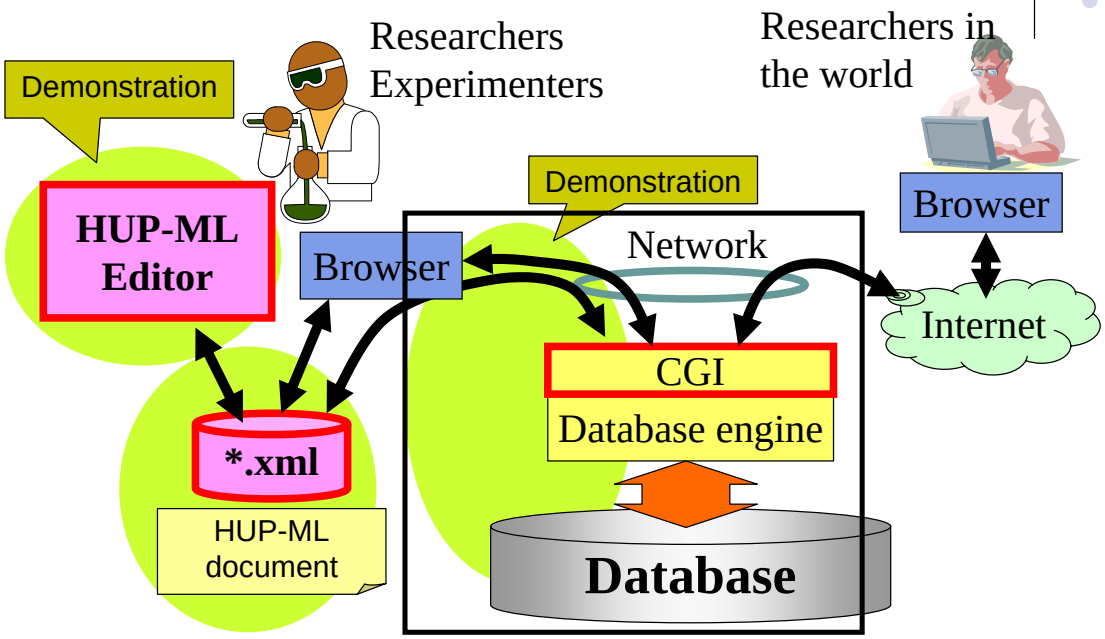
- Introduction
- XML features
- XMLs in life science
- HUP-ML concept
- Example of HUP-ML
- Details of HUP-ML

- Demonstration of “HUP-ML Editor”

