IFR 27 - DRDC

Département Réponse et Dynamique Cellulaires

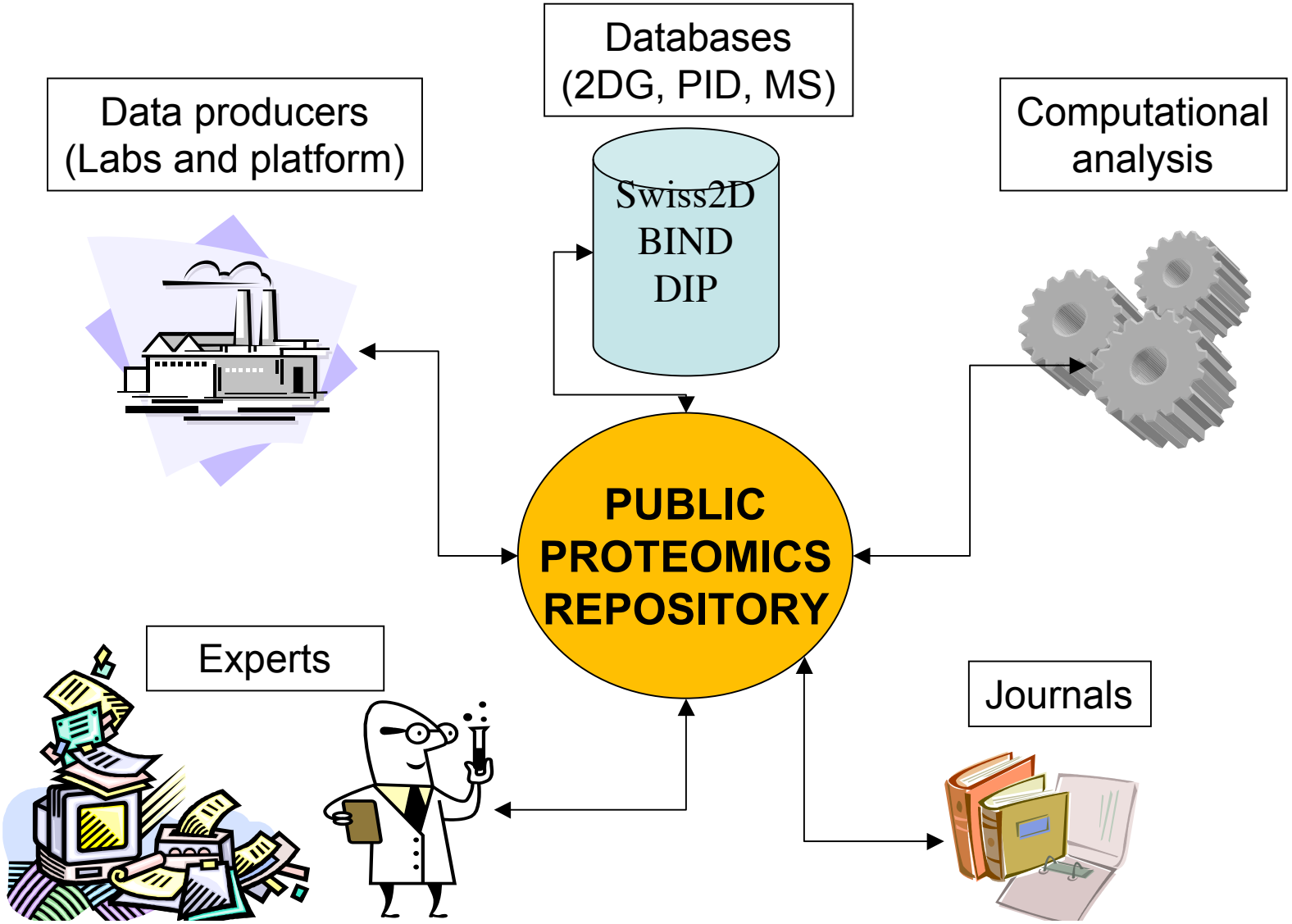
DATA MODELS AND EXCHANGE FORMAT FOR PROTEOMICS EXPERIMENTS

Quelle informatique et bioinformatique pour la protéomique?

1er juin 2004 - INRIA

yves.vandenbrouck@cea.fr

PROTEOMICS : THE PLAYERS



What about (raw or analyzed) data and metadata availability?

PROTEOMICS : THE STATE OF PLAY

The volume of generated proteome data is rapidly increasing

- Movement towards high–throughput approaches
- New experimental techniques and analyses (DiGE, ICAT, etc.)

