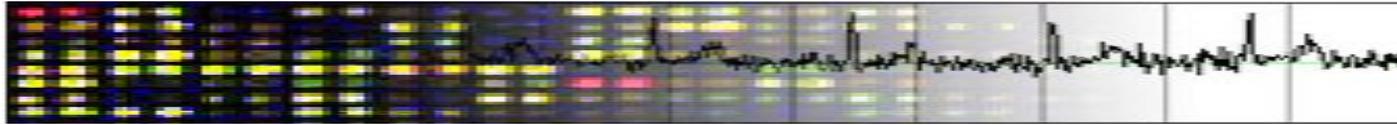


Biomedical Information Technology

2.771J BEH.453J HST.958J Spring 2005

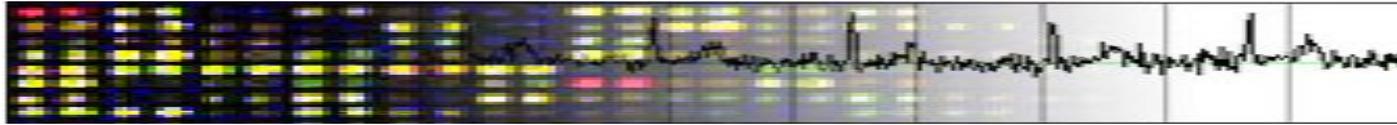
Lecture 28 April 2005

Data Integration and Analysis II: Biological Information Systems



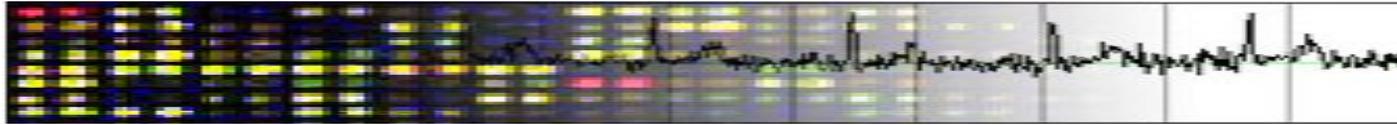
DATA INTEGRATION AND ANALYSIS II

- ❖ Integration in the biological environment
 - New standards are required
 - The I3C: A tale of good intentions
 - XML as the “medium AND the message”
- ❖ Database considerations
 - Strengths of relational databases
 - Weaknesses of the relational model
 - Solutions: the Semantic Web
 - Database federation
- ❖ Adding metadata to images and other records



New standards and methods are required

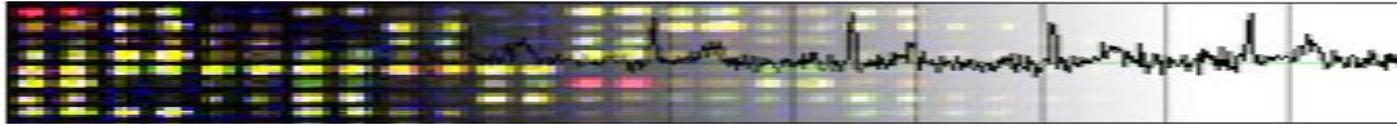
- ❖ Experimental data have grown to terabytes
- ❖ Analyses and other derived data abound
- ❖ Multiple data types exist
- ❖ Need to make transport of data neutral
 - Leave display/interpretation to the receiving program
 - Many programs must interact with the data
 - Allow cascading programs
 - Support maintenance and upgrades
- ❖ Archival storage required for publication
- ❖ Archival storage required for FDA traceability



