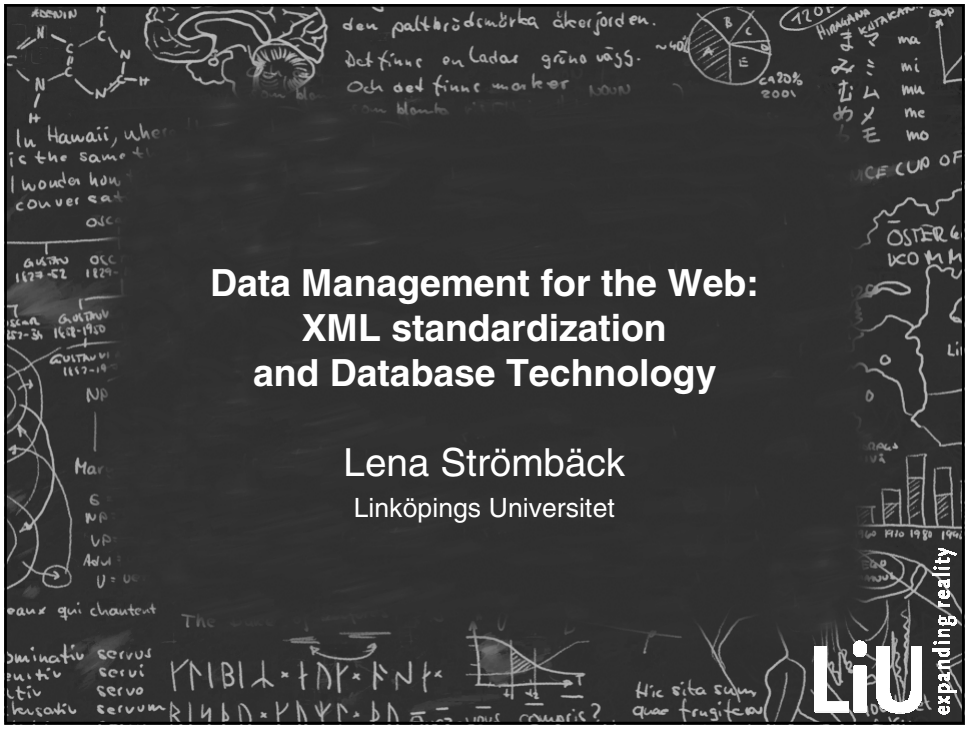
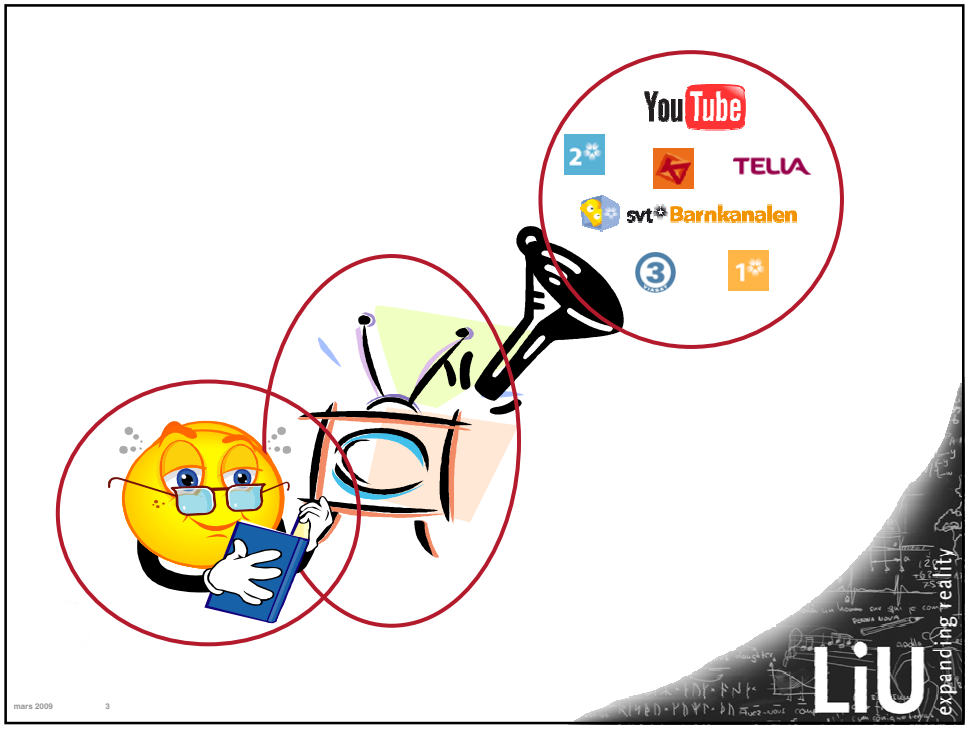
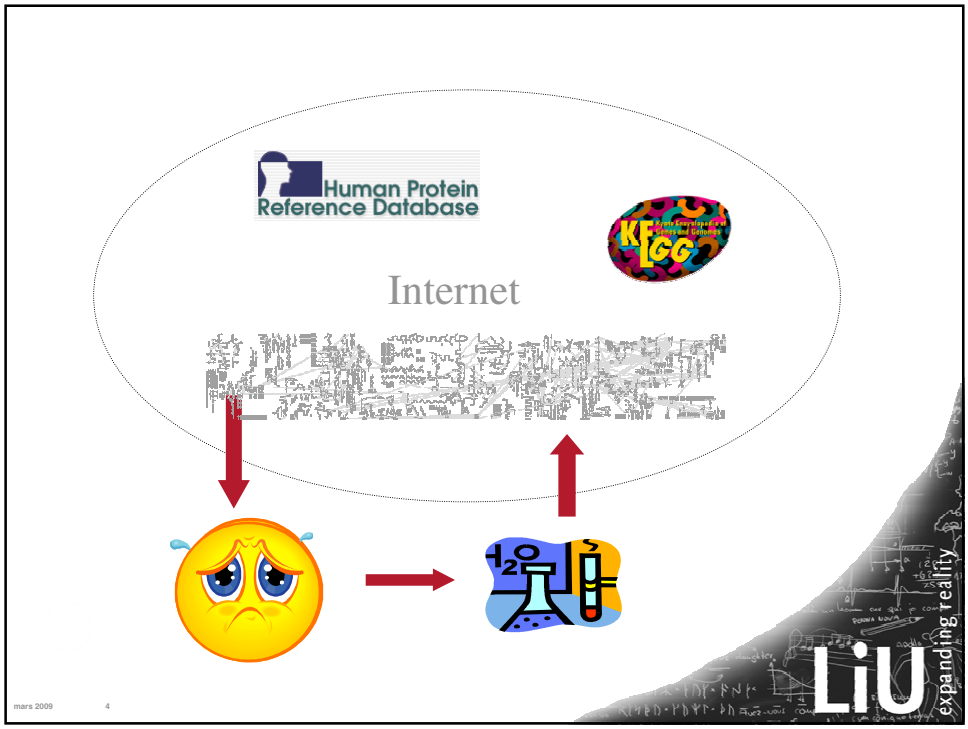
Data Management for the Web: XML standardization and Database Technology

