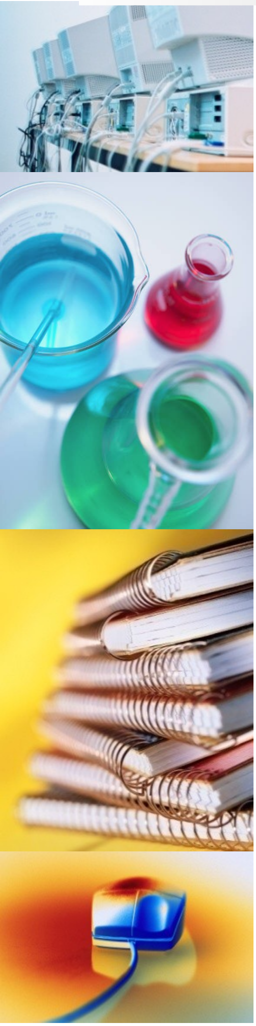


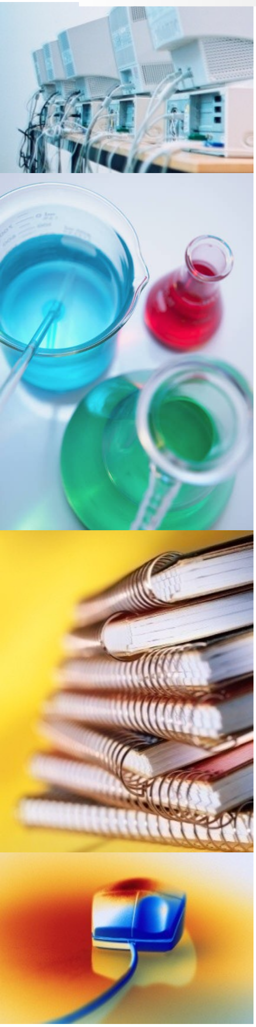
Protein information Management System: knowledge capture for protein production