I3C: Interoperable Information Infrastructure Consortium

I3C in Five Points

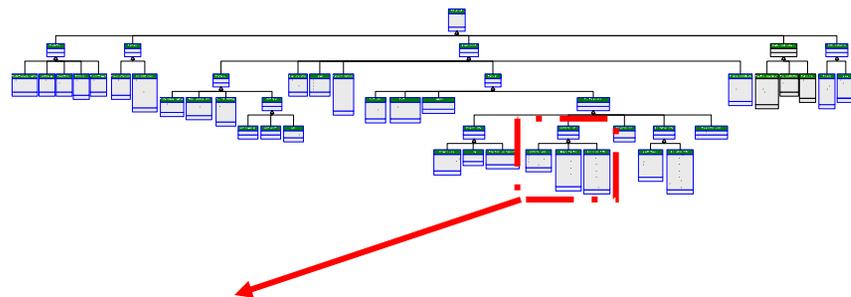
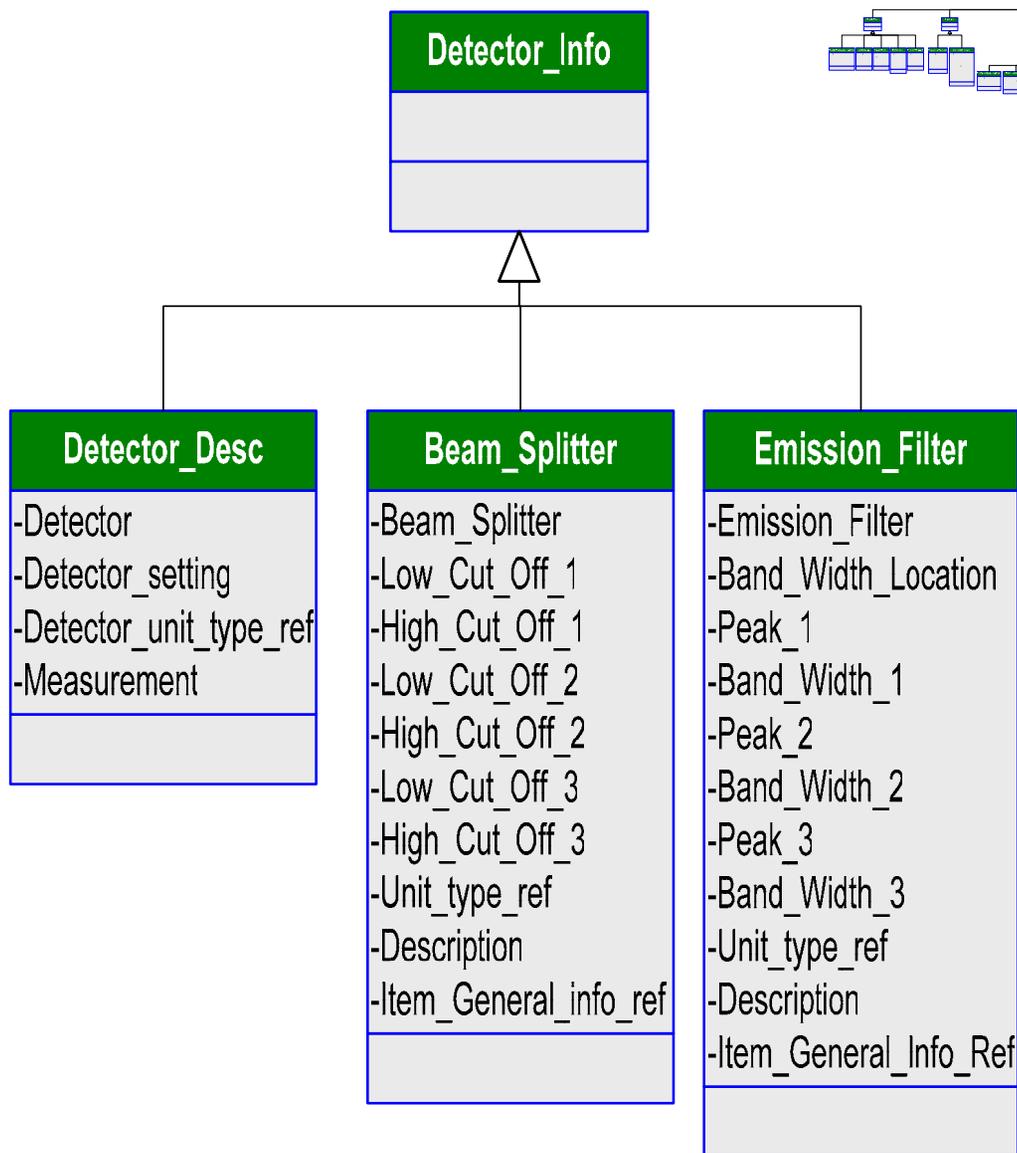
- 1. We are an open, global organization that coordinates and guides the design and development of methodologies and software that support data and tool interoperability.
- 2. Our goal is to accelerate discovery and development in life sciences.
- 3. We use scientific use cases that exemplify common bottlenecks to guide the development of fully documented recommendations or solutions.
- 4. We avoid duplication of effort by following methods, protocols and policies of other groups whenever possible.
- 5. Participation is open for any not-for-profit organization, academic or government research institution, or commercial organization focused on life sciences.

