Semantic Web and Ontologies

Bioinformatics of Large Systems

Andrew Gibson
a.p.gibson@uva.nl
Universiteit van Amsterdam
16th May 2011
Data Integration & The Web

Motivation

- System-level biomedical research
  - *Understanding* network effects over different levels of biomedical information
Data Integration & The Web

Motivation

- System-level biomedical research
  - Manual searching of the literature?

\[ y = 7.16E+06e^{4.01E-02x} \]
\[ R^2 = 9.99E-01 \]

Lu Z, Database 2011

© Lu Z, 2011. Published by Oxford University Press.
Data Integration & The Web

**Motivation**

- System-level biomedical research
  - ‘Distilled’ biomedical information in **databases**
Molecular Biology Databases

- Nucleic Acids Research – Database Edition

<table>
<thead>
<tr>
<th>Year</th>
<th>No. Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>1330</td>
</tr>
<tr>
<td>2010</td>
<td>1230</td>
</tr>
<tr>
<td>2009</td>
<td>1170</td>
</tr>
<tr>
<td>2008</td>
<td>1078</td>
</tr>
<tr>
<td>2007</td>
<td>968</td>
</tr>
</tbody>
</table>
Data Integration & The Web

Motivation

- **System-level biomedical research**
  - *Data needed from more than one DB to get system-level understanding*
  - *Possible, but very difficult*
Why is data integration so difficult?

1. Different names for biological things
2. Different identifiers
   ▫ How to tell if something is ‘the same’ in two data sources
3. Different data structures (schemas)
   ▫ Describing (properties of) things in different ways
4. Different data formats
5. No direct access to data
   ▫ HTML and downloadable data formats are usually difficult to integrate
   ▫ Queries for data usually defined by data provider in HTML interface

Further Reading:
1. Different names for biological things

- E.g. - Gene & Protein Names
  - Often named by their (presumed) original function or location
  - Renamed over time
    - Sometimes by nomenclature committees
  - Short names are often ambiguous
  - Variations in strings – commas, (brackets) - dashes

Source: Reactome
1. Different names for biological things
   E.g. Metabolites

Source: Human Metabolome Database

More specific compounds, but no IDs for them

Different compound
2. Different Identifiers

Source: Reactome.org
5. No *direct* access to data

- Example: Data in Spreadsheets
- Querying Databases
5. Example: Data in Spreadsheets
5. Example: Data in Spreadsheets

<table>
<thead>
<tr>
<th>Gene</th>
<th>ADF1</th>
<th>ACE2</th>
<th>CIN5</th>
<th>DIG1</th>
<th>FHL1</th>
<th>FK11</th>
<th>FK12</th>
<th>GCN4</th>
<th>GRF10(Pho2)</th>
<th>HIR1</th>
<th>HIR2</th>
<th>HSFI</th>
<th>MDPL</th>
<th>MCM1</th>
<th>MN1</th>
<th>NDD1</th>
<th>NRG1</th>
<th>PHD1</th>
<th>PHO4</th>
<th>RAP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAL024C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YAL026W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YAL029C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YAL040C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YAL052W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YAR019C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YAR099W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBL023C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBL032W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBL033C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBL061W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBL092C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBL093C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR037C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR038W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR068C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR070C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR071W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR077C</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR078W</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR092C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR104W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR109C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR110W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR113C</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR119W</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR157C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR168W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YBR162C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