Lena Strömbäck
Linköpings Universitet



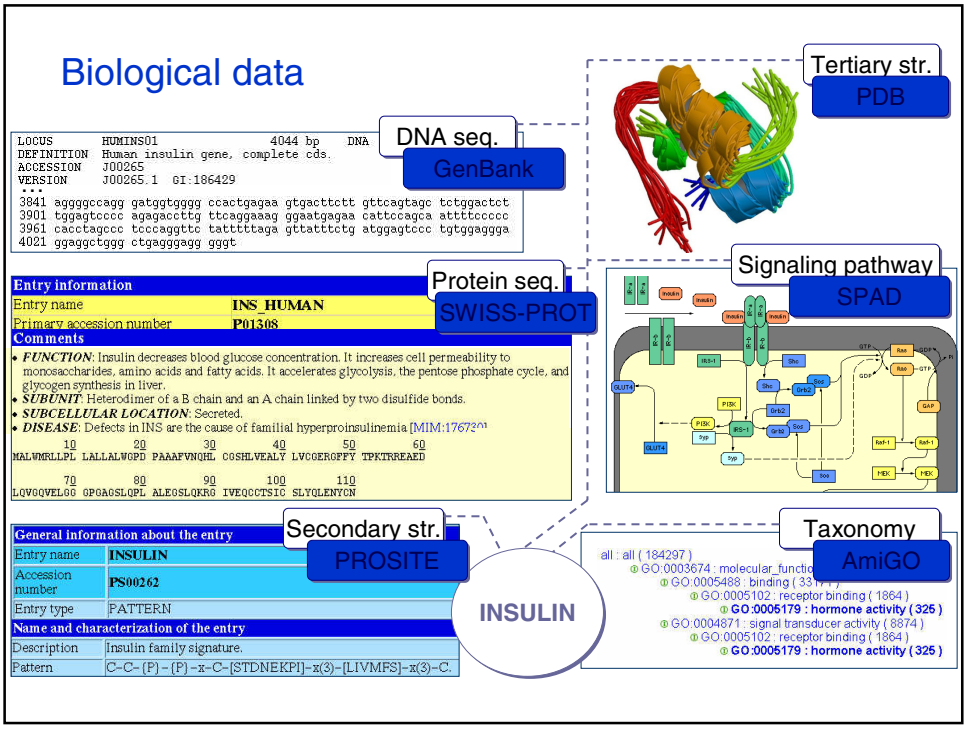


LiU expanding reality

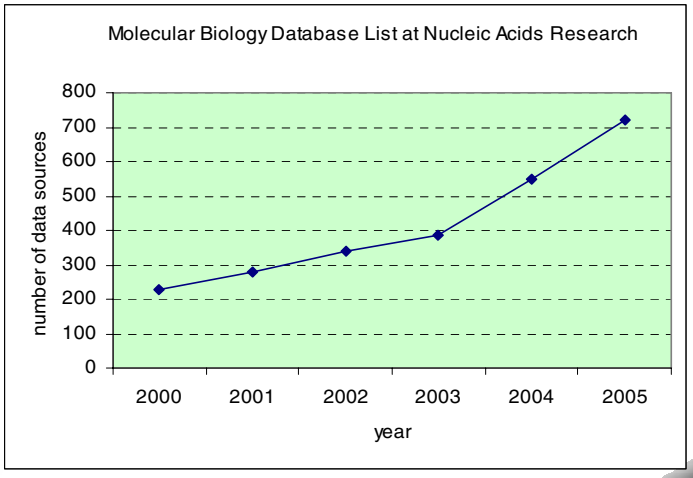


LiU expanding reality

Biological data



Biological data sources



Outline

- What is the problem?
- Standardization for the Web
- Impact on database technology

What is the problem?