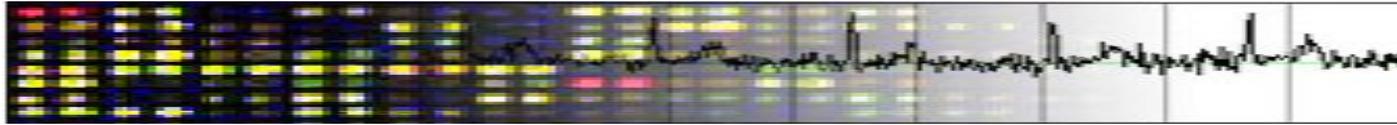


XML is the “medium and the message” (1)

- ❖ XML for schema representation
 - Works with all tree structures which are not multiply connected
 - Can be generated with “reasonable effort”
 - Can foresee exchange of schema and semantics in a form that is parseable and readable and independent of the means for implementation (e.g. SQL for databases)
 - Can easily “cut and paste” to modify schema, but still no tool to make clean object-oriented SQL code for databases

Details of schema



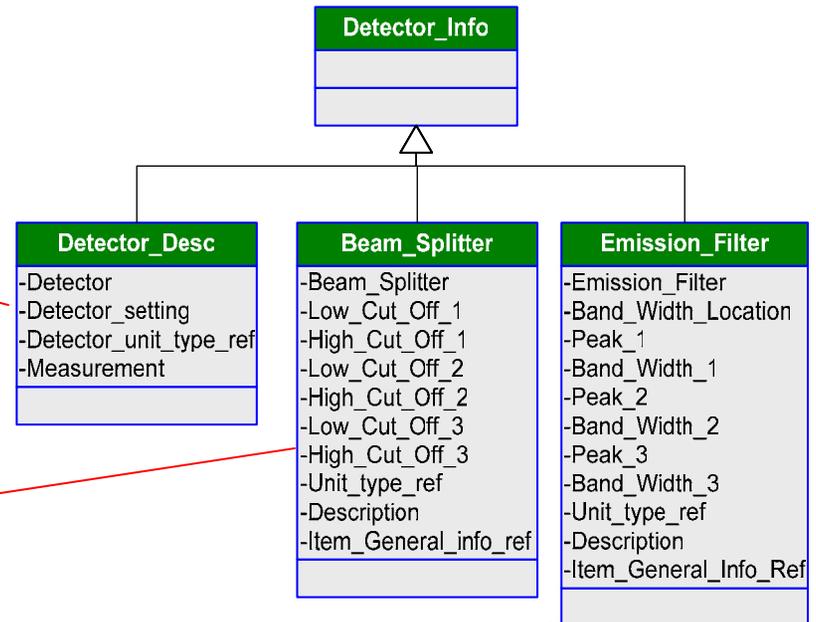


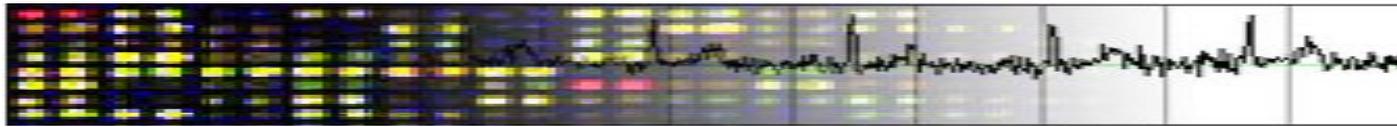
SQL code

```
CREATE TYPE detector_desc_t UNDER detector_info_t AS
(detector varchar(64),
detector_setting real,
detector_unit_pref REF(unit_prefix_t),
detector_unit REF(unit_t),
measurement varchar(64))
MODE DB2SQL;
```

```
CREATE TYPE beam_splitter_t UNDER detector_info_t AS
(beam_splitter varchar(64),
low_cut_off_1 real,
high_cut_off_1 real,
low_cut_off_2 real,
high_cut_off_2 real,
low_cut_off_3 real,
high_cut_off_3 real,
unit_prefix REF(unit_prefix_t),
unit REF(unit_t),
description varchar(64),
item_info REF(item_info_t))
MODE DB2SQL;
```

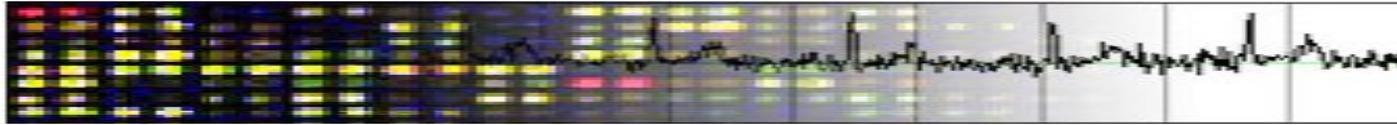
```
CREATE TYPE emission_filter_t UNDER detector_info_t AS
(emission_filter varchar(64),
band_width_loc varchar(16),
peak_1 real,
band_width_1 real,
peak_2 real,
band_width_2 real,
peak_3 real,
band_width_3 real,
unit_prefix REF(unit_prefix_t),
unit REF(unit_t),
description varchar(64),
item_info REF(item_info_t))
MODE DB2SQL;
```