Publicly available proteomics data is rather limited

- 2D–Gel image databases (e.g. SWISS–2DPAGE) contain little information about sample preparation, or analysis of results
- No widely used databases of mass spectrometry data or analyses
- Data fragmentation (not synchronized) and data formats incompatibility (PID)
- Incomplete data and noisy
- Need for cross-validation and comparative analysis

A robust, future proofed, standard representation of both methods and data from proteomics experiments is required (HUPO-PSI, Hermjakob and Apweiler, 2002)

THE PROTEOMICS STANDARD INITIATIVE (PSI-HUPO)

- Started october 2002

- Develop data format standards

- Data representation and annotations standards
  CVs, ontologies, standard reference sets

- Involve data producers, database providers, software designers, publishers

MIAPE : The Minimum Information About Proteomics Experiment
Analogous to MIAME (MGED)

PROTEOMICS DATA STANDARDS : STAKES AND BENEFITS

- **Users will know what to expect from datasets (formats etc.)**
- **Facilitate handling, exchange and dissemination of data**
- **Guide the development of effective search/analysis tools (Sequest, Mascot, ProteinLynx)**
- **Experiments comparison and data mining**
- **Journal encourage to support publications with MIAPE-compliant data (Analogous to the MIAME guidelines for transcriptomics)**
- **Federative boost for the community**

There are now several formal attempts at standardisation

GAML, HUP-ML, PEDRo, ProteinLynx, Sashimi, SpectroML, others…

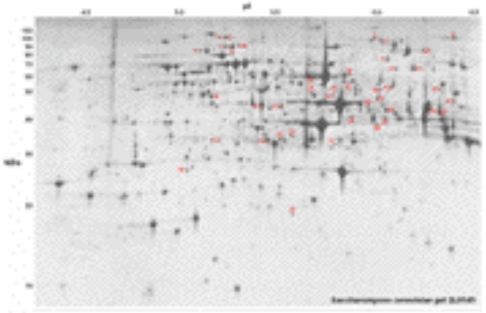
WHAT KIND OF DATA SHOULD BE STORED?

Sample generation

- Origin of sample
  - hypothesis, organism, environment, preparation, paper citations

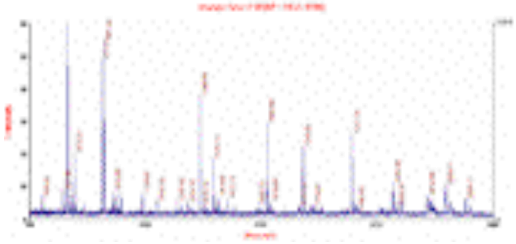
Sample processing

- Gels (1D/2D), columns, other methods
  - images, gel type and ranges, band/spot coordinates
  - stationary and mobile phases, flow rate, temperature, fraction details