441 Genes

Identifiers
5. Example: Data in Spreadsheets

| Gene   | ADF1 | ACE2 | CIN5 | DIG1 | FHL1 | FKIR1 | FKIR2 | GCN4 | GRF1(c) | HI1R | HI2R | HSFI1 | MDP1 | MCM1 | MSN4 | NDD1 | NRG1 | PHD1 | PHO4 | RAP1 |
|--------|------|------|------|------|------|-------|-------|------|---------|------|------|-------|------|------|------|------|------|------|------|------|------|
| YAL024C| 1    | 2    | 3    | 4    | 5    | 6     | 7     | 8    | 9       | 10   | 11   | 12    | 13   | 14   | 15   | 16   | 17   | 18   | 19   | 20   | 21   | 22   |
| YAL030W| 0    | 0    | 0    | 0    | 0    | 0     | 1     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YAL026C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YAL040C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YAL082W| 0    | 1    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YAR019C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 1       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YAR059W| 0    | 0    | 0    | 1    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBL023C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBL032W| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBL033C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBL031W| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBL032C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBL033C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR037C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR038W| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR096C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR070C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR079W| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR097C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR078W| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR092C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR104W| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR147C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR139W| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR168W| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR189C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| YBR162C| 0    | 0    | 0    | 0    | 0    | 0     | 0     | 0    | 0       | 0    | 0    | 0     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |

33 Transcription Factors - Names
5. Example: Data in Spreadsheets

ACE2 Regulates Transcription Of YAL062W
5. Example: Data in Spreadsheets

- Questions about the data:
  - Is this data still correct?
    - When was it published / updated?
  - Is the list of genes / TFs complete?
  - What does zero mean?
  - Where did the data come from?
    - Methodologies?
    - Papers?
  - Which of the genes (if any) are known to encode any of the transcription factors?
  - What biological functions are associated with the genes regulated by MBP1?
4. Different Data Formats & 5. Access to Data in Databases

<table>
<thead>
<tr>
<th>Protein/Gene Sequences (in FASTA Format)</th>
<th>Metabolizing Enzyme Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Set</strong></td>
<td><strong>Protein Sequences</strong></td>
</tr>
<tr>
<td>All Metabolite Metabolizing Enzymes</td>
<td>Download</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structures (in SDF Format)</th>
<th>Metabolite Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Set</strong></td>
<td><strong>SDF File</strong></td>
</tr>
<tr>
<td>Metabolite Structures</td>
<td>Download</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Flat Files (in MetaboCard Format)</th>
<th>MetaboCard Flat Files</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Set</strong></td>
<td><strong>MetaboCard File</strong></td>
</tr>
<tr>
<td>All Metabolites</td>
<td>Download (zip file)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spectra</th>
<th>MetaboCard File</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Set</strong></td>
<td><strong>MetaboCard File</strong></td>
</tr>
<tr>
<td>Mass Spectra Image Files</td>
<td>Download (zip file)</td>
</tr>
<tr>
<td>GC/MS Peak Lists</td>
<td>Download (zip file)</td>
</tr>
</tbody>
</table>
Focus On:
Semantic Web
Data Integration Problems
Semantic Web Solutions

• Different identifiers
  ▫ URIs, Linked Data

• Different names
  ▫ Controlled vocabularies

• Different data structures
  ▫ Ontologies, Open Schemas

• Different data formats
  ▫ RDF

• No direct access to data
  ▫ Linked Data, SPARQL

Further Reading:
URIs - Identity on the Semantic Web

- Documents have URLs to locate them on the Web
  - http://www.bdagroup.nl/content/People/andrew_gibson.php
    - A human readable Web page about me
    - HTML processed by your Web Browser

- URIs can also identify ‘real world’ things
  - Separate from:
    - HTML documents about the things
    - Images of the things
    - Data describing the things

▲ http://www.peroxisomekb.nl/v1.6/people/AndrewPGibson

A Uniform Resource Identifier (URI) for me
Data on the Web

This data about me is not *machine readable*

*I.e.* A computer does not know what the web page is about, that the picture is of a person, who my employer is, that it is in Amsterdam, who I work with...

Importantly: I cannot query this data as it is unstructured
Biochemistry of Mammalian Peroxisomes Revisited

Ronald J.A. Wanders and Hans R. Waterham

Department of Clinical Chemistry and Biochemistry, Laboratory Groningen Medisch Centrum, University of Groningen, P.O. Box 30.001, 9700 RB Groningen, The Netherlands, e-mail: r.janders@rug.nl, r.w.waterham@rug.nl

Key Words
fatty acid oxidation, phospholipids, reactive oxygen species, genetic disease