XML is the “medium and the message” (2)

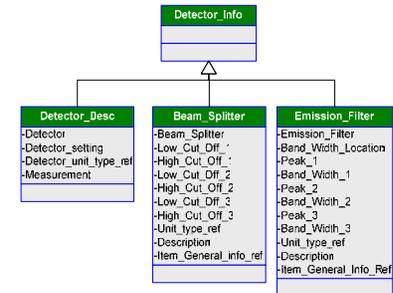
- ❖ XML for neutral transport
 - Self-describing
 - Carries no implicit or explicit presentation or use information
 - Contrast HTML, which carries explicit presentation information but no content meaning
 - Can embed instructions to render content, but that breaks neutrality
 - Can contain “blobs” to support raw data transport (special inefficient mime-type encoding)
- ❖ XML style sheets that conform to specific vocabularies for different application areas.

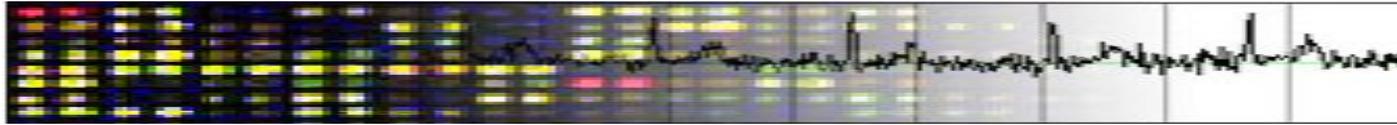


An example XML document

Here XML is used to transport data describing a specific experiment in the database

```
<?xml version="1.0" encoding="UTF-8"?>
<params:Parameter xmlns:params="parameters.xsd" xsi:schemaLocation="parameters.xsd">
  <Dectector_Info>
    <Detector>PMT</Detector>
    <Detector_Setting>600</Detector_Setting>
    <Detector_Units Prefix="none" Si_Unit_Name="volt"/>
    <Measurement>Flourescence</Measurement>
    <Beam_Splitter_Info Prefix="nano" Unit="meter">
      <Beam_Splitter>Dichroic_Reflect_Low</Beam_Splitter>
      <Low_Cut_Off_1>505</Low_Cut_Off_1>
      <Description>505DRLP</Description>
      <Item_General_Info>
        <Manufacturer>Omega Optical</Manufacturer>
        <Model_Name>XF2010</Model_Name>
      </Item_General_Info>
    </Beam_Splitter_Info>
    <Emission_Filter_Info Prefix="nano" Unit="meter">
      <Emission_Filter>Band_Block</Emission_Filter>
      <Band_Width_Location>unknown</Band_Width_Location>
      <Peak_1>535</Peak_1>
      <Band_Width_1>45</Band_Width_1>
      <Description>535AF45</Description>
      <Item_General_Info>
        <Manufacturer>Omega Optical</Manufacturer>
        <Model_Name>XF3084</Model_Name>
      </Item_General_Info>
    </Emission_Filter_Info>
  </Dectector_Info>
</params:Parameter>
```





XML is the “medium and the message” (3)

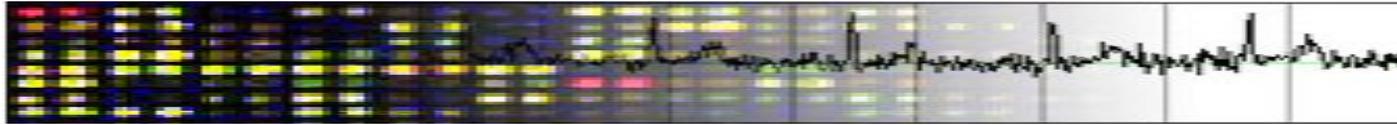
❖ XML for semantic application packages

➤ MathML

- “**2002-10-18: LaTeX to MathML converter.** Stéphan Sémirat has written WeM: an MathML editor that converts a subset LaTeX to MathML . It can be tested on line (<http://mathosphere.net/editeurml/WeM.html>).

➤ CellML

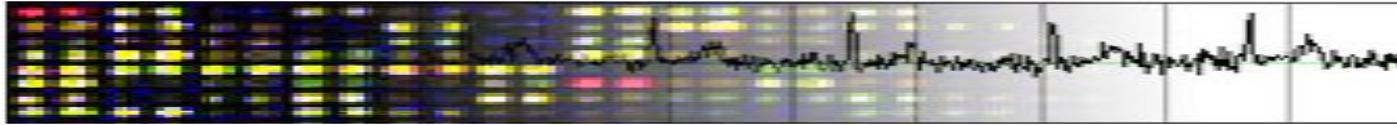
- “CellML is an Extensible Markup Language (XML) being developed by Physiome Sciences, the University of Auckland, and the CellML Working Group to provide a standard method for representing and exchanging computer-based biological models”.
- MathML embedded in CellML documents is used to define the underlying mathematics of models.