Mass Spectrometry

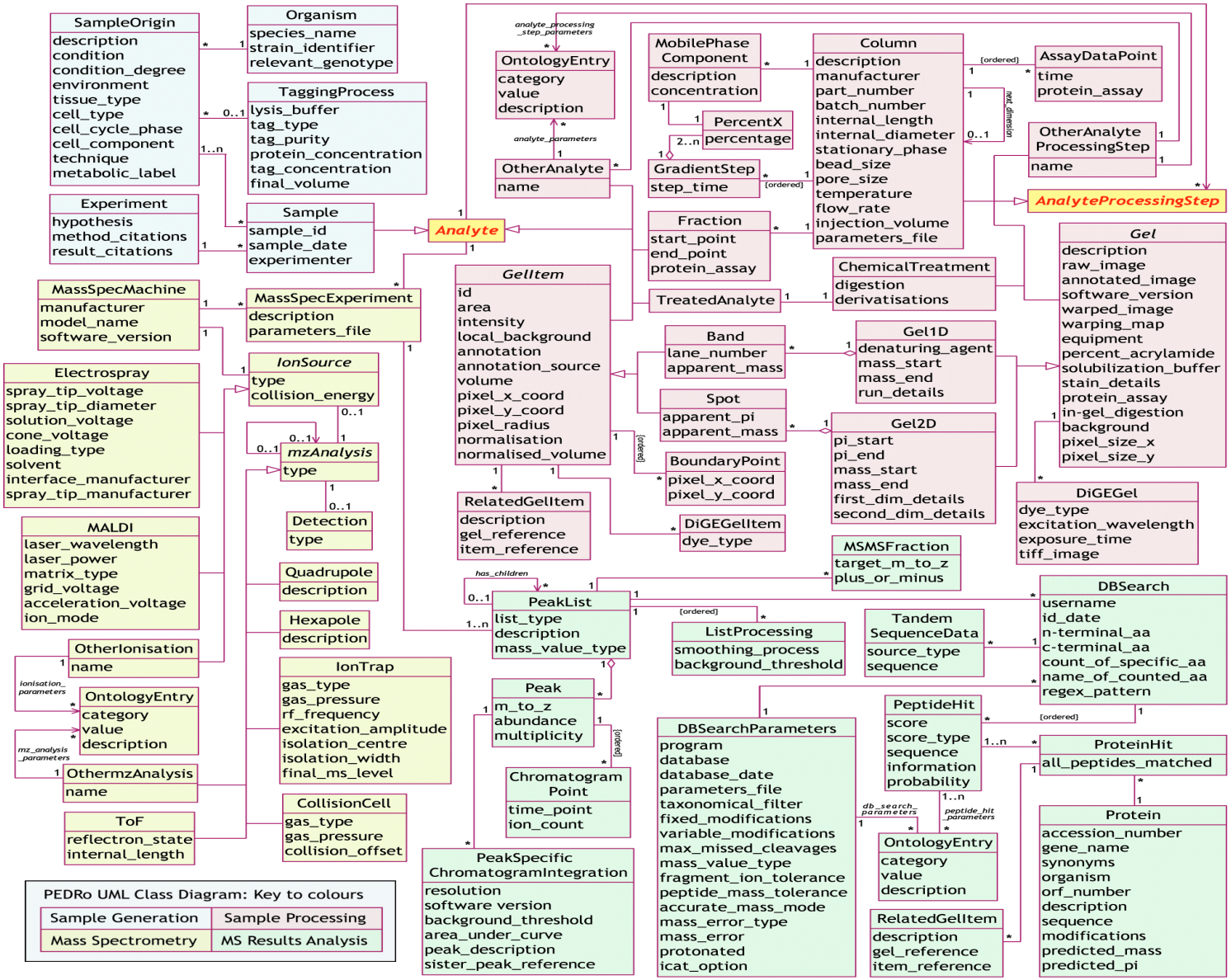
- machine type, ion source, voltages



In Silico analysis

- peak lists, database name + version, partial sequence, search parameters, search hits, accession numbers