Abstract
In this review, we describe the current state of knowledge about the biochemistry of mammalian peroxisomes. The identification and characterization of yeast mutants defective either in the fragmentation of peroxisomes or in one of its metabolic functions, notably fatty acid beta-oxidation, combined with the recognition of a group of genetic diseases in man, wherein these processes are also defective, have provided new insights in all aspects of peroxisomes. As a result of these and other studies, the independent role of peroxisomes in multiple metabolic pathways has been clarified, and many of the enzymes involved in these pathways have been characterized, purified, and cloned. Our insight into the topography of peroxisomes, which has remained ill-defined, is the transport of metabolites across the peroxisomal membrane. Although it is clear that mammalian peroxisomes under in vivo conditions are closed structures which require the active process of metabolic transporter proteins, much remains to be learned about the permeability properties of mammalian peroxisomes and the role of the four-four ATP-binding cassette (ABC) transporters therein.
**RDF and Statements**

- An RDF **resource** URI of something being described: `http://www.peroxisomekb.nl/v1.6/people/AndrewPGibson`
- An RDF **property resource** URI, used to link resources: `http://xmlns.com/foaf/0.1/img`
- An RDF **resource** URI which, in this case, is the URL of an image: `http://www.bdagroup.nl/images/people/Andrew_Gibson.jpg`
RDF and Statements Terminology

Subject

Predicate

Object

'http://www.peroxisomekb.nl/v1.6/people/AndrewPGibson'

'http://xmlns.com/foaf/0.1/img'

'http://www.bdagroup.nl/images/people/Andrew_Gibson.jpg'

seeAlso - 'http://www.w3.org/TR/rdf-syntax/#basicconcepts'
RDF and Statements
Terminology

http://www.peroxisomekb.nl/v1.6/people/AndrewPGibson

http://xmlns.com/foaf/0.1/page

http://www.bdagroup.nl/content/People/andrew_gibson.php

Same subject

Different predicate

Different object (web page)

seeAlso - http://www.w3.org/TR/rdf-syntax/#basicconcepts
RDF and Statements
For Life Science Too...

http://www.example.org/proteins/ACE2

http://mysite.org/ontology/regulatesTranscriptionOf

http://www.example.org/genes/YAL062W

An RDF triple:
Here URIs refer to biological concepts:
  • A gene
  • A protein,
  • A functional relationship
Graphs of RDF Statements

“Andrew”

http://xmlns.com/foaf/0.1/firstName

http://www.bioexpert.nl/people/febbede/AndrewPGibson

http://xmlns.com/foaf/0.1/knows

http://www.bioexpert.nl/people/ahcvankampen

http://xmlns.com/foaf/0.1/name

“Antoine van Kampen”

http://www.bioexpert.nl/people/febbede/antoine.gif

http://www.bioexpert.nl/people/img/Andrew.jpg

http://www.bioexpert.nl/people/img/antoine.gif
Schemas for the Web Predicates

- These predicates were from the “Friend of a Friend” schema (FOAF):
  - [shorthand](http://xmlns.com/foaf/0.1/img)
  - [http://xmlns.com/foaf/0.1/img](http://xmlns.com/foaf/0.1/img)
  - [http://xmlns.com/foaf/0.1/page](http://xmlns.com/foaf/0.1/page)
  - [http://xmlns.com/foaf/0.1/knows](http://xmlns.com/foaf/0.1/knows)

- Reuse of these properties across datasets makes it far easier to
  - Find and integrate data about people
  - Distinguish data about people from data about anything else

- They can respectively be used in RDF to describe:
  - foaf:img - That a Person has an Image
  - foaf:page - That a resource is described by a Document
  - foaf:knows - That a Person knows another Person

seeAlso - [http://xmlns.com/foaf/spec/](http://xmlns.com/foaf/spec/)
Stating that I am a Person is simple, but very useful because:

1. Some things are only true about people (semantics)
2. Now easy to make RDF queries about People (schema)
Schemas for the Web
RDFS and Lightweight Semantics

• rdfs:domain and rdfs:range
  • foaf:img  rdfs:domain  foaf:Person
  • foaf:img  rdfs:range  foaf:Image

See Also - http://www.w3.org/TR/rdf-schema/#ch_properties
Even More Semantics
Web Ontology Language (OWL)