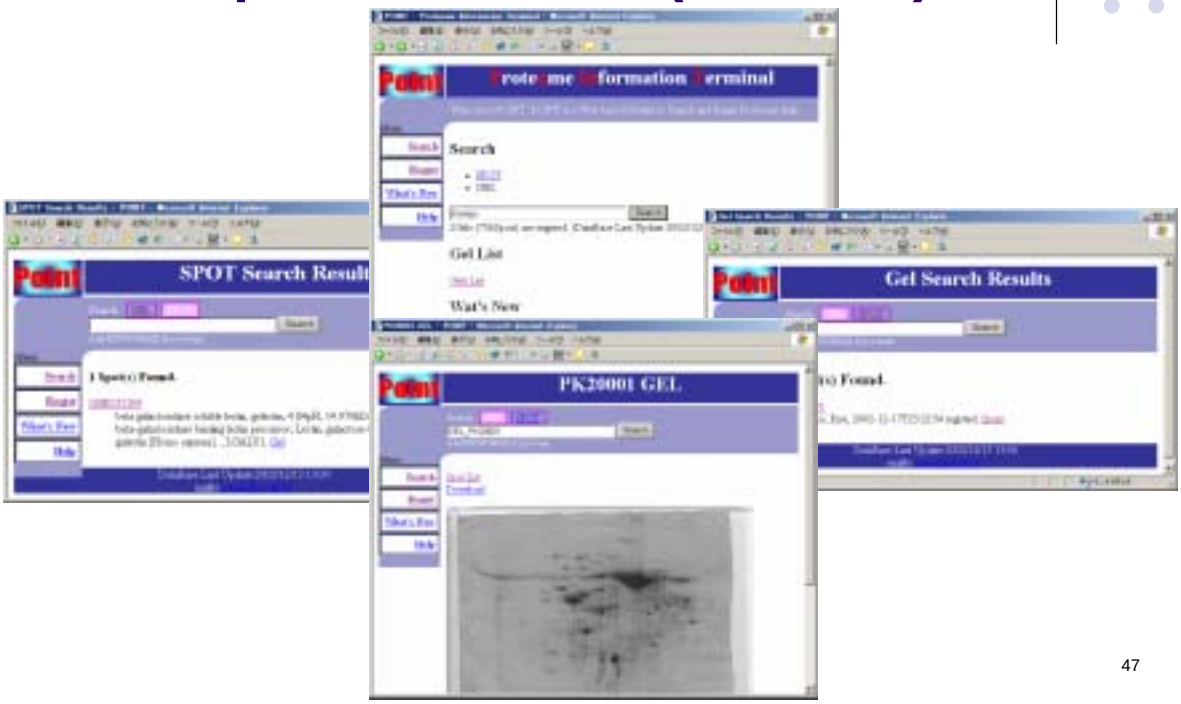
Developments

- HUP-ML DTD
  - Collaboration with AOHUPO
  - Open current version of DTD
- HUP-ML editor for free distribution
  - demonstrates later
- Prototype WWW-based management system
  - for registration, viewing, and retrieval of entries

HUP-ML document exchange scheme



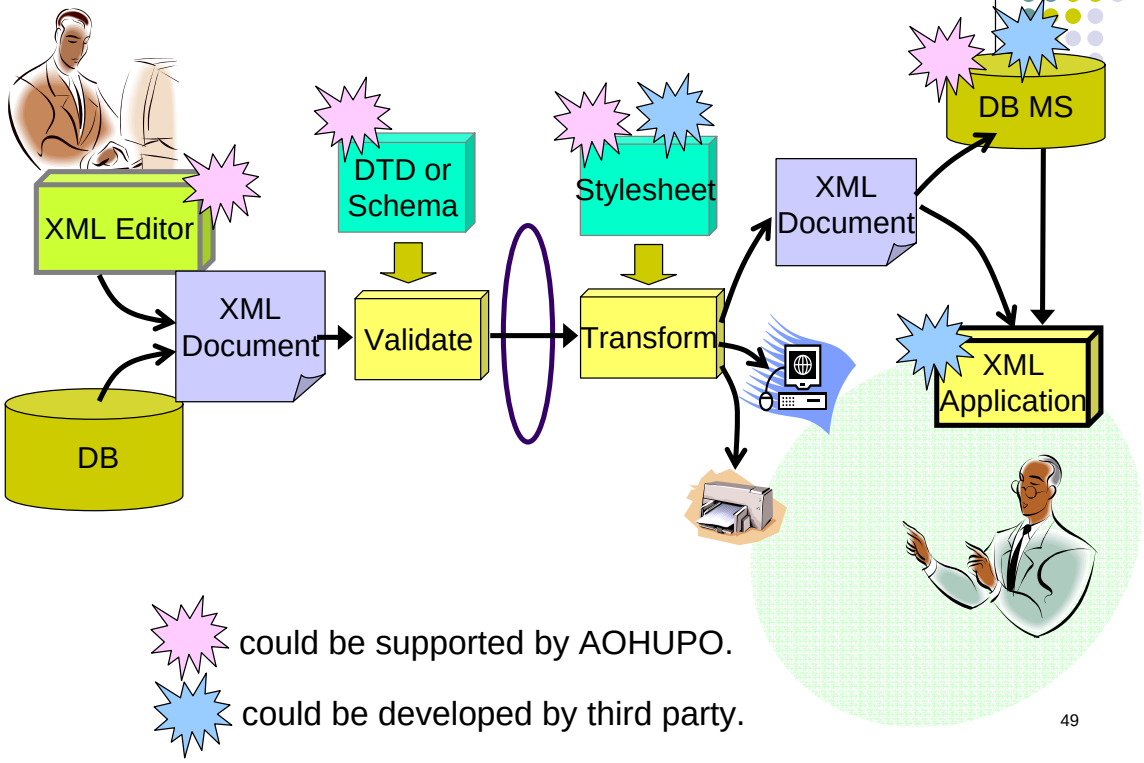
Example of WWW UI (Search)



Future work

- Refine HUP-ML specification and description items
  - Collaboration with AOHUPO
- Version up HUP-ML editor
- Convert from other XML formats
- Related tools
  - XML schema
  - XML Stylesheet for HUP-ML
- Relation to other analysis tools
  - image-analysis software
  - homology-analysis tools, etc.
- Collaboration with other standardization groups

Our XML Workflows



could be supported by AOHUPO.

could be developed by third party.