THE PEDRO OBJECT MODEL

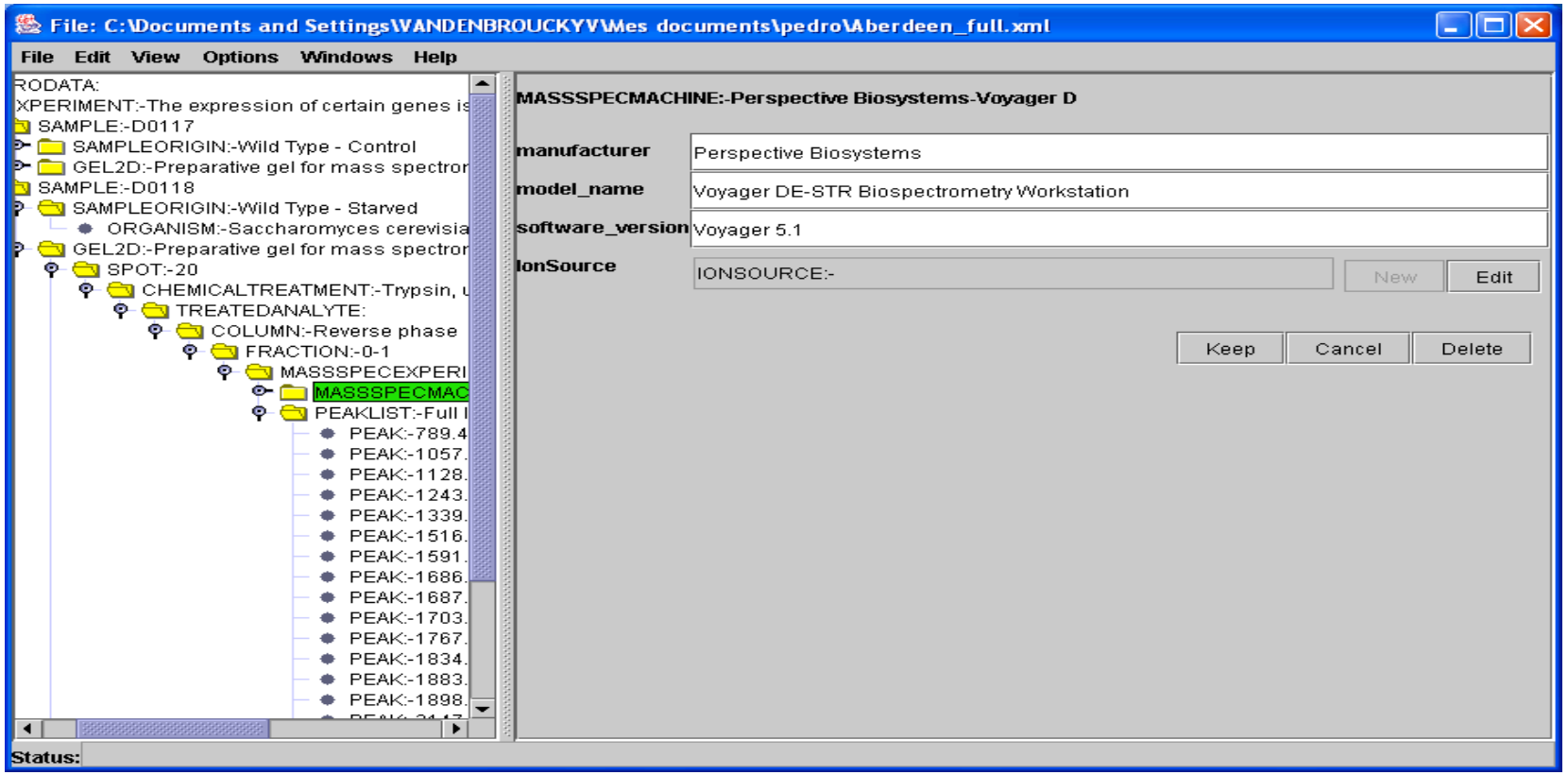


PEDRO : DERIVED IMPLEMENTATION STRUCTURES & TOOL

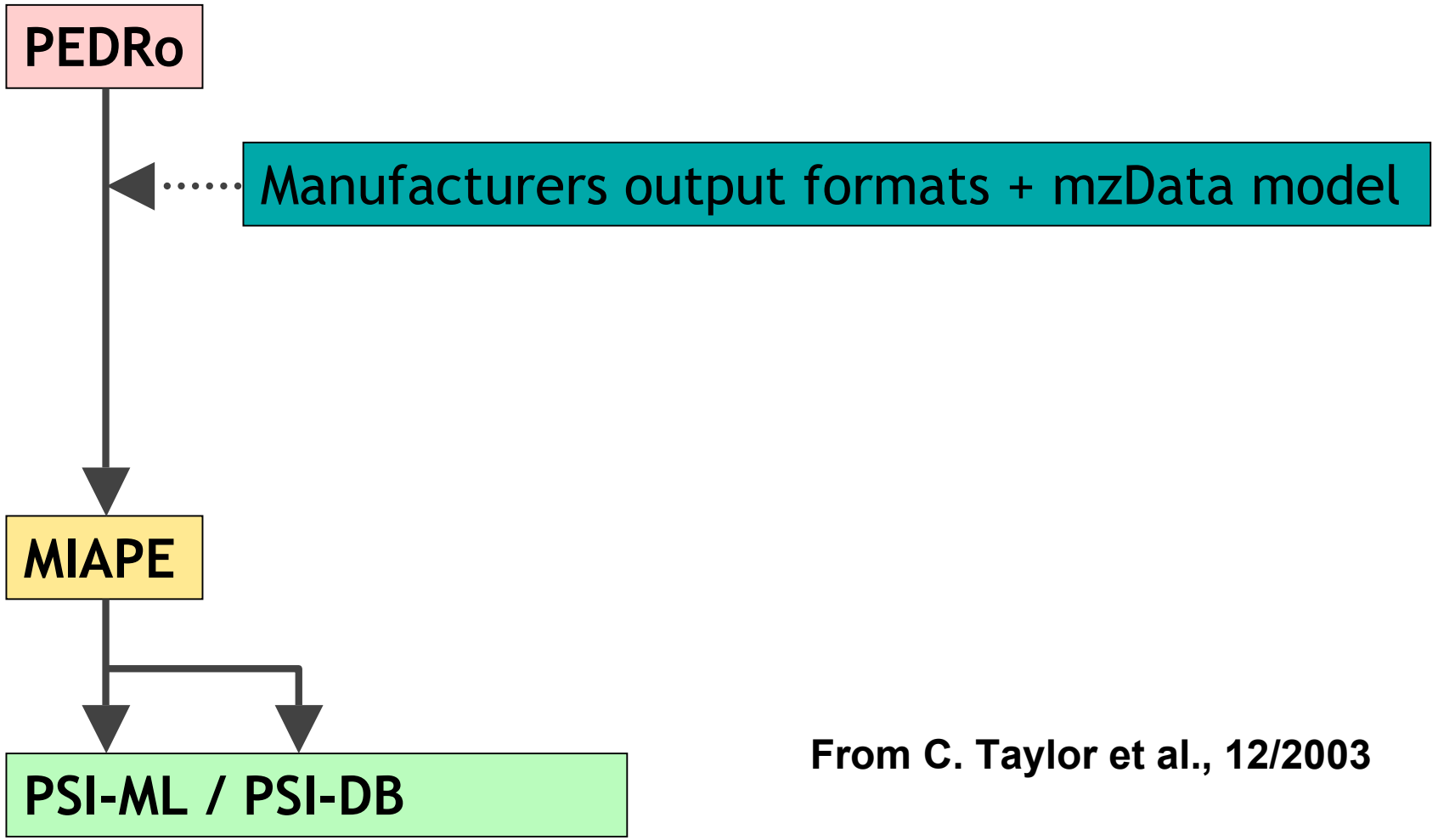
•XML Schema (PEML)