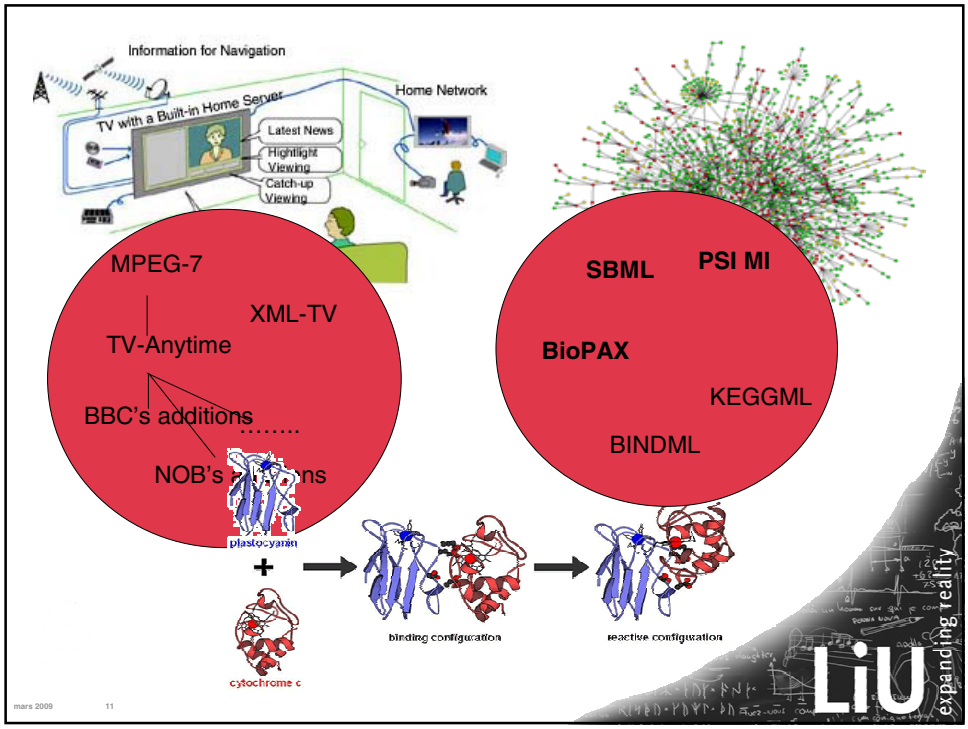
- The user's effort is not enough for the task
- The data describes complex real world objects
- The data is not easily human interpretable

A contradiction?

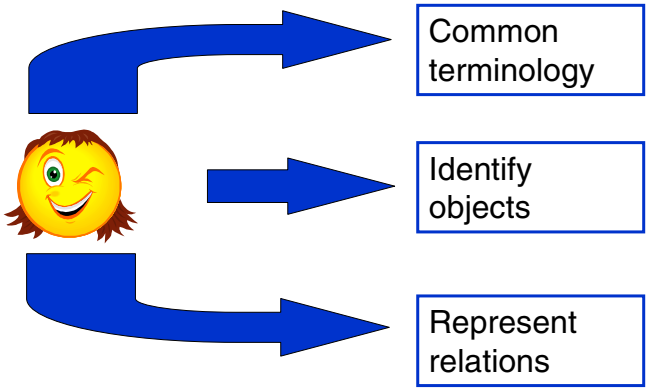
- Data integration is a hard problem
- Standardization and agreement of common format is a prerequisite for efficient data integration
- The Web is an ad-hoc platform where new data formats and actors occurs all the time

Standardization for the Web

- How are standards used?
- What do they contain?



What do we need to standardize?



How to represent data …

- Web data is often less structured than traditional applications
  - Semi structured
  - Integration
- XML is the most common language
- OWL and RDF are alternatives

The TV domain

TV-Anytime

```
<TVAMain>
 <ProgramDescription>
  <ProgramLocationTable>
   <Schedule serviceIDRef='SVT1'
          start='...' end='...'>
    <ScheduleEvent>
     <Program crid='crid:...'/>
    </ScheduleEvent>
   </Schedule>
  </ProgramLocationTable>
  <ProgramInformationTable>
   <ProgramInformation
        programId='crid:...'>
    <BasicDescription>
     <Title>SVT News</Title>
     <Genre href= ... >
      <Name>News</Name>
     </Genre> ...
    </BasicDescription>
   </ProgramInformation>
  </ProgramInformationTable>
</ProgramDescription>
</TVAMain>
```

XMLTV

```
<tv>
 <channel id="C1">
  <display-name lang="se">
       SVT1</display-name>
</channel>
<programme
        start="200006031633"
        channel="C1">
 <title lang="sv">Nyheterna</title>
 <title lang="en">News</title>
 <desc lang="sv"> ... </desc>
 <category>News</category>
 <country>SE</country>
</programme>
</tv>
```

Within molecular interactions ….

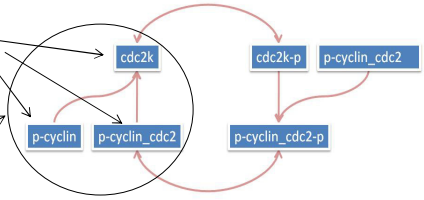
Name	Ver.	Year	Defined by	Purpose	Tools	Data
SBML	2	2003	Systems Biology Workbench development group.	A computer-readable format for representing models of biochemical reaction networks.	Many tools available.	Data available from many databases, for instance, KEGG and Reactome.
PSI MI	2.5	2005	Proteomics Standards Initiative.	A standard for data representation for protein-protein interaction to facilitate data comparison, exchange and verification.	Tools for viewing and analysis.	Datasets available from many sources, for instace IntAct, DIP and MINT.
BioPAX	2	2005	The BioPAX group.	A collaborative effort to create a data exchange format for biological pathway data.	Existing tools for OWL such as Protégé.	Datasets available from Reactome.
CellML	1.1	2002	University of Auckland and Physiome Sciences, Inc.	Support the definition of models of cellular and subcellular processes.	Tools for publication, visualization, creation and simulation.	CellML Model Repository (~240 models).
CML	2.2	2003	Peter Murray-Rust, Henry S. Rzepa.	Interchange of chemical information over the Internet and other networks.	Molecular browsers, editors.	BioCYC.
EMBLxml	1.0	2005	EBI.	More stability and fine-grained modelling of nucleotide sequence information.	API support in BioJavaX.	EMBL.
INSDseq	1.4	2005	International Nucleotide Sequence Database Collaboration.	The purpose of INSDSeq is to provide a near-uniform representation for sequence records.	API support in BioJavaX.	EMBL, DBJ and GenBank.
Seqentry	n/a	n/a	NCBI.	NCBI uses ASN.1 for the storage and retrieval of data such as nucleotide and protein sequences. Data encoded in ASN.1 can be transferred to XML.	SRI's BioWarehouse and ProteinStructureFactory's ORFer.	Entrez.
BSML	3.1	2002	Labbook.com.	Facilitate the interchange of data for more efficient communication within the life sciences community.	Labbook's Genomic Browser and Sequence Viewer. Converters.	Previously provided by EMBL.
HUP-ML	0.8	2003	JHUPO.	A proteomics-oriented markup language for exchanging proteome data between researchers.	HUP-ML Editor.	
MAGE-	1.1	2003	MGED.	To facilitate the exchange	Converters.	