• A language that extends RDF and RDFS with even more semantics

• Allows ‘richer’ statements to be made in OWL ontologies
  ▫ Specifications of the structure of knowledge

• Supports computational reasoning (logic)
  ▫ over data described by ontologies
    • Infer new statements based on ontology
    • Consistency checking – looking for errors in data

• There are whole tutorials on OWL
  ▫ Not enough time here
Semantic Web:

Linked Data
My background is in biology, and I am now mostly interested in developing emerging technologies from computer science and trying to make them more accessible to scientists, and as such am interested in what myExperiment is doing.
My background is in biology, and I am now mostly interested in developing emerging technologies from computer science and trying to make them more accessible to scientists, and as such am interested in what myExperiment is doing.

Other contact details:
Not specified
Interests:
Bioinformatics
OWL Ontology Development
Semantic Web
e-Science

Field/Industry: Academia
Occupation/Role(s): Postdoctoral Researcher
Organisation(s):
University of Amsterdam

Non-Information Resource URL: http://www.myexperiment.org/users/37

Alternative Formats

What is this?
This XML file does not appear to have any style information associated with it. The document tree is shown below.

```xml
<rdf:RDF>
  <rdf:Description rdf:about="http://www.myexperiment.org/users/37.rdf">
    <foaf:primaryTopic rdf:resource="http://www.myexperiment.org/users/37"/>
    <rdf:Description>
      <mebase:User rdf:about="http://www.myexperiment.org/users/37">
        <foaf:homepage rdf:resource="http://www.myexperiment.org/users/37.html"/>
        <dcterms:hasFormat rdf:resource="http://www.myexperiment.org/users/37.rdf"/>
        <dcterms:hasFormat rdf:resource="http://www.myexperiment.org/users/37.xml"/>
        <sioc:name rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Andrew Gibson</sioc:name>
        <mebase:country rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Netherlands</mebase:country>
        <mebase:country rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Amsterdam</mebase:country>
        <dbpedia:residence rdf:resource="http://dbpedia.org/resource/Amsterdam"/>
        <mebase:email rdf:resource="mailto:gibson@uwa.nl"/>
        <mebase:shasum rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2001d9736373956a61a2e55f95c044c6d5c8c</mebase:shasum>
        <mebase:interests rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Academia</mebase:interests>
      </mebase:User>
    </rdf:Description>
  </rdf:Description>
</rdf:RDF>
```
Linking Data on the Semantic Web

A machine could now:

find my full name from integrating these two statements,

know that both images are of the same person

“Andrew”

foaf:firstName

http://www.peroxisomekb.nl/v1.6/people/AndrewPGibson

http://www.w3.org/2002/07/owl#sameAs

http://www.myexperiment.org/users/37

foaf:surname

“Gibson”

Statements from peroxisomekb.nl

Statements from myexperiment.org
RDF: Data Publishing and RDFization

- Data in databases can stay relational, and also be published as RDF at the same time:
RDF: Data Publishing and RDFization

- Data in databases can stay relational, and also be published as RDF at the same time
- RDF can be published as separate RDF files
- RDF can be embedded in HTML (RDFa)
- Sometimes data needs to be “RDFized”
  - Get access to data from one or more databases
  - Develop an ontology that specifies and unifies the meaning of the data
    - Types of things, relationships between things, metadata
  - Convert the data into RDF and fit it to ontology
  - Store in an RDF ‘triplestore’ (RDF database)
Linked Data Cloud

Break?

• Coming Up
  ▫ Pathway Data and Standards
  ▫ Gene Ontology and Biomedical Vocabularies
  ▫ Large scale RDFization, Integration and Querying of Life Science Data
Focus On:

Pathway Data and Standards
Aspects of Pathway Data
(Think about semantics)

• Metabolic pathways, Signalling pathways

• Different viewpoints
  ▫ Descriptions of pathways (entities involved)
  ▫ Mathematical models (kinetics)

• Pathway Entities (IDs)
  ▫ Pathways, Reactions, Enzymes, Complexes, Locations, Metabolites

• Complex to describe
  ▫ Different representations of how pathway components interact
Pathway Data on the Web
Database Interactions

Network: Pathway Databases

This network shows the links among many databases in Pathguide.