XML : a technology for data description and structuration (tree) (!= HTML)

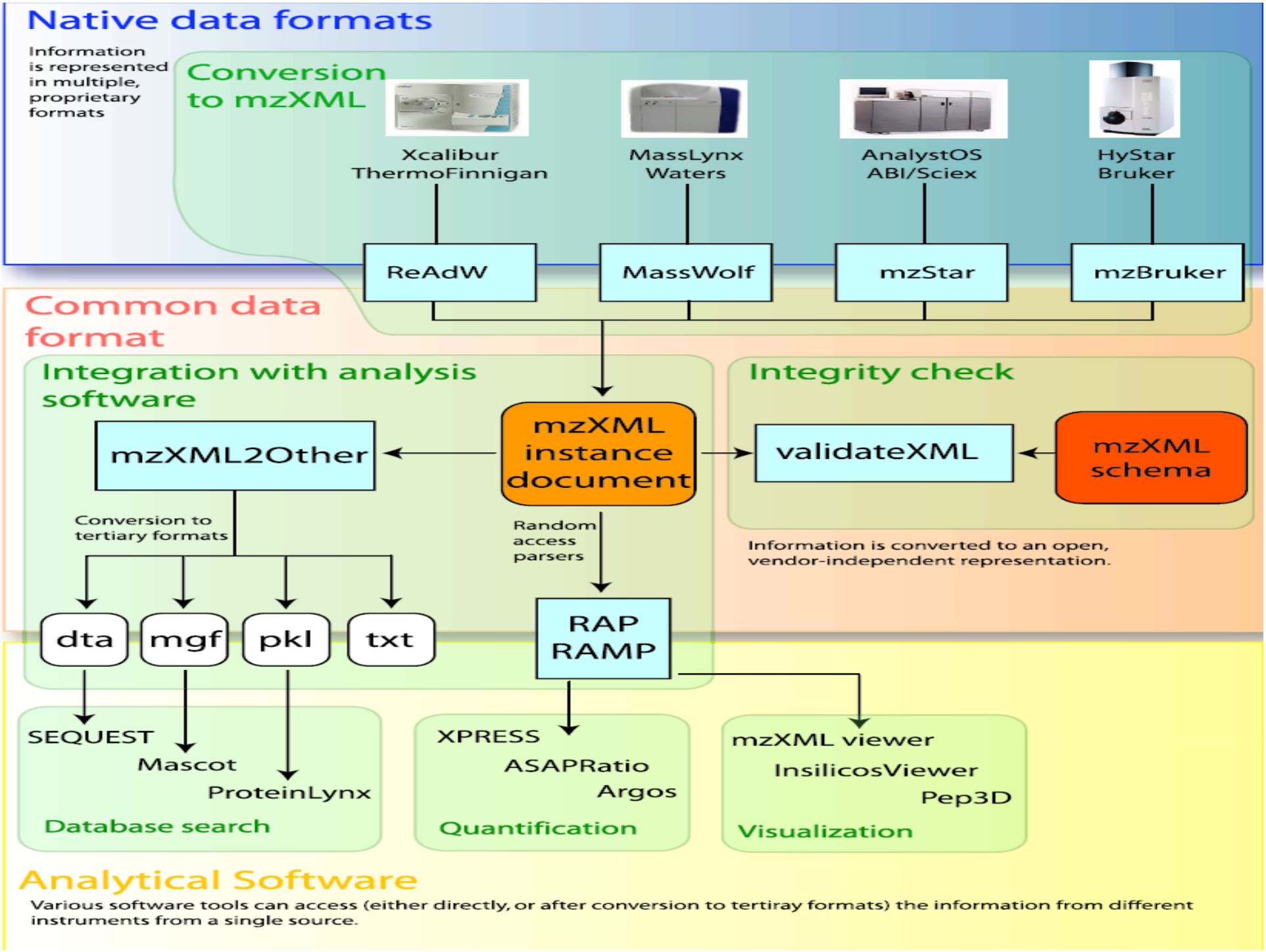
•RDBMS schema (SQL)

