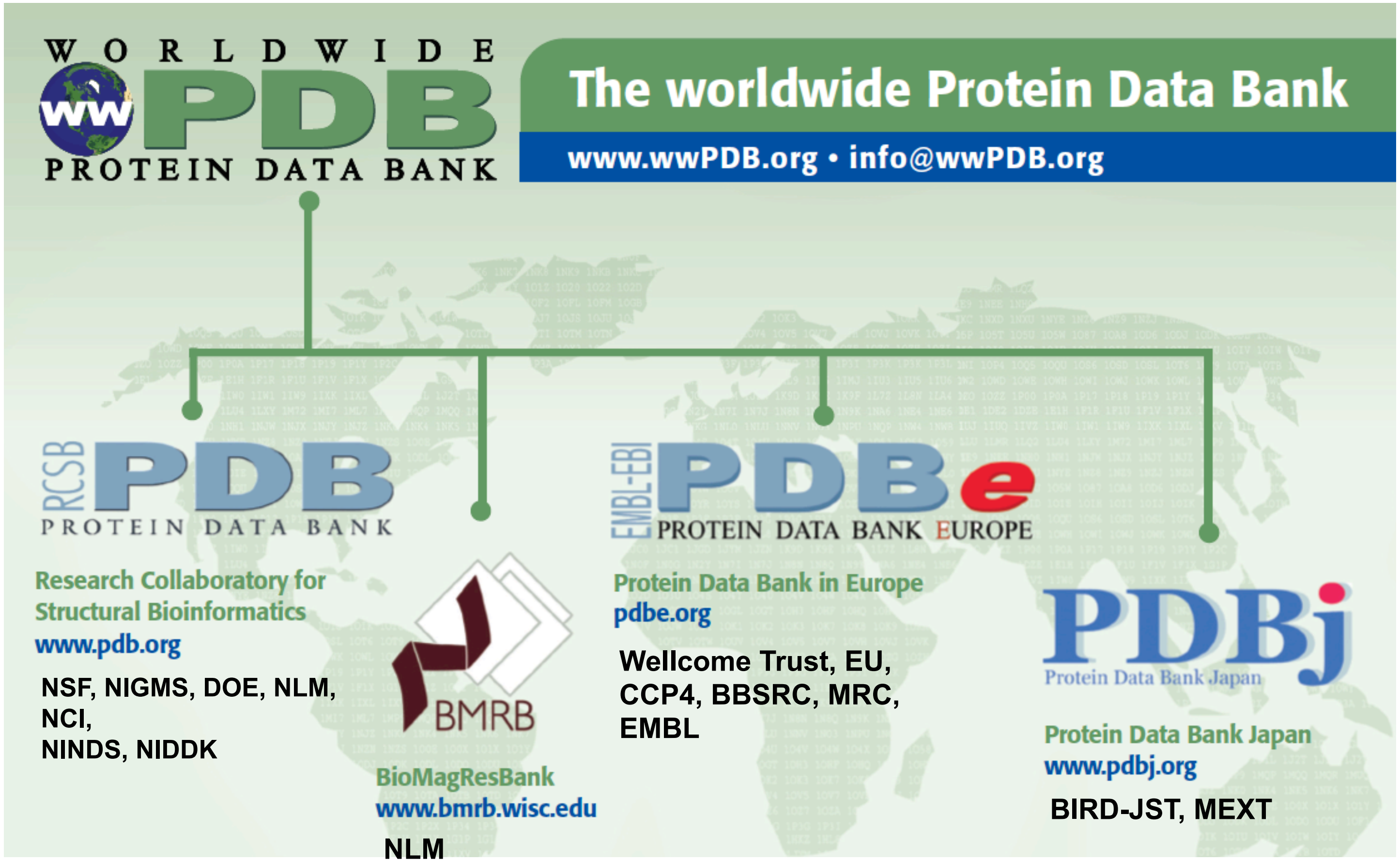
Some Economic Considerations for Managing a Centralized Archive of Raw Diffraction Data