Details of the XML representation of metadata in the Open Microscopy Environment

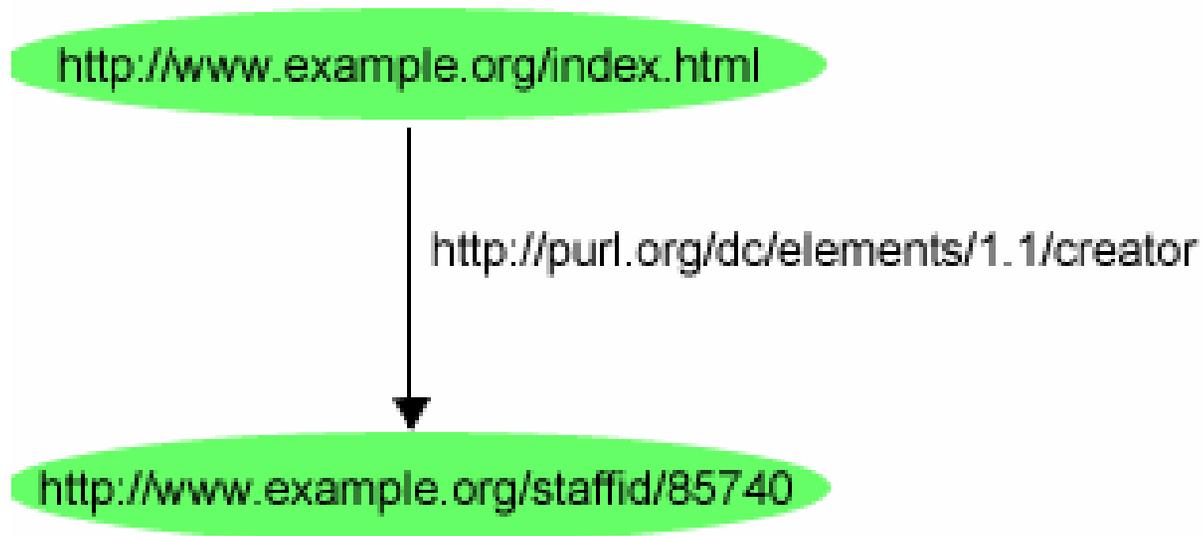
```
- xs:attribute name="Mode">
  - <xs:simpleType>
    - <xs:restriction base="xs:string">
      <xs:enumeration value="Wide-field" />
      <xs:enumeration value="Laser Scanning Microscopy" />
      <xs:enumeration value="Laser Scanning Confocal" />
      <xs:enumeration value="Spinning Disk Confocal" />
      <xs:enumeration value="Slit Scan Confocal" />
      <xs:enumeration value="Multi-Photon Microscopy" />
      <xs:enumeration value="Structured Illumination" />
      <xs:enumeration value="Single Molecule Imaging" />
      <xs:enumeration value="Total Internal Reflection" />
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
```

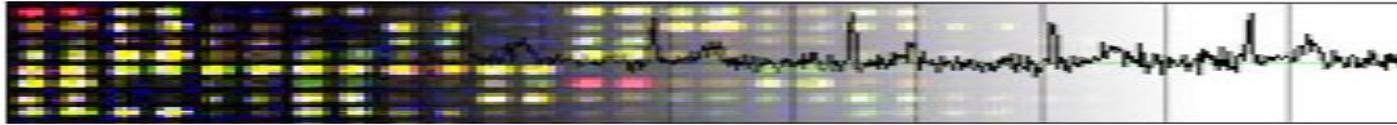
Ref: www.openmicroscopy.org



The role of the Semantic Web

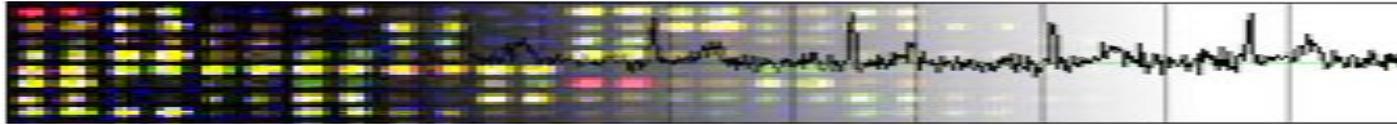
- ❖ LSIDs are absolutely necessary
- ❖ RDF for transport?
 - <http://www.example.org/index.html> has a creator whose value is John Smith
 - the RDF terms for the various parts of the statement are:
 - the *subject* is the URL <http://www.example.org/index.html>
 - the *predicate* is the word "creator"
 - the *object* is the phrase "John Smith"