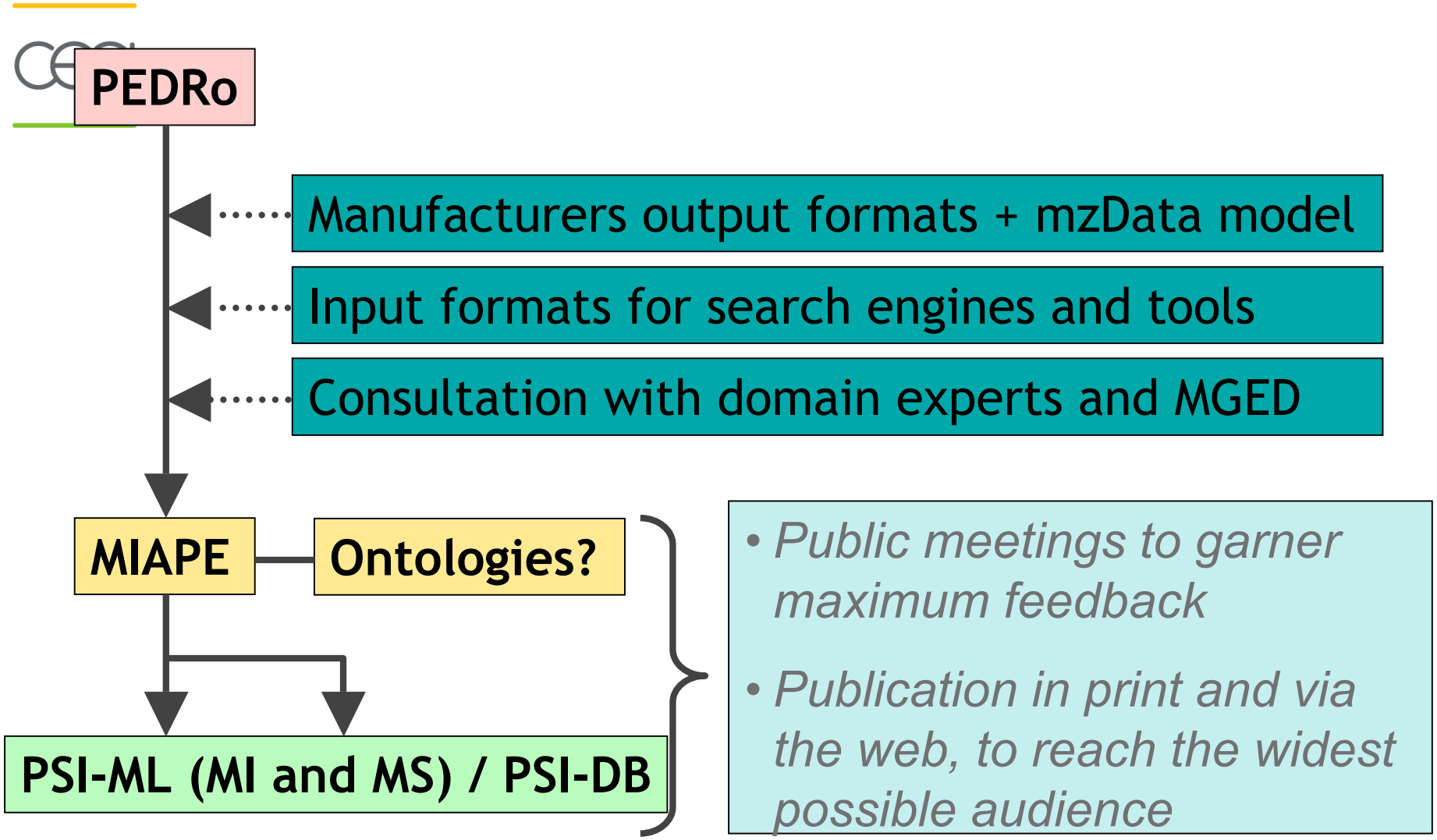


PEDRo MIAPE / PSI-ML / PSI-DB



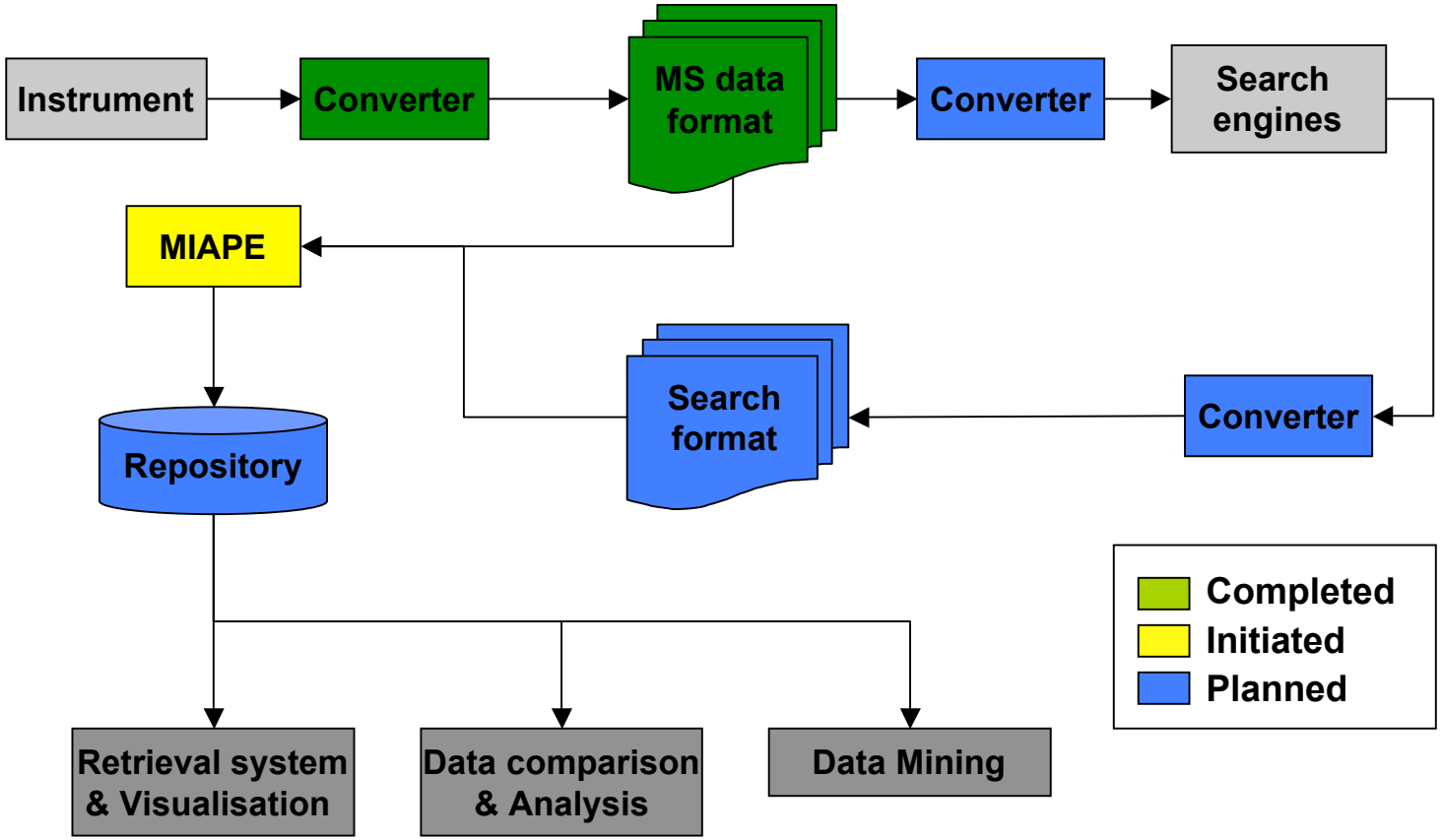
From C. Taylor et al., 12/2003


Native data formats
Information is represented in multiple, proprietary formats
Conversion to mzXML
Xcalibur ThermoFinnigan
MassLynx Waters
AnalystOS ABI/Sciex
HyStar Bruker
ReAdW
MassWolf
mzStar
mzBruker
Common data format
Integration with analysis software
mzXML2Other
mzXML instance document
Integrity check
validateXML
mzXML schema
Information is converted to an open, vendor-independent representation.
Conversion to tertiary formats
Random access parsers
dta
mgf
pkl
txt
RAP RAMP
SEQUEST
Mascot
ProteinLynx
Database search
XPRESS
ASAPRatio
Argos
Quantification
mzXML viewer
InsilicosViewer
Pep3D
Visualization
Analytical Software
Various software tools can access (either directly, or after conversion to tertiray formats) the information from different instruments from a single source.