What do the standards contain?

- Information about objects:
  - Proteins/Complexes
  - Genes/DNA
  - Other molecules
- Interaction information
- Information about experiments
  - Kind of experiment
  - Evidence of the experiment
- More ….

SBML model example

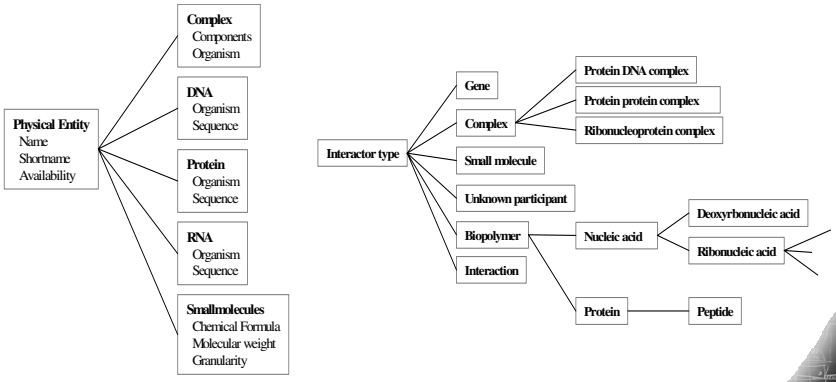
```
<model id="Tyson1991CellModel_6"
       name="Tyson1991_CellCycle_6var">
 <listOfSpecies>
   + <species id="C2" name="cdc2k" compartment="cell">
   + <species id="M" name="p-cyclin_cdc2" compartment="cell">
   + <species id="YP" name="p-cyclin" compartment="cell"> …
  </listOfSpecies>
  <listOfReactions>
   <reaction id="Reaction1" name="cyclin_cdc2k dissociation">
    <annotation>
      <rdf:li rdf:resource="http://www.reactome.org/#REACT_6308"/>
      <rdf:li
     rdf:resource="http://www.geneontology.org/#GO:0000079"/>
    </annotation>
    <listOfReactants>
      <speciesReference species="M"/>
    </listOfReactants>
    <listOfProducts>
      <speciesReference species="C2"/>
      <speciesReference species="YP"/>
    </listOfProducts>
    <kineticLaw>
      <math xmlns="http://www.w3.org/1998/Math/MathML">
        <apply> <times/> <ci> k6 </ci> <ci> M </ci>  </apply></math>
      <listOfParameters>   <parameter id="k6" value="1">
      </listOfParameters>
    </kineticLaw>
   </reaction>
   + <reaction id="Reaction2" name="cdc2k phosphorylation">
   ... more reactions
  </listOfReactions>
 </model>
</sbml>
```



Representation of objects

	SBML: Species	PSI MI: Interactor		CellML:component	
1	id	id		name	1
2	name	names		dc:title	2
				dcterms:alternative	
3		xref		cmeta:bio_entity	3
4	speciesType	interactortype			4
5		organism			5
6			ncbiTaxId		6
7			names	cmeta:species	7
8			celltype		8
9	compartment		compartment	(~ group)	9
10			tissue		10
11				cmeta:sex	11
12		sequence			12
13				variable	13
14	initialAmount				initialvalue 14
15	initialConcentration				15
16	substanceUnits				units 16
17	spatialsizeUnits				17
18	hasonlysubstanceUnits				18
19	boundaryCondition				19
20	charge				20
21	constant				21

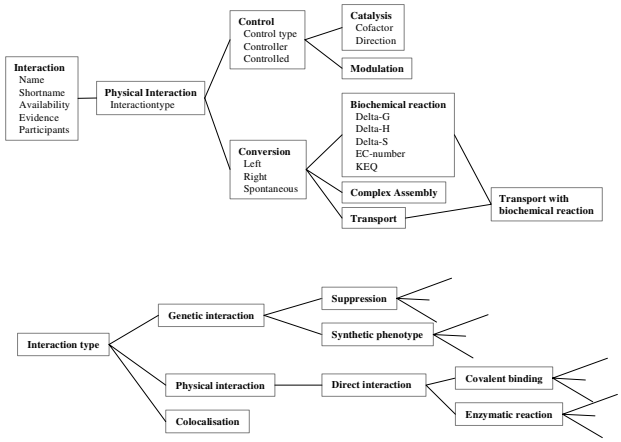
Type of objects - ontologies



Representation of interactions

	CellML: component	SBML: Reaction	PSI MI: Interaction
1			imexID
2	name	id	id
3		name	
4			xref
5	variable		
6	reaction		
7		sboTerm	interactiontype
8			experimentList
	variable-ref	reactant	participantList
		product	
9		modifier	
10		id	id
11		name	names
12			experimental-role
13	role	sboTerm	biological-role
14			participantidentification
15			experimentalpreparation
16			confidencelist
17	direction		
18	delta_variable		
19	stoichiometry	stoichiometry	
20		kineticLaw	
21			inferredInteractionlist
22			participants
23			modelled
24			confidencelist
25	reversible	reversible	
26		fast	

Interaction types



Bioinformatics:
Minimal Information Standards.