Selecting node(s) shows a summary of database information below the network, with links to database details from Pathguide, and to the database itself.
Pathway Standards

Pathguide: the pathway resource list

Complete Listing of All Pathguide Resources
Pathguide contains information about 325 biological pathway related resources and molecular interaction related resources. Click on a link to go to the resource home page or BioPAX interaction page. Databases that are free and those supporting BioPAX CellML, PSI-MI or SSML standards are respectively indicated.

If you know of a pathway resource that is not listed here, or have other questions or comments, please send us an e-mail.

Protein-Protein Interactions

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Details</th>
<th>Free</th>
<th>PSI-MI</th>
<th>SSML</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D-ID</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADAN</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIFuse</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aMAZE</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APID</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASEdb</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASFID</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aPIID</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BID</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIND</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BioGRID</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRITE</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C41Neuron</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer Cell Map</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellCircuits</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellCircuits</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellCircuits</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellCircuits</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellCircuits</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellCircuits</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellCircuits</td>
<td>Details</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BioPAX
BioPAX Motivation

Pathway Database 1

Internal Pathway Representation (schema)

Entity XREFs (IDs)

Integrated Pathways

BioPAX Export Format

Pathway Database 2

Internal Pathway Representation (schema)

BioPAX Export Format
Pathway Standards
BioPAX - Shared Data Schema

BioPAX

- BioPAX is RDF / OWL based
  - Easier to integrate into RDF based environments
- Detailed model of ‘Pathway Semantics’
- Covers metabolic pathways, genetic interactions, degradation
  - Lots of discussion on best way to approach each issue
- Limited support for mathematical modelling
  - SBML more suited for this
Focus On:

Gene Ontology & Controlled Vocabularies
Gene Ontology Motivation

Database 1 (D. melanogaster)
- Gene
- Protein
- Function: “geranyltranstransferase”

Database 2 (H. sapiens)
- Gene
- Protein
- Function: “farnesyl pyrophosphate synthetase”

Same enzyme activity
Different nomenclature
Gene Ontology Motivation

Database 1 (D. melanogaster)

Gene

Protein

Function

GO:0004337
term - geranyltranstransferase activity
synonym - farnesyl pyrophosphate synthetase activity

Database 2 (H. sapiens)

Gene

Protein

Function

Gene Ontology
Genes involved with fatty acid metabolism?

Database 2 (H. sapiens)

Query Database

Look up GO term

GO:0006631: fatty acid metabolic process

Gene Ontology

Gene

Protein

Function
Gene Ontology Content

- **Controlled vocabularies describing**
  - **Cellular Components**
    - Where a gene product may be localized in the cell
    - Cellular structures, also functional complexes (e.g. ‘ribosome’)
    - 2831 defined
  - **Biological Processes**
    - A series of events accomplished by one or more ordered assemblies of molecular functions.
    - E.g. ‘fatty acid metabolism’, ‘signal transduction’
    - 20,753 defined
  - **Molecular Functions**
    - Elemental activities, such as catalysis or binding, describing the actions of a gene product at the molecular level.
    - Catalytic activities, transporter activities, binding
    - 9010 defined
Formats for GO

- ‘Native’ format is OBO – a flat file format:
  - Requires parsing for integration

```
[Term] id: GO:0000016
name: lactase activity
namespace: molecular_function
def: "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose."
[EC:3.2.1.108]
synonym: "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
synonym: "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
xref: EC:3.2.1.108
xref: MetaCyc:LACTASE-RXN
xref: Reactome:20536 "lactase activity"
is_a: GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds
```

- GO is also available as an OWL ontology
  - Easier to integrate into RDF applications
GO Structure

- Terms are related in a **hierarchical** tree
- Relationships include
  - is a
  - part of
Applying GO - Annotations

• **Thing being annotated**
  ▫ A gene OR gene product

• **The annotation term**
  ▫ a **GO term**

• **The date** the annotation was assigned

• **An evidence code**
  ▫ Each annotation must include an evidence code to indicate how the annotation to a particular term is supported