(From Taylor et al., 12/2003)

PROGRESS ON THE MIAPE TOOLKIT (Dec. 2003)



USEFUL LINKS AND REFERENCES

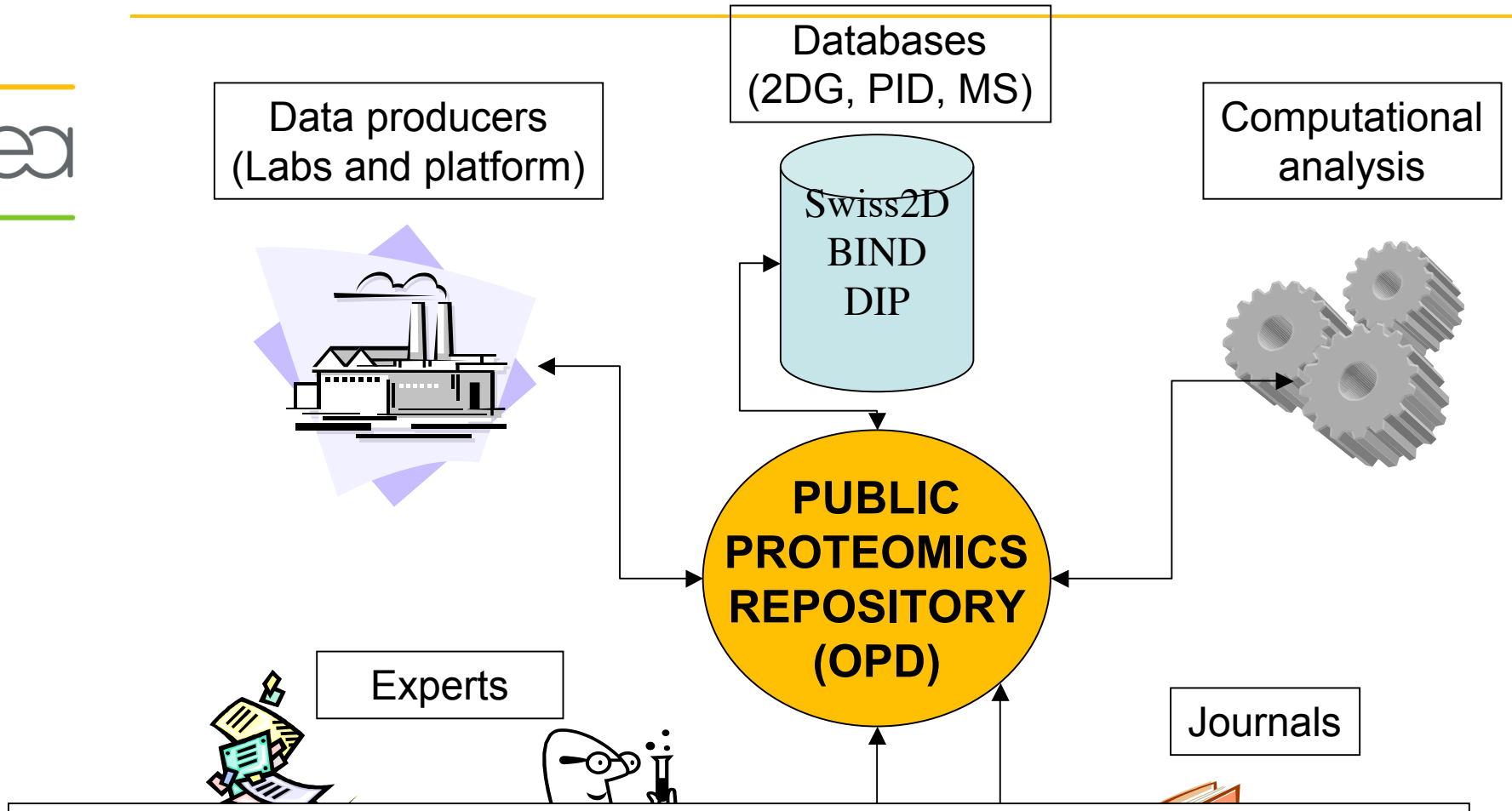
URL :

- **World-2Dpage - http://au.expasy.org/ch2d/2d-index.html**
- **Open Proteomics Database - http://bioinformatics.icmb.utexas.edu/OPD/**
- **PSI (HUPO) - http://psidev.sourceforge.net/gps/index.html**
- **PEDRO : http://pedro.man.ac.uk/index.html**
- **Sashimi project : http://sashimi.sourceforge.net/software.html**

Bibliography :

- **Prince, J.T. et al. The need for a public proteomics repository. 2004. Nat. Biotechnol. 22 : 471-472.**
- **Hermjakob, H. et al. The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. Nat. Biotechnol. 22 : 177-183.**
- **Taylor, C. et al. A systematic approach to modeling, capturing and disseminating proteomics experimental data. 2003. Nat. Biotechnol. 21 : 247-254.**

PROTEOMICS DATA DISSEMINATION : THE QUESTIONS



French Genopole Network Context :
Data exchange : necessity (and constraints), agreement?
Model & Data format : involvement about current attempts?