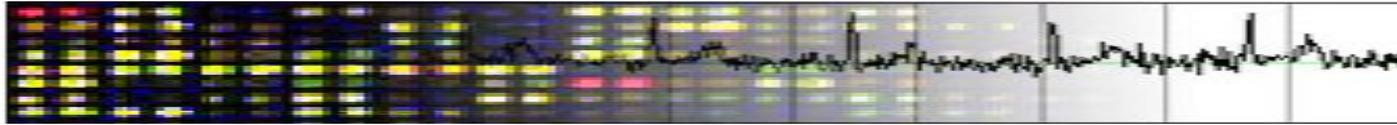
OWL – Web Ontology Language

- *An expressive and uniform way of defining meaning for terms used to transmit data*
- *Can be used for many key purposes*
 - *Guarantee that two definitions are the same*
 - *Discover that two terms are synonymous*
 - *Encode complete object descriptions in RDF*
 - *Define unambiguous database schema*
- *Accessing OWL repositories requires new tools*
 - *Appropriate databases (see <http://www.alphaworks.ibm.com/tech/snobase>)*
 - *Appropriate parsing engines (not Jena)*



Database considerations for biology

- ❖ Strengths of relational databases
- ❖ Weaknesses of the relational model
- ❖ Solutions
- ❖ Database federation

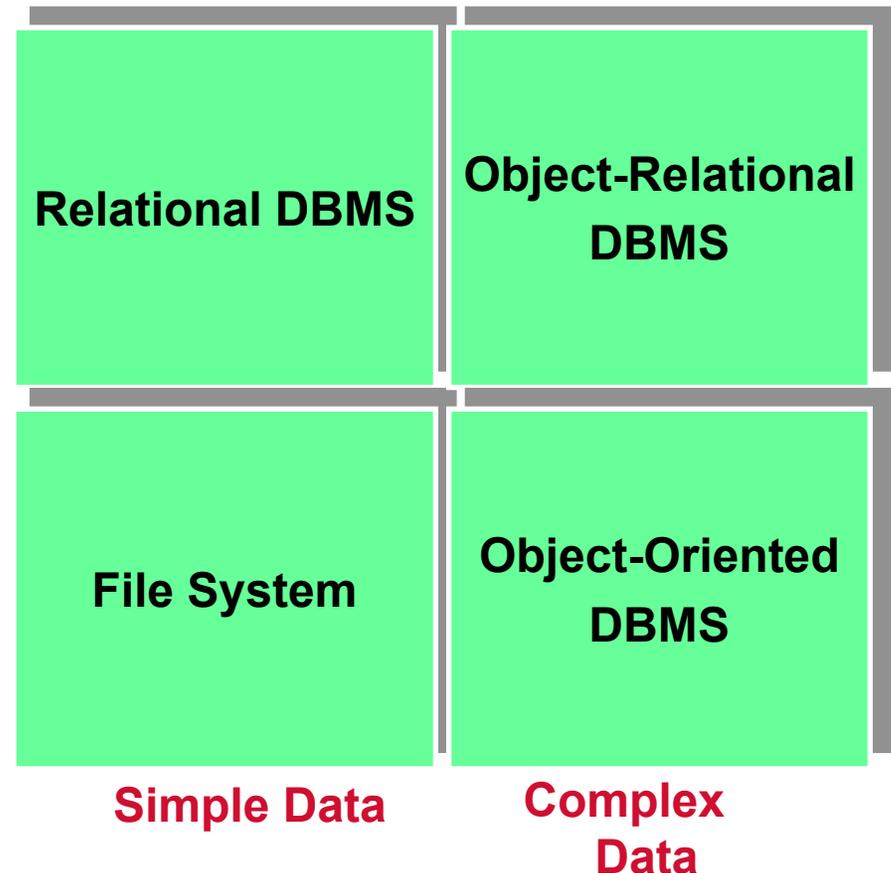


The Object-Relational DBMS

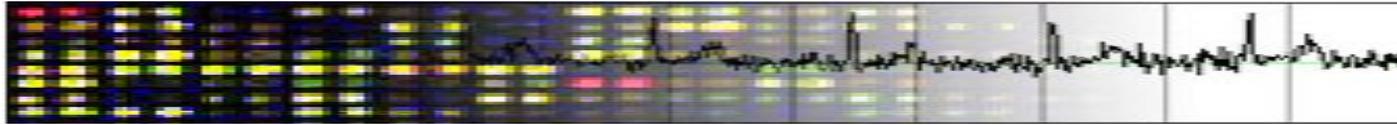
- Supports Queries
- Supports Complex Data
- Supports Standards
 - SQL-3
 - Legacy data
 - Client-server
 - Development tools
- Supports Open Tools
 - ODBC
 - Java JDBC
 - Internet