- **MIRIAM : Minimal Information Requested in the Annotation of biochemical Models**
- **MIAPE: The Minimum Information About a Proteomics Experiment**
  - **MIAPE: GE**(Gel Electrophoresis)
  - **MIAPE: MS** (Mass Spectrometry)
  - **MIAPE: CC** (Column Chromatography)
  - **MIAPE: CE** (Capillary Electrophoresis)
  - ….
- **MIMIx: The minimum information required for reporting a molecular interaction experiment** ….
- ….

XML and storage

- XML provides a data model
- The valid XML data structures can be defined by
  - XML Schema
  - DTD
- XML has its own query languages
  - XPath
  - XQuery
- XML is richer than the relational model
  - Tree structure,
  - Order
  - …
- Vary from highly structured to unstructured

Expressing Queries in XQuery

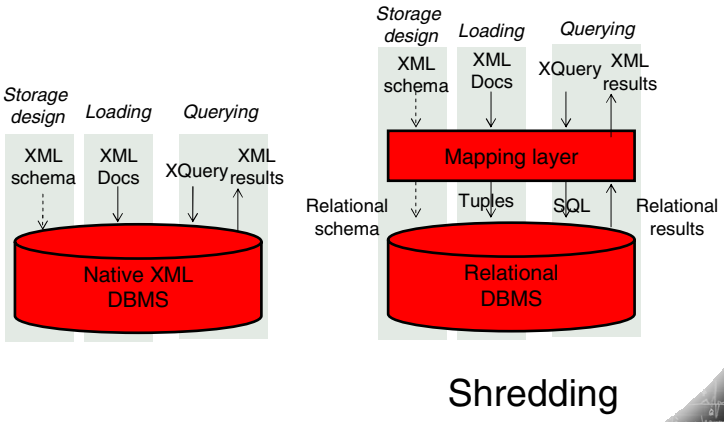
Find information on a given protein. Protein id is given.

```
document("rat_small.xml")//proteinInteractor[@id="EBI-77471"]
```

Find the protein information for the proteins that participate in a given interaction. Interaction id is given.

```
for $ref in document("rat_small.xml")//interaction
    [names/shortLabel="interaction1"]
    /participantList/proteinParticipant/proteinInteractorRef/@ref
  return document("rat_small.xml")//proteinInteractor[@id=$ref]
```

Storage possibilities for XML



Shredding

How to shred XML?

```
<?xml version="1.0" encoding="UTF-8"?>
<families
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance>
   <family>
      <parent>
         <name>Lena</name>
         <job>Lektor</job>
      </parent>
      <child>
         <name>Ludvig</name>
         <school>Skolan</school>
      </child>
   </family>
</families>
```

Source	Ordinal	attrName	isValue	Value
0	1	Families	False	1
1	1	Family	False	2
2	1	Parent	False	3
3	1	Name	True	Lena
3	2	Job	True	Docent ..

Families

Id	Pid
0	-

Family

Id	Pid
1	0

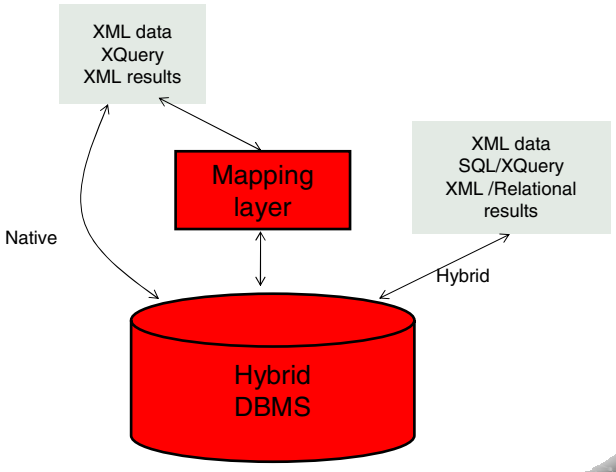
Parent

Id	Pid	Name	Job
2	1	Lena	Lektor

Child

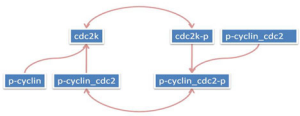
Id	Pid	Name	School
3	1	Ludvig	Skolan

Hybrid XML Storage



New possibilities…

```
<model id="Tyson1991CellModel_6"
       name="Tyson1991_CellCycle_6var">
 <listOfSpecies>
   + <species id="C2" name="cdc2k" compartment="cell">
   + <species id="M" name="p-cyclin_cdc2" compartment="cell">
   + <species id="YP" name="p-cyclin" compartment="cell"> …
  </listOfSpecies>
  <listOfReactions>
   <reaction id="Reaction1" name="cyclin_cdc2k dissociation">
     <annotation>
       <rdf:li rdf:resource="http://www.reactome.org/#REACT_6308"/>
       <rdf:li
      rdf:resource="http://www.geneontology.org/#GO:0000079"/>
     </annotation>
     <listOfReactants>
       <speciesReference species="M"/>
     </listOfReactants>
     <listOfProducts>
       <speciesReference species="C2"/>
       <speciesReference species="YP"/>
     </listOfProducts>
     <kineticLaw>
       <math xmlns="http://www.w3.org/1998/Math/MathML">
         <apply> <times/> <ci> k6 </ci> <ci> M </ci>  </apply></math>
       <listOfParameters>   <parameter id="k6" value="1">
       </listOfParameters>
     </kineticLaw>
   </reaction>
   + <reaction id="Reaction2" name="cdc2k phosphorylation">
   ... more reactions
  </listOfReactions>
 </model>
</sbml>
```



Species:

Id	Name	Compartment
C2	cdc2k	cell
M	p-cyclin_cdc2	cell
YP	p-cyclin	cell
….	….	….