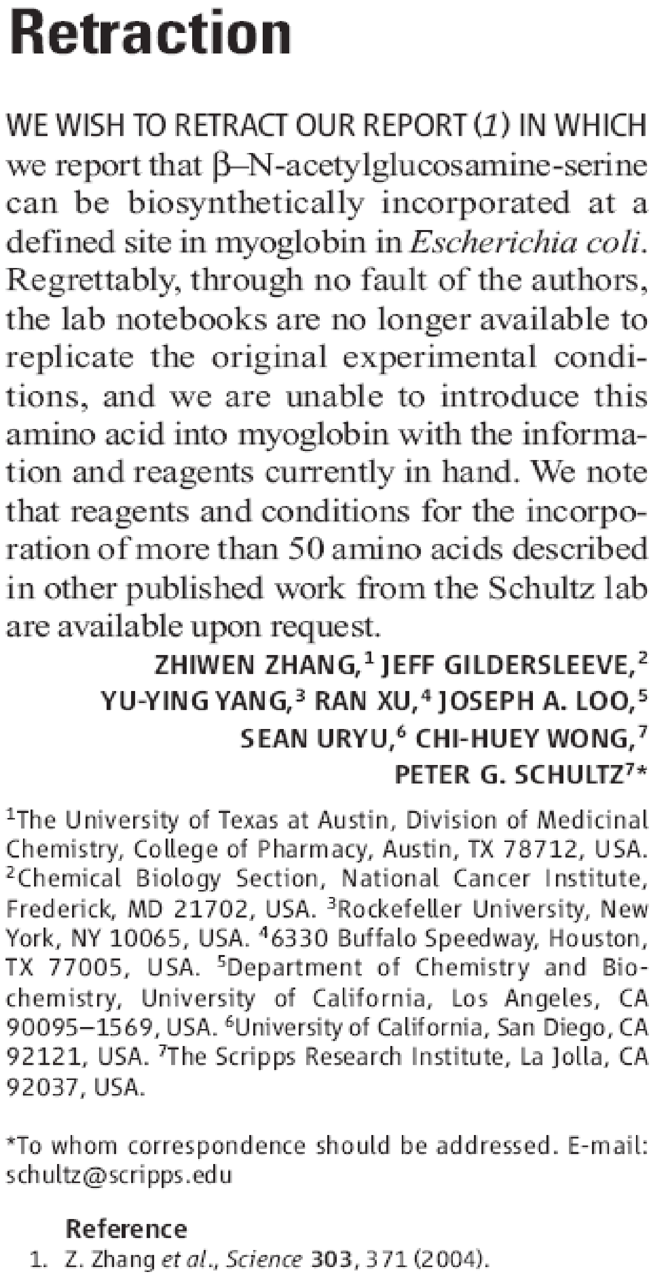
Chris Morris

STFC…

…and the PIMS development team

September 2010





Retraction

WE WISH TO RETRACT OUR REPORT (*1*) IN WHICH we report that β–N-acetylglucosamine-serine can be biosynthetically incorporated at a defined site in myoglobin in *Escherichia coli*. Regrettably, through no fault of the authors, the lab notebooks are no longer available to replicate the original experimental conditions, and we are unable to introduce this amino acid into myoglobin with the information and reagents currently in hand. We note that reagents and conditions for the incorporation of more than 50 amino acids described in other published work from the Schultz lab are available upon request.

ZHIWEN ZHANG,[1] JEFF GILDERSLEEVE,[2] YU-YING YANG,[3] RAN XU,[4] JOSEPH A. LOO,[5] SEAN URYU,[6] CHI-HUEY WONG,[7] PETER G. SCHULTZ[7]*

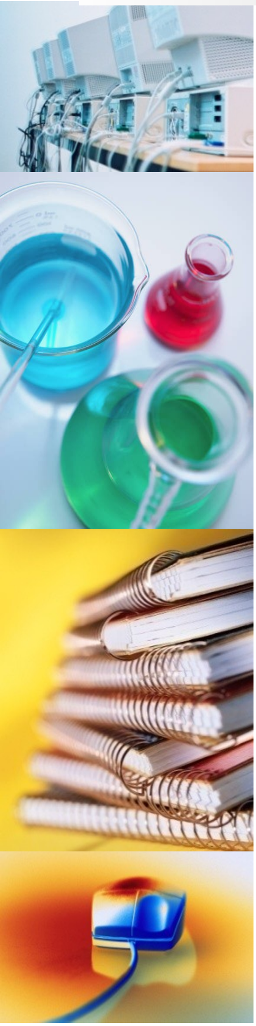
[1]The University of Texas at Austin, Division of Medicinal Chemistry, College of Pharmacy, Austin, TX 78712, USA. [2]Chemical Biology Section, National Cancer Institute, Frederick, MD 21702, USA. [3]Rockefeller University, New York, NY 10065, USA. [4]6330 Buffalo Speedway, Houston, TX 77005, USA. [5]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095–1569, USA. [6]University of California, San Diego, CA 92121, USA. [7]The Scripps Research Institute, La Jolla, CA 92037, USA.

*To whom correspondence should be addressed. E-mail: schultz@scripps.edu

Reference

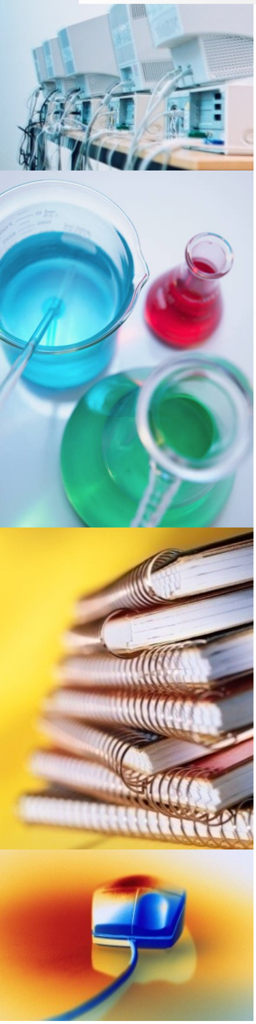
1. Z. Zhang *et al.*, *Science* **303**, 371 (2004).