John Westbrook

www.wwpdb.org

Overview

- PDB as a community partner
- Challenges and scope of archiving primary data
- Some technical and cost alternatives
- Possible incremental strategy


WORLDWIDE wwPDB PROTEIN DATA BANK
The worldwide Protein Data Bank
www.wwPDB.org • info@wwPDB.org
RCSB PDB PROTEIN DATA BANK
Research Collaboratory for Structural Bioinformatics
www.pdb.org
BMRB
BioMagResBank
www.bmrb.wisc.edu
EMBL-EBI PDBe PROTEIN DATA BANK EUROPE
Protein Data Bank in Europe
pdbe.org
PDBj Protein Data Bank Japan
Protein Data Bank Japan
www.pdbj.org

NSF, NIGMS, DOE, NLM, NCI,
NINDS, NIDDK

NLM

Wellcome Trust, EU, CCP4, BBSRC, MRC, EMBL

BIRD-JST, MEXT

A unique scientific collaboration providing the authoritative global resource for experimentally determined 3D structures of important macromolecules.

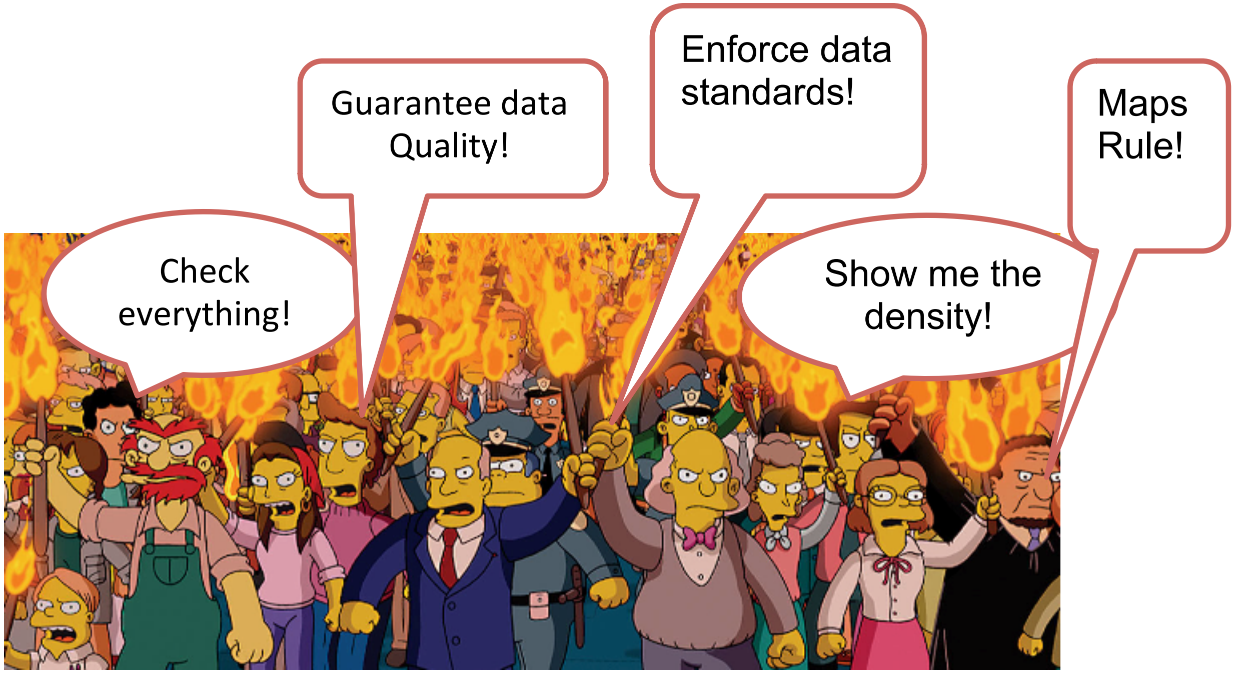
- Formalization of current working practice
- MOU signed July 1, 2003
- Announced in *Nature Structural Biology* November 21, 2003
- **Each partner funded locally**

