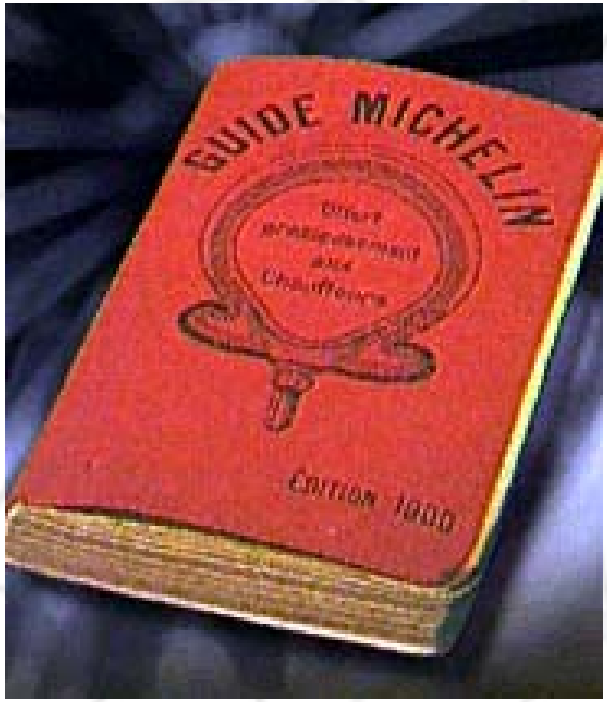
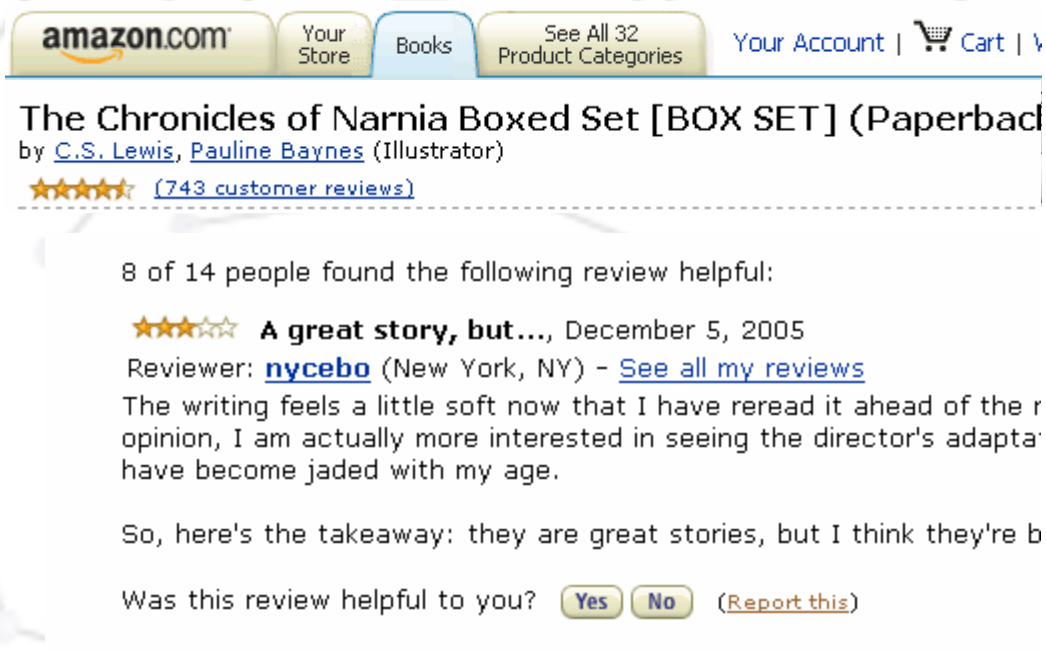
NCI Proteomics Reagents workshop

Building a Database





Henning Hermjakob, European Bioinformatics Institute

Overview

- **Use cases:**
  ***Who* are the users and *What* do they want?**
- **Data:**
  ***Which* data elements should be stored?**
- **Implementation:**
  ***How* could it be approached?**

Use case 1: Simple lookup

- Scientist looks up the available antibodies for his favorite protein.
  - Anonymous access
  - HTML web interface

Use case 2: Feedback

- Scientist used the antibody in his study and reports success, supported by a journal citation
  - Login required
  - Data input required
  - Data entered can be complex
    - Pubmed id
    - Experiment classification from controlled vocabulary
    - Comment

Use case 3: Target request

- Scientist didn’t find a validated AB for his protein, requests addition to the “targets” list
  - Login
  - Data input
  - Anonymity of request?!
  - Potential extension of targets list
  - Prioritisation algorithm for target requests from many sources required