Outline of talk



- Needs of academic molecular biologists
- Collaboration in drug discovery
- Some lessons learnt

The Scientific Process



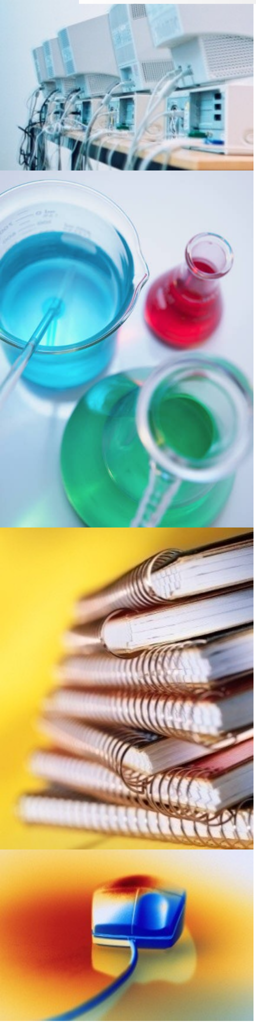
- Select “target” (using GenBank etc)
- DNA processing
- Transfection
- Expression
- Purification
- Crystallogenesis

Scope of PiMS/xtalPiMS

- X-ray diffraction
- Structure solution
- Deposition in PDB
- Structure interpretation

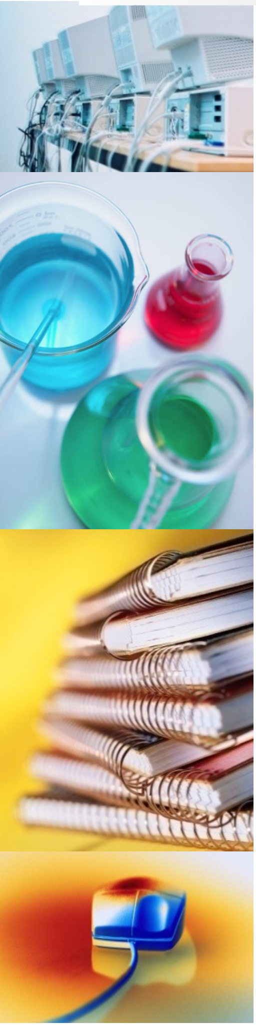
And/or NMR
And/or EM
And/or biochemistry
And/or ...

Why use a LIMS for protein production?



- Proof of priority
- Traceability, preservation of assets
- Searchability, Manageability, Continuity, Integration
- Flexibility, Future Proofing
- - lucky that we worked for **two** consortia
- Collaboration with security
- Publication and archiving
- Methods improvement
- Growth of “Integrative Structural Biology”

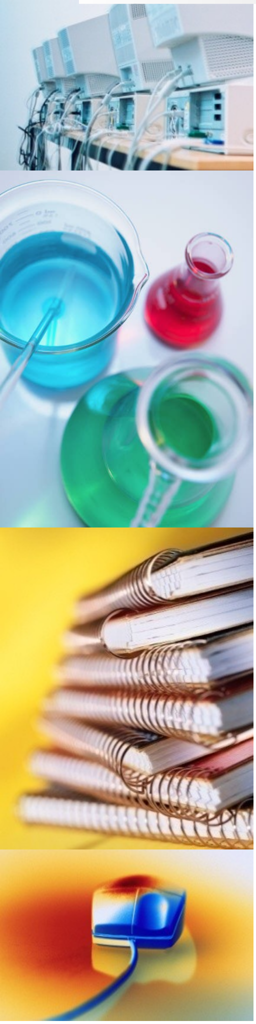
The long term vision



A unified and extensible set of software tools for molecular biology, offering seamless data transfer and a consistent user experience, from target selection to extraction of biological significance from the structure