Query

No Query

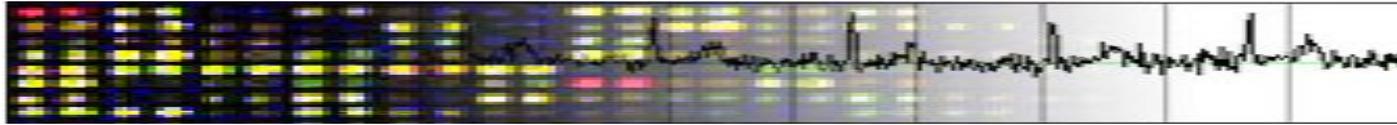


Adapted from *Object-Relational DBMSs: The Next Wave* by Michael Stonebreaker, Morgan Kaufman, Publ., San Francisco, 1996.



Weaknesses of the pure relational model

- ❖ Just tables – poor native support for complex objects or connections between objects
- ❖ SQL is very unfriendly and a limited programming language. No support for “nesting”
- ❖ SQL-3 was designed but only partially implemented
- ❖ Database federation is not a supported concept



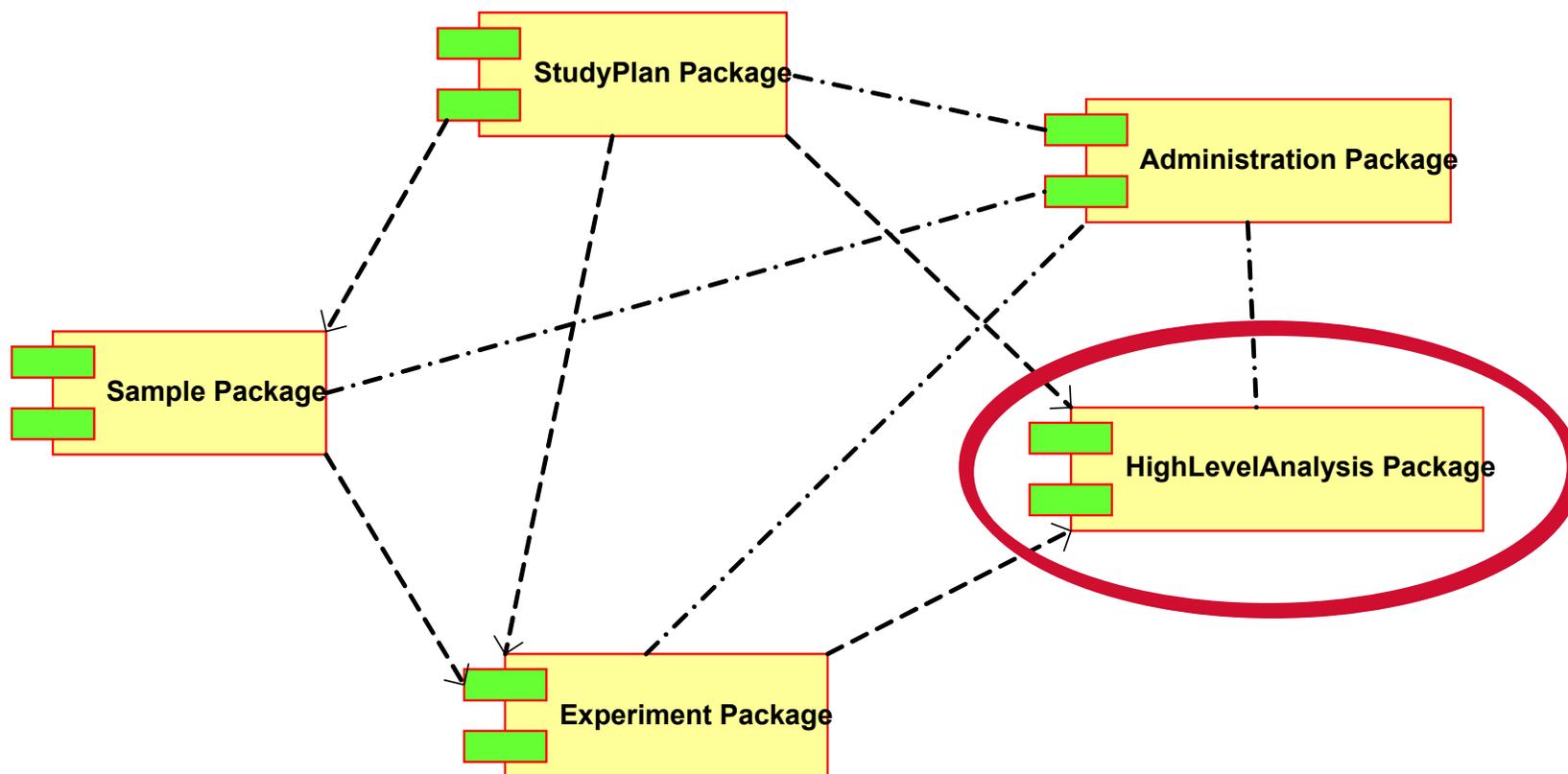
How to design an optimal database language[§]