• **A reference**
  ▫ Data or statements which support the annotation
  ▫ Description of the method by which the annotation was assigned
## GO Annotations

### UniProt Example

<table>
<thead>
<tr>
<th>Gene Ontology (GO)</th>
<th>Annotation</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biological process</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>branching involved in ureteric bud morphogenesis</td>
<td>inferred from sequence or structural similarity</td>
<td>UniProtKB</td>
</tr>
<tr>
<td>cell cycle arrest</td>
<td>traceable author statement</td>
<td>ProtInc</td>
</tr>
<tr>
<td>cell proliferation</td>
<td>traceable author statement</td>
<td>ProtInc</td>
</tr>
<tr>
<td>cellular ion homeostasis</td>
<td>traceable author statement</td>
<td>ProtInc</td>
</tr>
<tr>
<td>positive regulation of mesenchymal cap mesenchymal cell proliferation</td>
<td>inferred from sequence or structural similarity</td>
<td>UniProtKB</td>
</tr>
<tr>
<td>regulation of telomere maintenance</td>
<td>inferred from mutant phenotype</td>
<td>BHF-UCL</td>
</tr>
<tr>
<td>regulation of transcription from RNA polymerase II promoter</td>
<td>traceable author statement</td>
<td>ProtInc</td>
</tr>
<tr>
<td>response to drug</td>
<td>inferred from expression pattern</td>
<td>UniProtKB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Cellular component</strong></th>
<th>Annotation</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>nucleolus</td>
<td>inferred from direct assay</td>
<td>UniProtKB</td>
</tr>
<tr>
<td>nucleoplasm</td>
<td>inferred from direct assay</td>
<td>UniProtKB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Molecular function</strong></th>
<th>Annotation</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-box binding</td>
<td>inferred from direct assay</td>
<td>UniProtKB</td>
</tr>
<tr>
<td>protein binding</td>
<td>inferred from physical interaction</td>
<td>UniProtKB</td>
</tr>
<tr>
<td>sequence-specific DNA binding</td>
<td>transcription factor activity</td>
<td>traceable author statement</td>
</tr>
</tbody>
</table>
Evidence Codes Ontology

- Terms to describe the way a GO annotation was derived
  - Mostly assigned by curators

- Experimental
  - Annotations for particular laboratory techniques
  - Usually high quality

- Computational
  - IEA – Inferred from electronic annotation
    - Not reviewed by a curator! Treat with caution

- Author Statement
  - Traceable – read in a paper with a reference
  - Non-traceable – read in a paper, no reference

- No Evidence
  - Inferred by the curator with no evidence

Evidence Codes:

UniProt Example

TIP: ALWAYS look at the evidence
Gene Ontology in Data Analysis

- Transcriptomics
  - Emerging around the same time as GO

- Use of Gene Ontology
  - Interpretation and Visualization (see left)
  - Many tools exist to support this process

- Proteomics data analysis
  - More recent use of GO

Other Biomedical Vocabularies Worth Mentioning

• **Medical Subject Headings (MeSH)**
  - PubMed citations – annotation and search!

• **Unified Medical Language System (UMLS)**
  - Integrates >100 different biomedical vocabularies

• **SNOMED CT** *(Systematized Nomenclature of Medicine Clinical Terms)*
  - “Clinical terminology with comprehensive scientifically validated content”
  - 200,000 concepts

• **National Cancer Institute (NCI) Thesaurus**
  - 61,000 classes
  - “The NCI Thesaurus is an ontology-like vocabulary to meet the needs of the cancer research community, which provides unambiguous codes and definitions for concepts used in cancer research”
Even More Biomedical Vocabularies

‘BioPortal’

- Collection of 267 biomedical controlled vocabularies
- Over 5 million biomedical terms
Focus On:

Large Scale RDFization, Integration and Querying
Bio2RDF.org

Semantic web atlas of postgenomic knowledge

Send your comments to the project team and see the Bio2RDF wiki for instructions on the download and use of this software. You can find a demonstration set of Bio2RDF URIs here.