? -> PiMS -> xtalPiMS

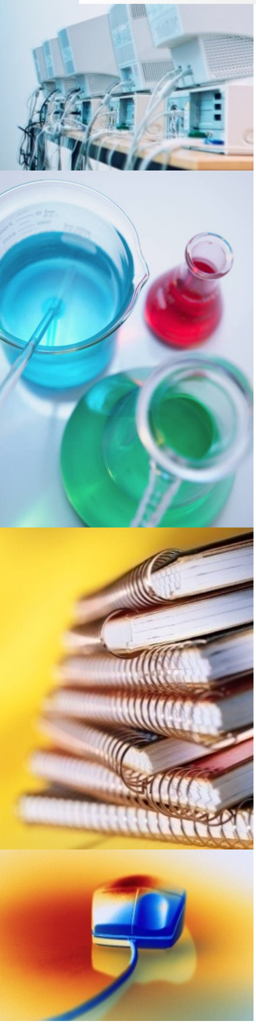
-> ISPYB -> Xia -> CCP4 -> ?



Crisis in Drug Discovery

- Opportunities:
  - Aging population in West
  - More consumers in East
- But over last ten years:
  - Big Pharma's R&D spend up 125%
  - Number of new chemical entities getting FDA approval down 40%
- 10,000 small biotech companies …

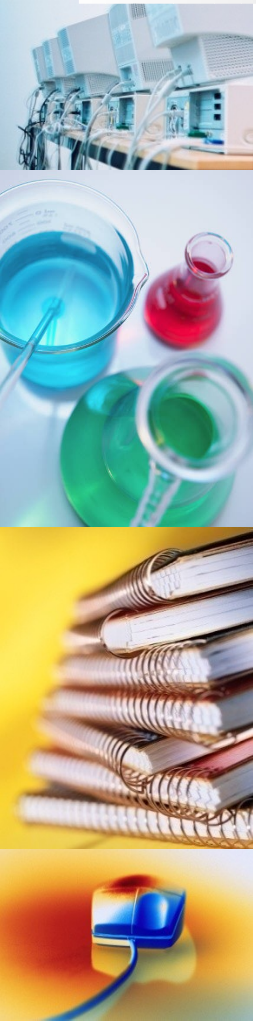
Responses to Crisis



- Closure of many in-house R&D groups
- Sharing risk with academic groups
- Market in drug candidates
  - Product could include application pack
- Use of Contract Research Organisations
  - Many in India
- Pistoia Alliance standards process

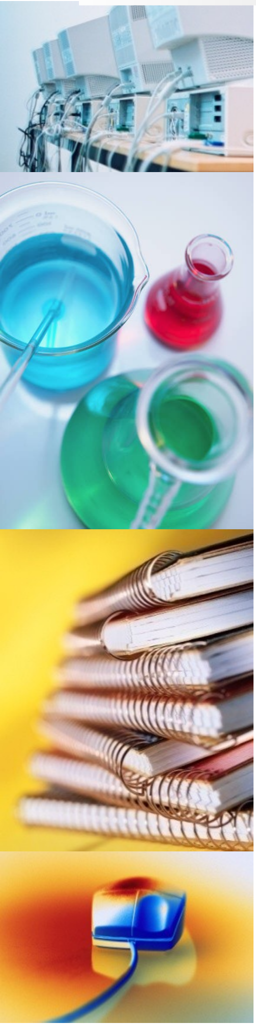
Need for collaborative knowledge management solutions

Lesson Learnt: Data Modelling is hard



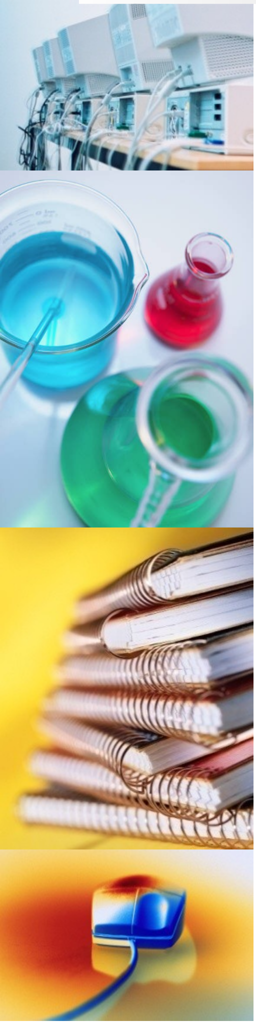
- Temin: “Intellectually I felt that the central dogma was true, but that it didn't explain my results ... Since this is biology, I didn't have any philosophical problems with my results being an exception – biology doesn't have the force of physics.”
- Tanford and Reynolds: “A common bond was a tacitly agreed permissiveness - *carte blanche* for whatever your vision to future progress might be.”