Reaction:

Id	Name	Annotation	Formula
Reaction1	cyclin_cdc2k dissociation	<annotation …..>	<kinetic_law …..>
Reaction2	cdc2k phosphorylation	<annotation ….>	<kinetic_law ….>
….	….	….	….

Reactants:

Id	Species
Reaction1	M
Reaction2	….
….	….

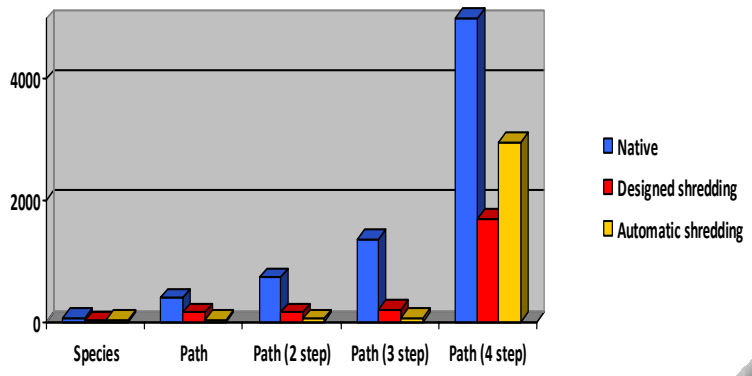
Products:

Id	Species
Reaction1	C2
Reaction1	YP
….	….

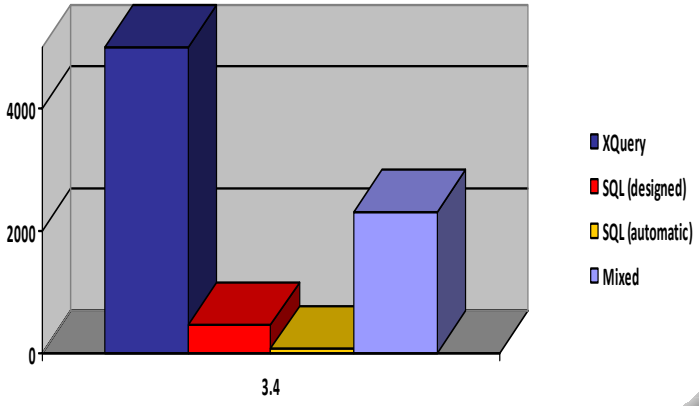
SQL and Xpath/XQuery

```
select, r.name, s.name
from reaction r, products p, species s
where r.id = p.id and p.species=s.id;

xquery
for  $y in db2-fn:xmlcolumn('SBML_DATA.SBML_DOC')
          /model/listOfReactions/reaction/listOfModifiers/modifierSpeciesReference,
     $z in db2-fn:xmlcolumn('SBML_DATA.SBML_DOC')/model/listOfSpecies/species[@id = $y/@species]
return <product> {$y/../../@name} {$z/@name} </product>

SELECT  p.reaction, species.name
from species,
(SELECT X.*
  FROM reactome_data,
  XMLTABLE ('$d/model/listOfReactions/reaction/listOfModifiers/modifierSpeciesReference' passing
      reactome_doc as "d"
  COLUMNS
  product VARCHAR(200) PATH '@species',
  reaction VARCHAR(200)                PATH '../../@id') AS X) p
where p.product=species.id
```

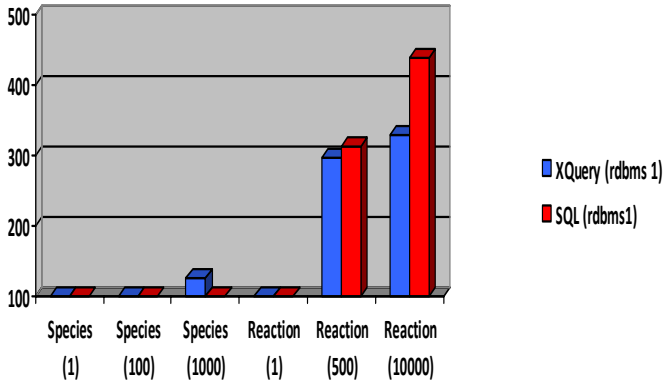

Efficiency: Increasing query complexity



Efficiency:
Combining representations

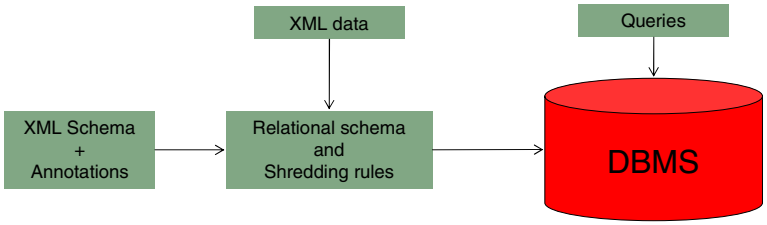


Efficiency:
Return the result as XML



Tool development: HShreX

- Need a tool to speed up the process



- (Further development of work by Amer-Yahia et al.)