Bio2RDF.org Blog Feed

Bio2RDF return to Japan
by Marc-Alexandre Nolin - Oct 08, 2010
Bio2RDF is returning in Japan again this year. We will give a talk about Bio2RDF at Biocuration 2010. Biocuration is from ...

Biocuration 2010 presentation
Bio2RDF return to Japan
Video on Bio2RDF and Beyond! Large Scale, Distributed Biological Data Integration
Bio2RDF Cognoscope presentation at BioHackathon 2010 in Tokyo

Home
Last modified 20-April-2009
Data Warehousing - Semantic Web Style
Linked Life Data

- **LinkedLifeData - statistics:**
  - Number of RDF statements: 4,193,400,044
  - Number of explicit RDF statements: 2,741,704,569
  - Number of entities (resources): 582,691,283
  - 20+ data sources

- **Platform to automate the process:**
  - Infrastructure to store and inferences
  - Transform the structured data sources to RDF
  - Provide web interface to access the data

- **Publicly available at:**
  [http://www.linkedlifedata.com](http://www.linkedlifedata.com)
<table>
<thead>
<tr>
<th>Database</th>
<th>Dataset</th>
<th>Schema</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniprot</td>
<td>Curated entries</td>
<td>Original by the provider</td>
<td>Protein sequences and annotations</td>
</tr>
<tr>
<td>Entrez-Gene</td>
<td>Complete</td>
<td>Custom RDF schema</td>
<td>Genes and annotation</td>
</tr>
<tr>
<td>iProClass</td>
<td>Complete</td>
<td>Custom RDF schema</td>
<td>Protein cross-references</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>Complete</td>
<td>Schema by the provider</td>
<td>Gene and gene product annotation thesaurus</td>
</tr>
<tr>
<td>BioGRID</td>
<td>Complete</td>
<td>BioPAX 2.0 (custom generated)</td>
<td>Protein interactions extracted from the literature</td>
</tr>
<tr>
<td>NCI - Pathway Interaction Database</td>
<td>Complete</td>
<td>BioPAX 2.0 (original by the provider)</td>
<td>Human pathway interaction database</td>
</tr>
<tr>
<td>The Cancer Cell Map</td>
<td>Complete</td>
<td>BioPAX 2.0 (original by the provider)</td>
<td>Cancer pathways database</td>
</tr>
<tr>
<td>Reactome</td>
<td>Complete</td>
<td>BioPAX 2.0 (original by the provider)</td>
<td>Human pathways and interactions</td>
</tr>
<tr>
<td>BioCarta</td>
<td>Complete</td>
<td>BioPAX 2.0 (original by the provider)</td>
<td>Pathway database</td>
</tr>
<tr>
<td>KEGG</td>
<td>Complete</td>
<td>BioPAX 1.0 (original by the provider)</td>
<td>Molecular Interaction</td>
</tr>
<tr>
<td>BioCyc</td>
<td>Complete</td>
<td>BioPAX 1.0 (original by the provider)</td>
<td>Pathway database</td>
</tr>
<tr>
<td>NCBI Taxonomy</td>
<td>Complete</td>
<td>Custom RDF schema</td>
<td>Organisms</td>
</tr>
</tbody>
</table>
Query Language for RDF
SPARQL and Open Data

SPARQL Query

Select all proteins that are linked to a curated molecular interaction, to inflammatory response and to a target of existing drug.
Challenges for the Semantic Web and Life Science Data

- Extracting data from
  - Papers
  - Files
  - Databases

- Interfaces and tools for
  - Generation and publication of RDF
  - Use of data from Semantic Web

- Generating more ontologies that describe
  - Things in biology (vocabularies)
  - Descriptions and relationships

- Increasing metadata about
  - Where data came from
  - Who authored particular data and when
  - What sort of evidence supports a particular statement
Summary

• Biomedical data is complex and widely distributed

• Currently, data is difficult to integrate, but strategies exist:
  ▫ Exchange formats, controlled vocabularies

• The Semantic Web offers new approaches to publishing and integrating data
Questions?