Development of the xtalPiMS schema



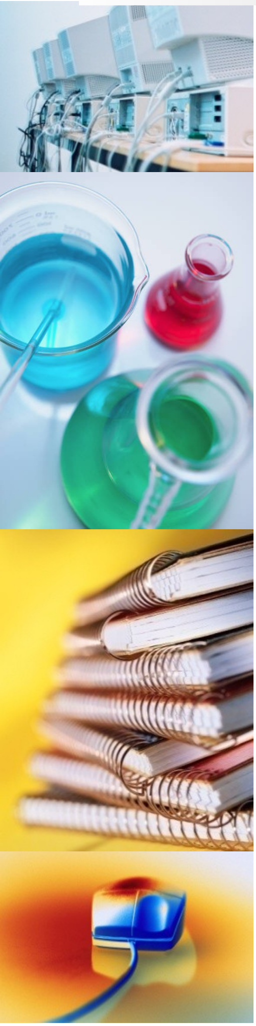
- Crystallogenesis is a well defined experiment
  - Until Cubic Phase techniques
- Protein expression is more varied
  - DNA chemistry simple, protein chemistry complex
- Uses of soluble protein not yet all modelled

Development of the xtalPiMS schema



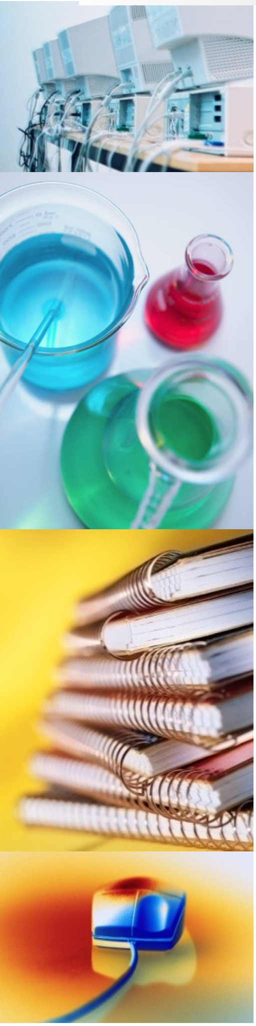
- Crystallogenesis is a well defined experiment
  - Until Cubic Phase techniques
- Protein expression is more varied
  - DNA chemistry simple, protein chemistry complex
- Uses of soluble protein not yet all modelled
- There is always a “not yet”

Technologies used



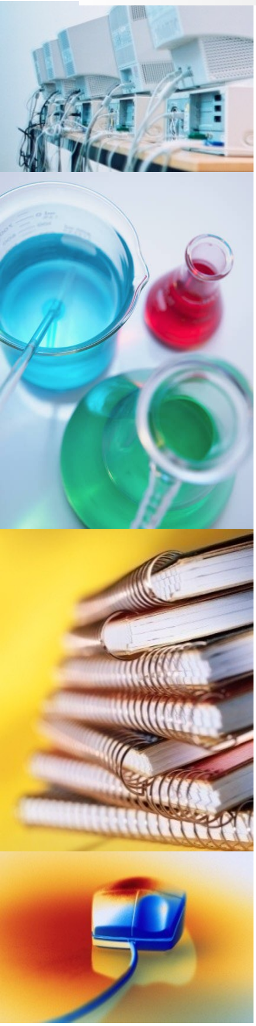
- RDBMS, with schema in DDL
- Java, with business objects, DAOs, DTOs
- UML data model
- All inherently strong schemata

e-IRG recommendations



- R4: persistence of metadata
  - Give everything a doi
  - Make business plans for continuation
- R7: interoperability
  - One digital research object can usefully be annotated in more than one metadata language
  - Decouple metadata stores from research object stores
- R12: non-discipline-specific frameworks
  - WebDAV for research object stores
  - Properties record “metadata available at ...”
- R20: ontologies problematic
  - Especially for Life Sciences