- ❖ Skeleton: a small number of type constructs
 - Ex: `set(1,2,3)` `bag(0,2,2,2,4)` `list(5,6,6,7)`
- ❖ Operations: constructors and de-constructors
 - Ex: `SetU(x ∈ S); if Pred(x) then set (Exp(x))`
- ❖ **Compositionality**: the meaning of the whole is a function of the meaning of the parts. Parts can be replaced with equivalent ones. (SQL is not compositional.)
- ❖ In silico discovery requires workflows built from transformations and queries.

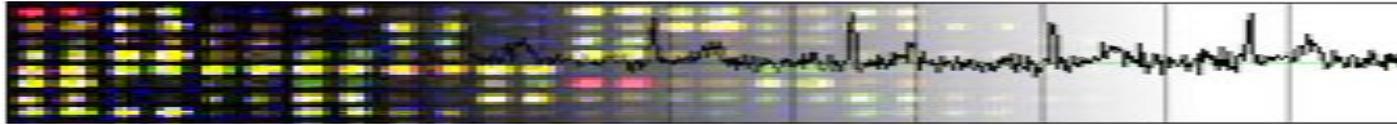
[§] Adapted from *What Makes Bioinformatics Data Special* by Susan Davidson and Val Tannen. GeneticXchange seminar August 8, 2002 (www.geneticxchange.com).



Adding metadata to images and other records

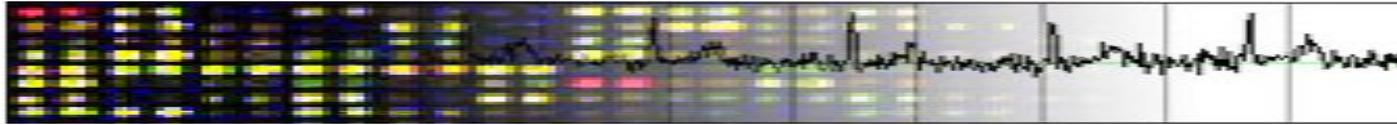


A->B: Dependency. The changes of A can cause changes in B.
-.-: Reference



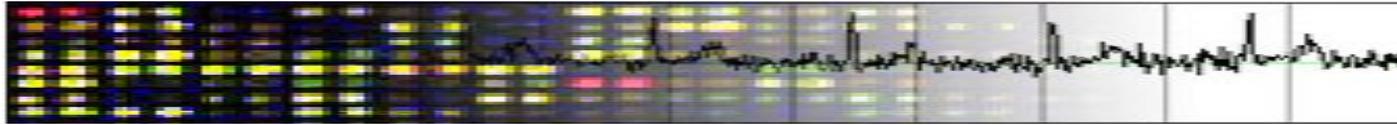
General design rules for metadata

- ❖ Use the “**Study Plan**” and **UIDs** to connect data
 - Each analysis package requires (in general) unique parameters to describe the output
- ❖ Embedded XML “**Style Sheets**”?
 - Can be read by other programs
 - Can be used to generate database schema
- ❖ Choose an appropriate database
 - Strengths of relational databases
 - Weaknesses of the relational model
 - Solutions
 - Database federation
- ❖ Add metadata to images and other records



Conclusions (1)

1. A single database schema specification covering gel electrophoresis, microarrays, microscopy images, mass spec data, and other experimental modalities looks quite feasible.
2. Packages for each experimental method can be constructed from the global schema, with useful commonality and overlap between methods.
3. A key ingredient is the liberal use of LSID identifiers to find common semantics and allow tight definitions of attributes. OWL is the repository for such definitions.
4. A logical object-level schema design will promote a shorter learning time and the development of cooperative software.



Conclusions (2)

5. Particular attention should be given to the support of external analysis packages following the OME model.
6. XML can be used as a transport medium, with semantic information embedded in all appropriate files. Data can be parsed by external programs without explicit knowledge of database itself. RDF may become important.
7. Because individual images are most often parts of larger collections of image objects, study and series information must be embedded into every image information object.
8. Security is included through the Administration metadata package.