Working with HShreX:

```
<?xml version="1.0" encoding="UTF-8"?>
    <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
      xmlns:shrex="http://www.cse.ogi.edu/shrex">

      <xs:element name="families">
        <xs:complexType>
          <xs:sequence maxOccurs="unbounded">
            <xs:element name="family" type="familyType"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>

      <xs:complexType name="familyType">
        <xs:sequence>
          <xs:element name="parent" type="parentType" >
          <xs:element name="child" type="childType" >
        </xs:sequence>
      </xs:complexType>

      <xs:complexType name="parentType">
        <xs:sequence>
          <xs:element name="name" type="xs:string"/>
          <xs:element name="job" type="xs:string"/>
        </xs:sequence>
      </xs:complexType>

      <xs:complexType name="childType">
        <xs:sequence>
          <xs:element name="name" type="xs:string"/>
          <xs:element name="school" type="xs:string"/>
        </xs:sequence>
      </xs:complexType>
    </xs:schema>
```

Families

Id	Pid
0	-

Families_family

Id	Pid
1	0

Families_family_parent

Id	Pid	Name	Job
2	1	Lena	Lektor

Families_family_child

Id	Pid	Name	School
3	1	Ludvig	Skolan

Working with HShreX:

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:shrex="http://www.cse.ogi.edu/shrex">

  <xs:element name="families">
    <xs:complexType>
      <xs:sequence maxOccurs="unbounded">
        <xs:element name="family" type="familyType"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:complexType name="familyType">
    <xs:sequence>
      <xs:element name="parent" type="parentType" >
      <xs:element name="child" type="childType"
                                  shrex:maptoxml="true">
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="parentType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="job" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="childType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="school" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>
</xs:schema>
```

Families

Id	Pid
0	-

Families_family

Id	Pid	Child
1	0	<child> <name>Ludvig</name> <school>Skolan/school> </child>

Families_family_parent

Id	Pid	Name	Job
2	1	Lena	Lektor

Working with HShreX:

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:shrex="http://www.cse.ogi.edu/shrex">

  <xs:element name="families">
    <xs:complexType>
      <xs:sequence maxOccurs="unbounded">
        <xs:element name="family" type="familyType"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:complexType name="familyType">
    <xs:sequence>
      <xs:element name="parent" type="parentType"

        shrex:tablename="person">
      <xs:element name="child" type="childType"

        shrex:tablename="person" >
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="parentType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="job" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="childType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="school" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>
</xs:schema>
```

Families

Id	Pid
0	-

Families_family

Id	Pid
1	0

Person

Id	Pid	Name	Job	School
2	1	Lena	Lektor	
3	1	Ludvig		Skolan

Working with HShreX:

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:shrex="http://www.cse.ogi.edu/shrex">