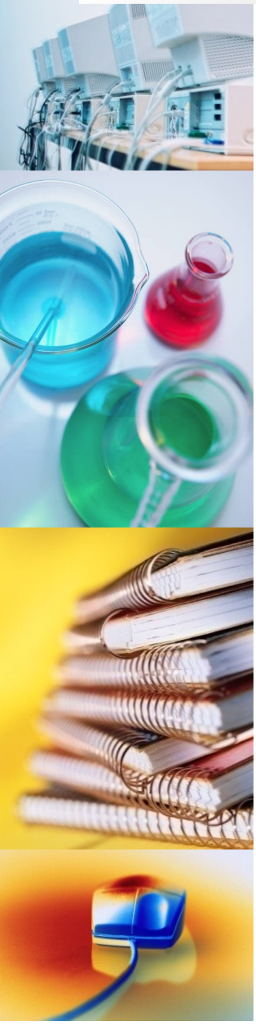
Benefits of RDF



- RDF is used by Quixote project (data management for computational chemistry)
- Research outputs are scattered
- Annotations are scattered and conflicting
- There will never be a complete schema
- RDF is designed for this
- This is what I wish I had done

PiMS

Acknowledgements

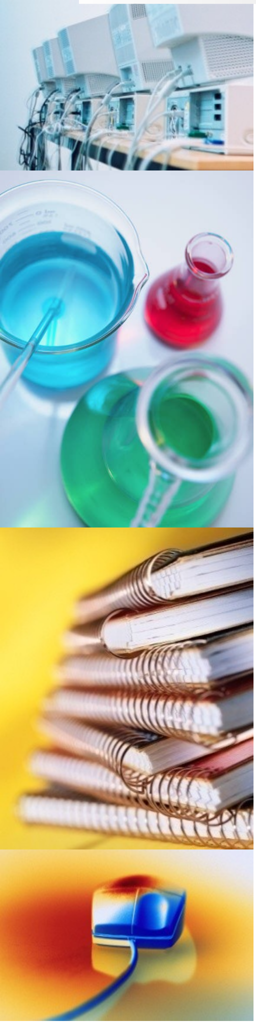


- Johan van Niekerk, Dundee
- Susy Griffiths, YSBL
- Anne Pajon, EBI
- Ekatarina Pilicheva, Marc Savitsky, Jon Diprose, Robert Esnouf OPPF
- Bill Lin, Ed Daniel, Peter Troshin STFC
- ... all who told us what PiMS should do