Use case 4: Curator approval

- In principle, a website operator is responsible for the contents, even for contributions from third parties
  - Curator required
    - All needs to be approved, or
    - Only obviously wrong things will be removed

Use case 5: Curator contribution

- Not all relevant data will be entered by third parties.
- Example: Polanski paper: 1200 candidates
- Requires curator
- Requires curation guidelines

Use case 6: Large scale lookup I

- Lab for large scale protein interaction studies, or Pat Brown for his prototype AB array, want the available 4 star antibodies for a list of 500 proteins. They want the updated status every month.
- Ideally with a price tag!
- Requires automated large-scale read access
- Anonymous access
- In fact, anonymity is a key requirement! Pat might not want others to know what will be on his array.

Use case 7: Large scale lookup II

- NCI contract validation lab looks up next batch of antibodies to validate
- Needs potential antigen sources
- “Status flag” necessary to mark ABs which are in the validation pipeline

Use case 8: Large scale target list

- 200 new promising targets from the microarray study of lab X, with associated publication
- Lab X will update the list every quarter or so over the next two years

- NCI grantee Y requests 200 new targets from the apoptosis pathway

Use case 9: Large scale validation

- Pat reports the success of his AB array:
  - 1 technology description
  - 1 publication
  - 500 data points
- He probably doesn't want to type it all, but send it in a defined format
- Same action, different data next month
- "Pat" might also be NCI contract lab for validation
- Update problems: Between Pat's data retrieval and the publication of results, 2 ABs have gone off the market, and 4 protein sequences have been updated in UniProt

Use case 10: International collaboration

- Mathias Uhlen provides validation information for a few thousand ABs, monthly updated, with extensive documentation in the Protein Atlas
- The Chinese Liver Proteome Project AB database provides validation for 400 ABs against existing targets, plus validation for 200 ABs against targets not in the targets list, again with detailed information in the local DB.
- The European CaBinders project provides validation for 200 ABs based on Affybodies, a technology not yet on the list of technologies.

Data:
Which data elements should be stored?

Targets

- Defined with respect to a well-known resource, e.g. UniProt
- Sequence, Accession number, and Version are essential!
- Species attribution
- Target ranking, and associated evidence, some of it potentially confidential
- Antigen source

Antibodies

- Name
- Source
- Price
- Quantity
- Mono/polyclonal
- System, e.g. mouse, rabbit, ..
- ….
- Look in Biocompare.com for more attributes