  <xs:element name="families">
    <xs:complexType>
      <xs:sequence maxOccurs="unbounded">
        <xs:element name="family" type="familyType"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:complexType name="familyType">
    <xs:sequence>
      <xs:element name="parent" type="parentType"
                  shrex:withparent="true">
      <xs:element name="child" type="childType" >
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="parentType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="job" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="childType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="school" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>
</xs:schema>
```

Families

Id	Pid
0	-

Families_family

Id	Pid	Name	Job
1	0	Lena	Lektor

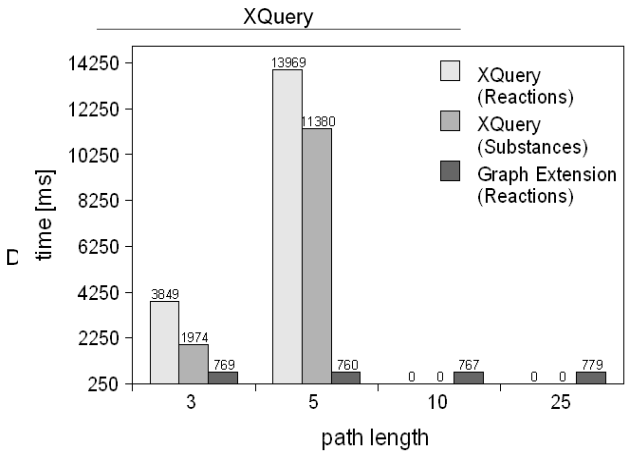
Families_family_child

Id	Pid	Name	School
3	1	Ludvig	Skolan

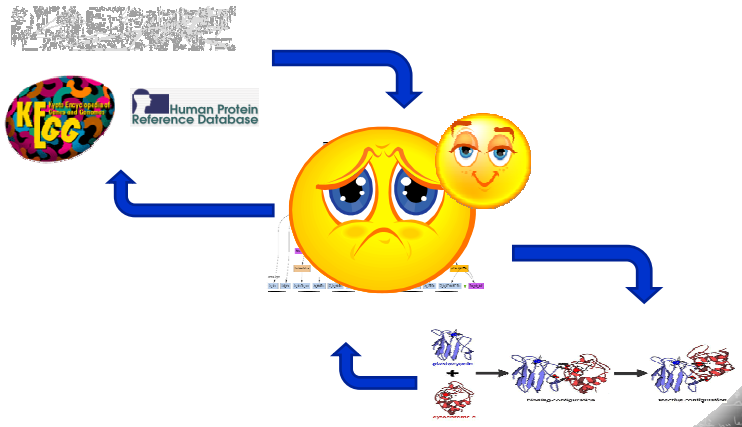
Future challenges

- Hybrid storage, to be as efficient as possible in each case.
- Special solutions for networks and hierarchies

XML with graph queries:



Workflows for data exploration



Capturing provenance

- Provenance of scientific artifacts is necessary to reproduce, validate and share scientific results
- Provenance can be as important as the results!

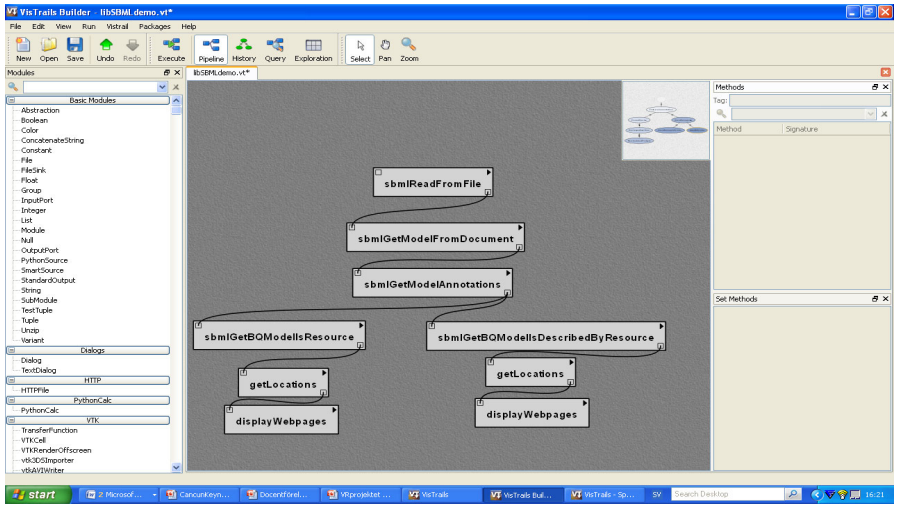
Dictionary

prov•e•nance |ˈprävənəns|
noun
the place of origin or earliest known history of something : *an orange rug of Iranian provenance.*
• the beginning of something's existence; something's origin : *they try to understand the whole universe, its provenance and fate.*
See note at ORIGIN .
• a record of ownership of a work of art or an antique, used as a guide to authenticity or quality : *the manuscript has a distinguished provenance.*

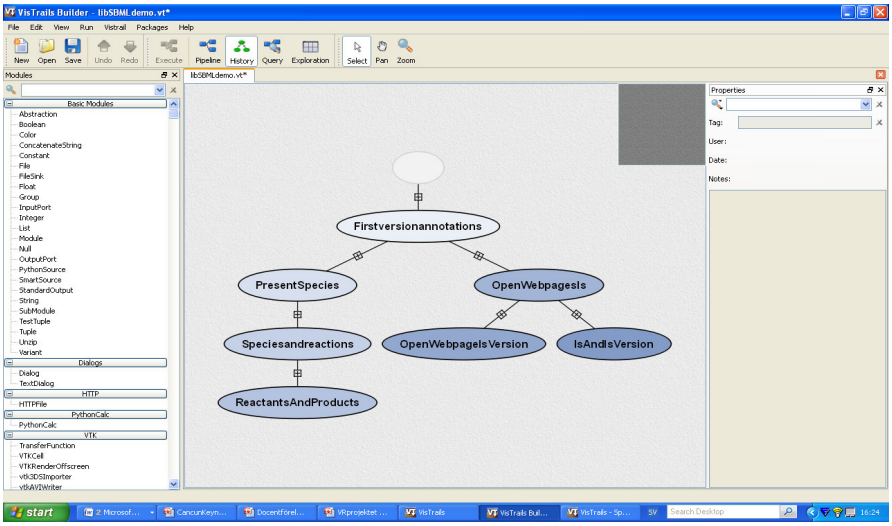
ORIGIN late 18th cent.: from French, from the verb ***provenir 'come or stem from,'*** from Latin ***provenire***, from ***pro- 'forth' + venire 'come.'***

LiU expanding reality

Specifying my steps….



…and remembering my actions

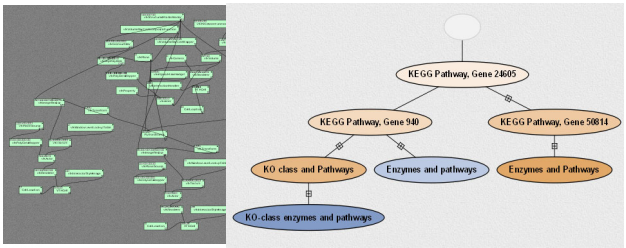


The Vistrails system
(Freire et al. University of Utah)

- *Vision: Provenance enable the world*
- Comprehensive provenance infrastructure for computational tasks
  - Captures provenance transparently
  - Provides intuitive query interfaces for exploring provenance data
  - Supports collaboration
- Designed to support *exploratory tasks such as* visualization and data mining
- VisTrails system is open source: www.vistrails.org
  - >2,000 downloads since beta release in Jan 2007
  - 100% Python--runs on Windows, Linux and Mac

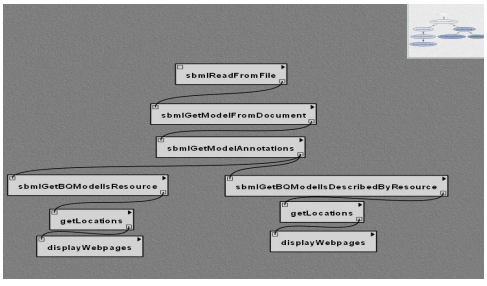
Interesting issues:

- Reuse of other peoples efforts
  - Provenance server
  - Presentation of search results
  - Co-work
  - Interoperability between systems



Interesting issues:

- Bio-specific version of Vistrails
  - Handling XML – LibSBML, XQuery
  - Annotations – use of ontologies
  - Easy to use module library for bioinformatics
  - Combining webdata, own results and visualization



Questions?

People that have contributed to this work:

Linköpings Universitet: Patrick Lambrix, José M Pena, He Tan, Tommy Ellqvist, David Hall, Mikael Åsberg

University of Utah: Juliana Freire, David Koop, Lauro Lins, Claudio Silva

University of Rostock: Dagmar Köhn, Andreas Heuer, Adelinde Uhrmacher, Stefan Schmidt

Thanks!