bbsrc
biotechnology and biological
sciences research council

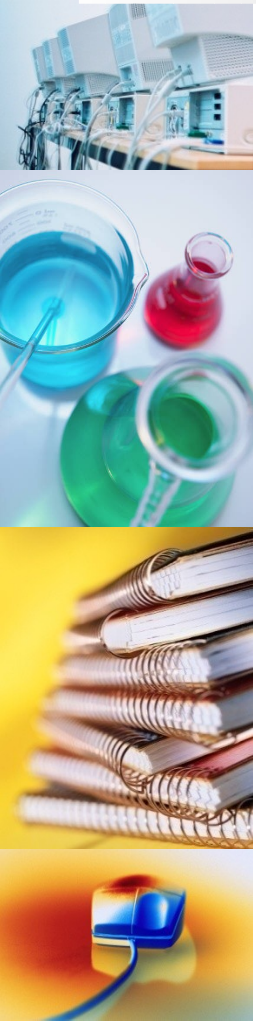
CCP4

Technologies used

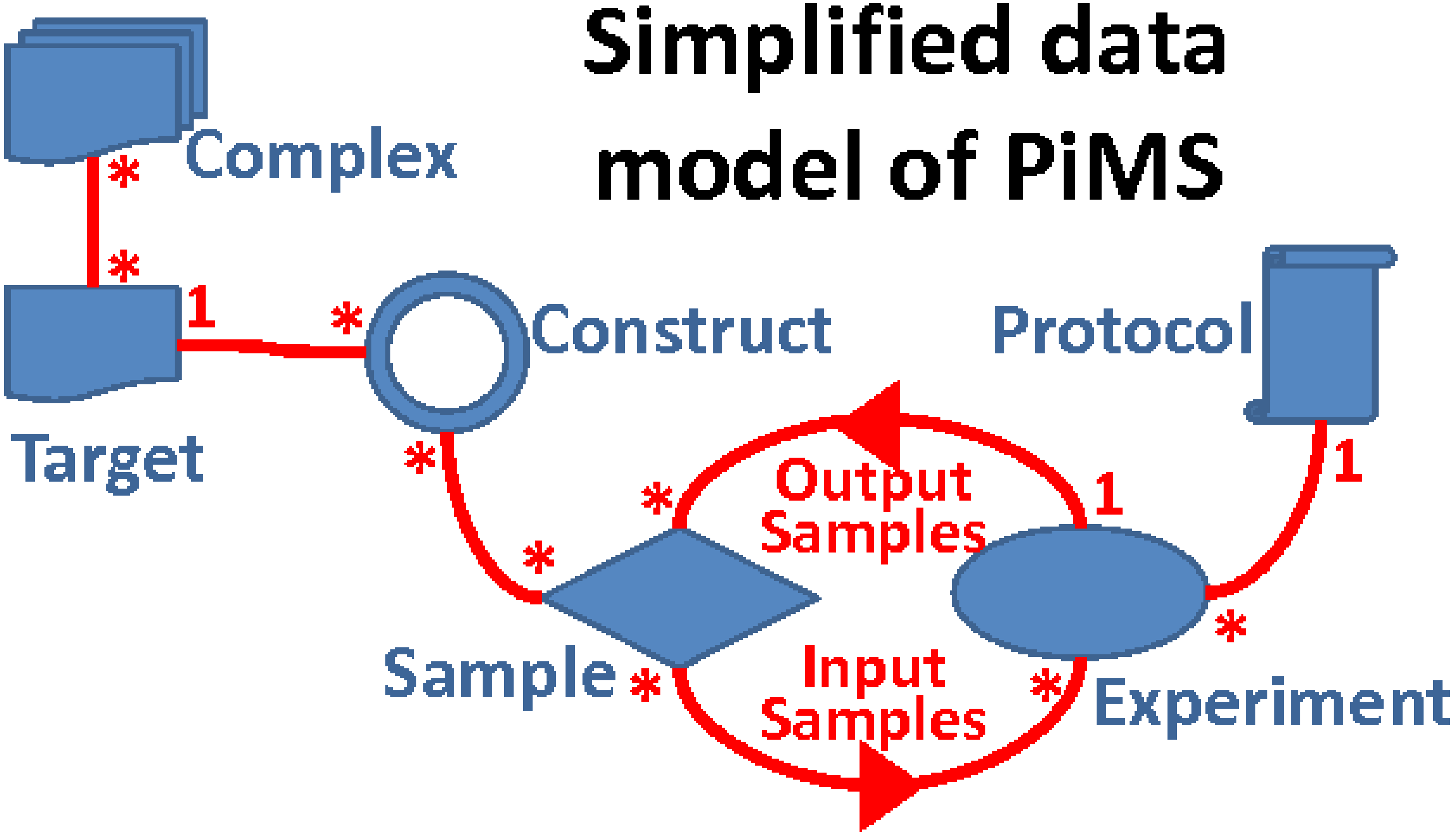


- PIMS is used from a web browser
  - Mozilla Firefox or Internet Explorer
  - No client software to install (perhaps plugins)
  - Windows, Macintosh and Linux clients
- PIMS requires a web and database server
  - Typically the same machine
  - Web server Apache Tomcat
  - Development on free PostgreSQL
  - Now available for Oracle
  - Windows and Linux servers
- Technologies used by developers
  - Java1.5, Hibernate, JUnit, BioJava, dot, batik, AJAX, ...