Technologies

- Technology description
- Method publication
- Application publication

- Technical: Necessary data items for validation with this technology

Antibody validation

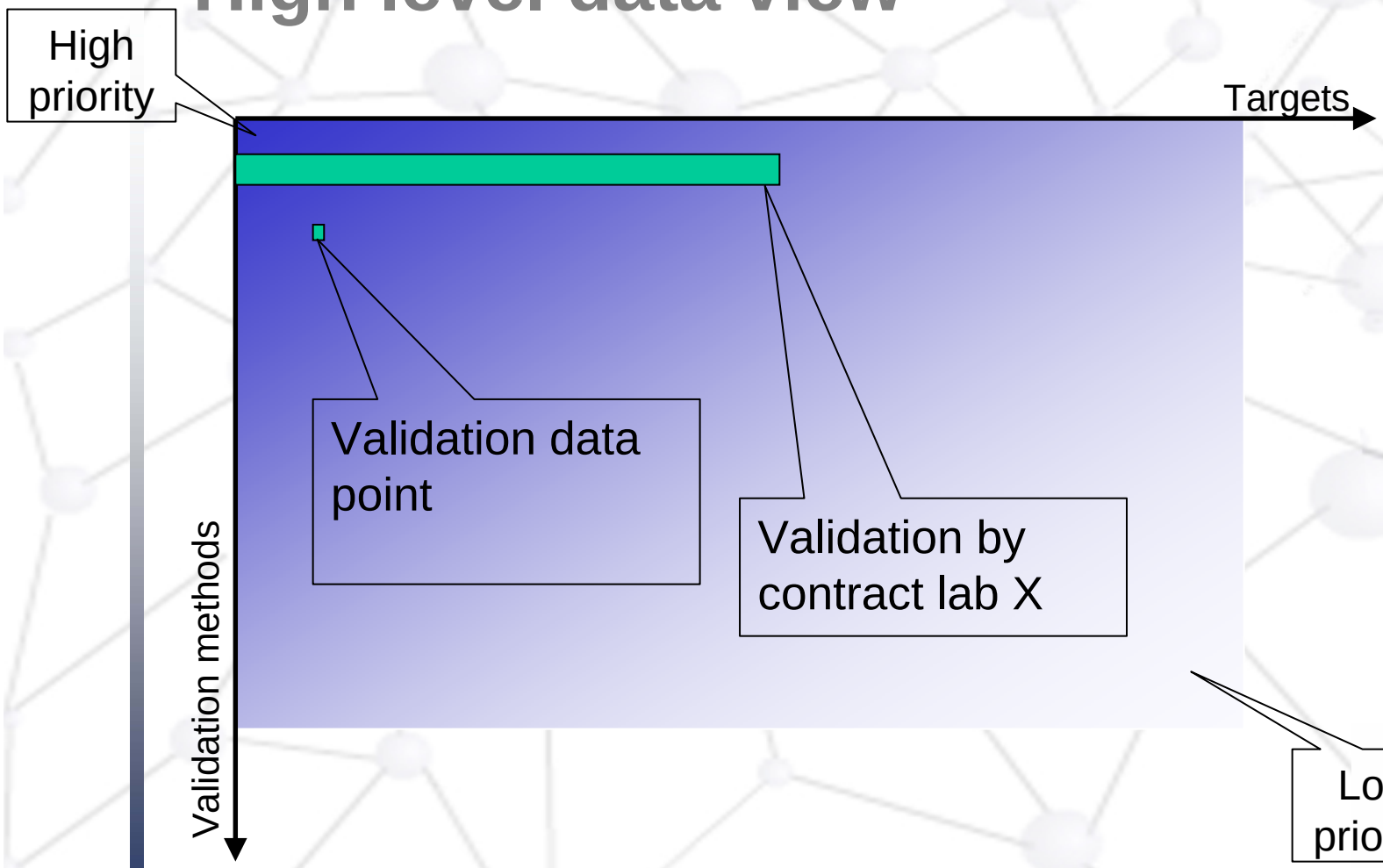
- Relevant antigen (sequence) features, e.g. PrESTs
- Technology
- Lab
- Score
- Supporting evidence, potentially external, e.g. Western images
- Quantitation
- Technology-dependent parameters

- Many validations for one AB/antigen pair
- Relationship between AB, antigen, technology
- More than one AB in a validation assay, e.g. sandwich assays

Implementation

How could it be approached?

High level data view



Access control

- Read-only access
- Write access by different entities:
  - Anonymous opinion statement?
  - Registered external user
    - Anonymous target request
  - Internal curator
  - International collaborators

- Detailed access control for individuals and groups necessary

Automated access

- Automated, repetitive access to large data sections
- Complex queries (“Cost for my 500 target ABs?”)
- Read and write access

- XML interface
- Web services interface (SOAP style)?

- Full database download

Data exchange format

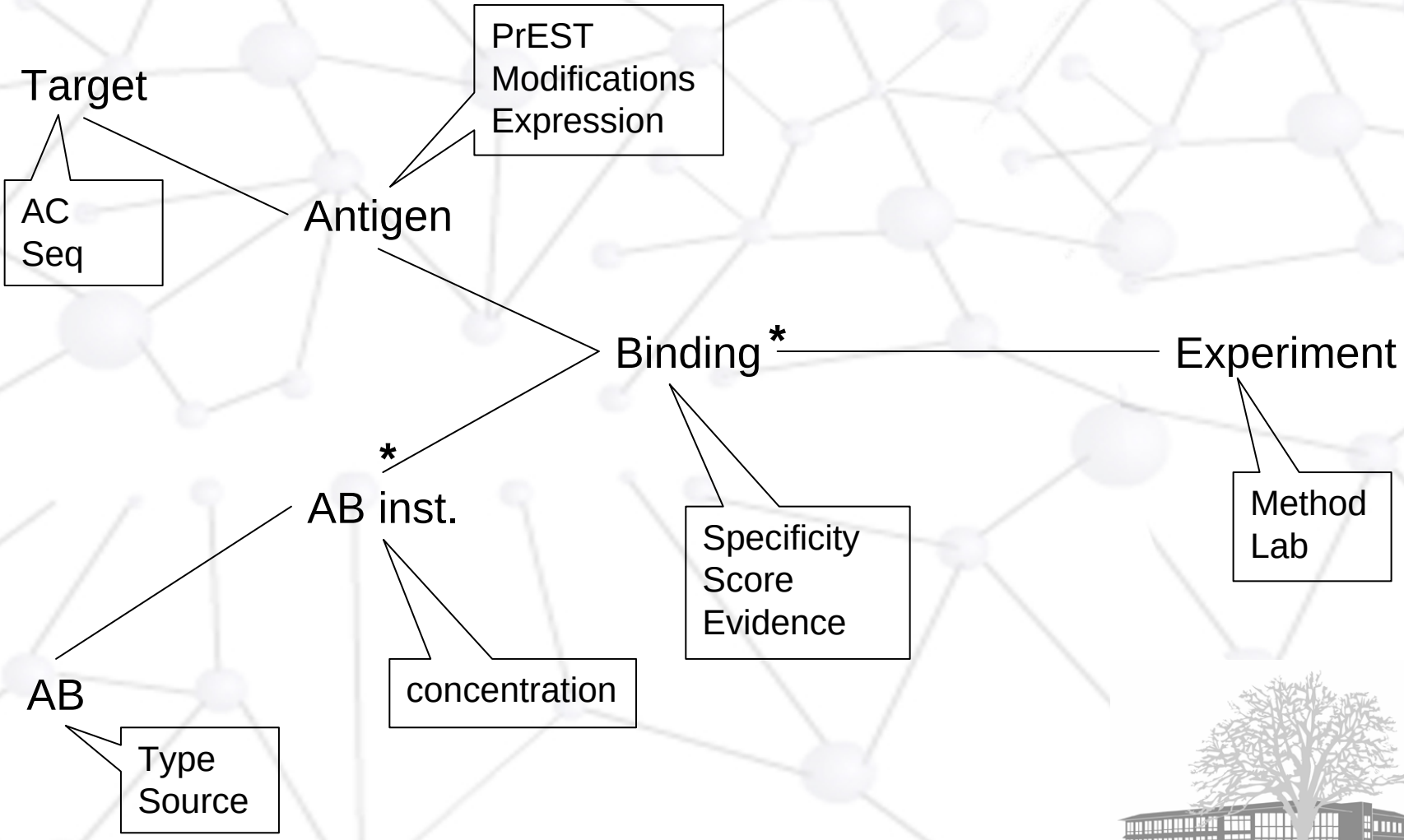
- **XML based format for large scale data exchange**

- ***Why reinvent the wheel when you can just polish the spokes***

OR

- ***If you have a hammer, everything looks like a nail***

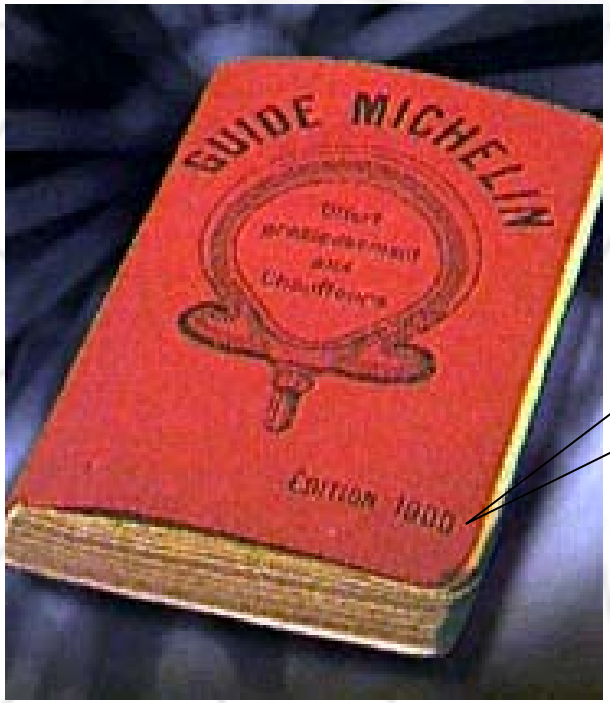
Possible data exchange format structure



Proposal: HUPO PSI-MI XML format

- **Community standard for Molecular Interactions**
- **XML schema and detailed controlled vocabularies**
- ***Exchange* format, not internal format**
- **Jointly developed by major data providers:**
  BIND, CellZome, DIP, GSK, HPRD, Hybrigenics, IntAct, MINT, MIPS, Serono, and others
- **Version 1.0 published in February 2004**
  The HUPO PSI Molecular Interaction Format - A community standard for the representation of protein interaction data.
  Hermjakob et al, Nature Biotechnology 2004, 22, 176-183.
- **Version 2.5 (Released December 5, 2005)**
  covers flexible molecule types, and parameters
- **http://psidev.sf.net/mi/rel25**

Sustainability



"Le Guide" has been around for 105 years …

Thank you!