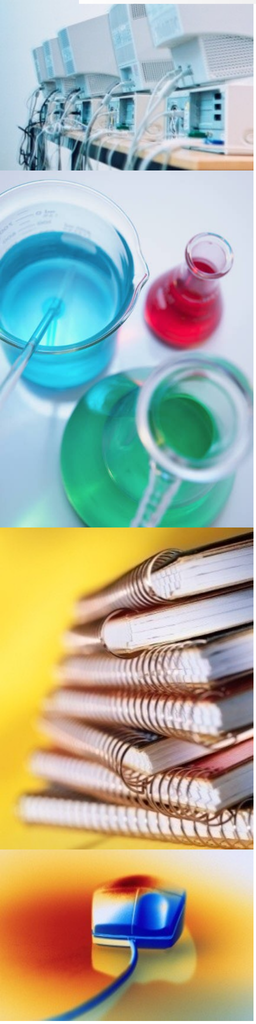
Basic concepts of PIMS



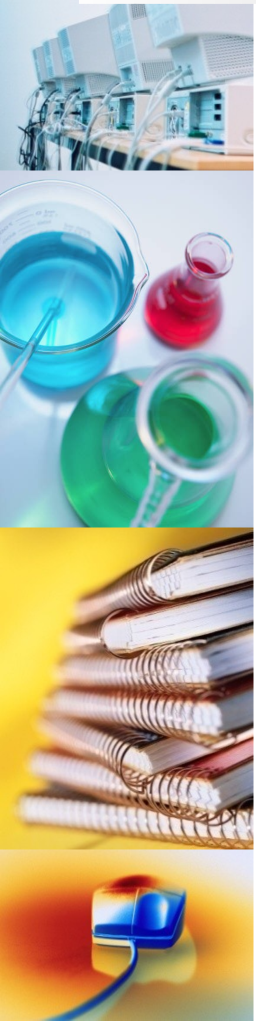
Simplified data model of PiMS



Origins of PiMS

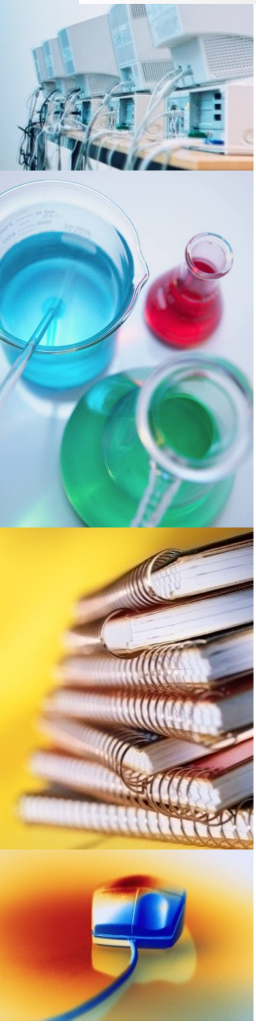


- Membrane Protein Structure Initiative
  - 1064 Targets, 2536 Experiments, 3467 Samples
- Scottish Structural Proteomics Facility
  - 392 Targets, 3709 Experiments, 1344 Samples
- Research funding to develop PiMS for them
- 38 person years work
- Structure Based Drug Design is similar to academic protein science



Who else uses PiMS?

- OPPF-UK turned off Nautilus
- IRB (Barcelona), CSIRO (Australia), Albert Einstein College (USA), EMLB Hamburg, IQTB (Lisbon), ...
- A Pharma/CRO collaboration installing it
- A hosting company evaluating it
- 18 academic licences
- 50 registered users on academic hosted service



Future work

- Support for Gel Scanner, Cartesian, ...
  - The best way to enter data is automatically
- Public repository
- Reporting
  - Commercial enhancements to academic PiMS
- Import and export of data
  - Pistoia Alliance standards process
- Conformance to Dublin Core, CERIF
- Salary committed to July 2013