Changing View of PDB

A Generally Hands-Off Role

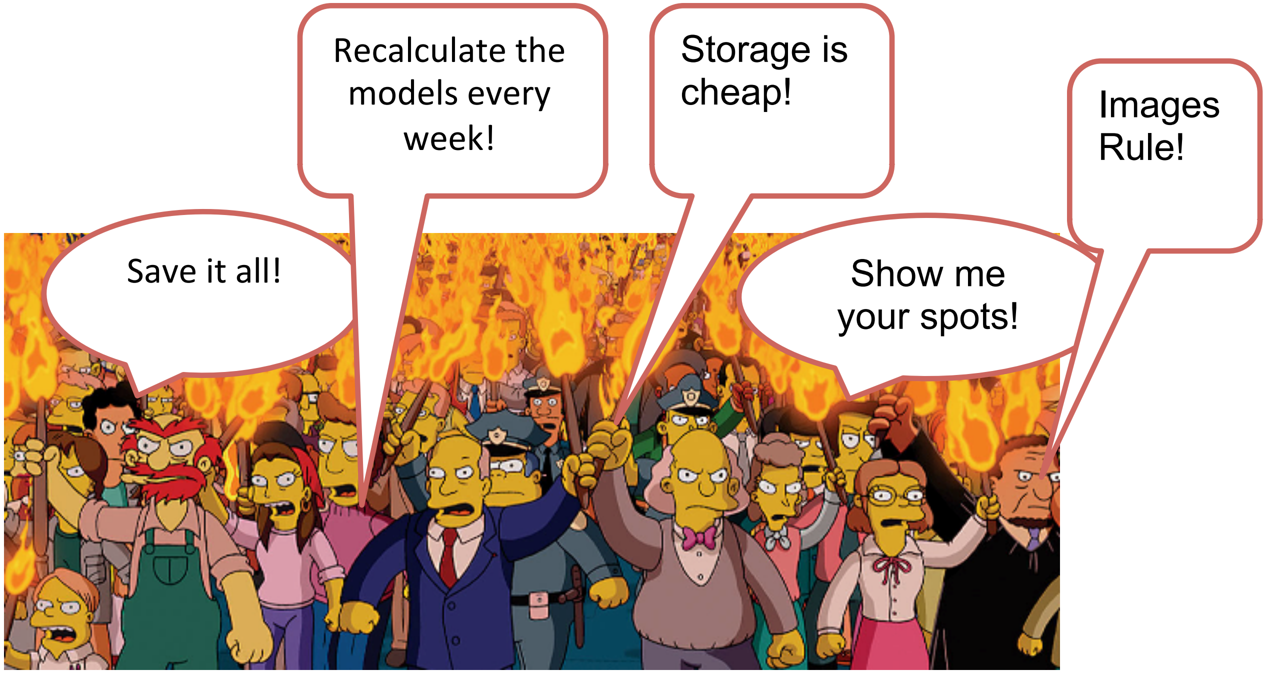
Changing View of PDB

Increasing Emphasis on Data Quality



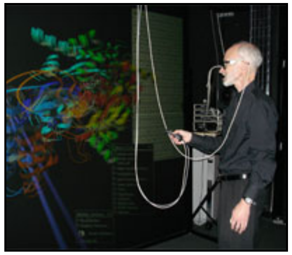
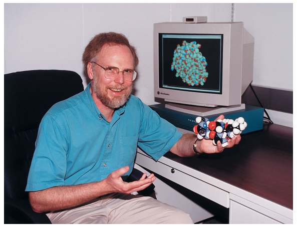
Changing View of PDB

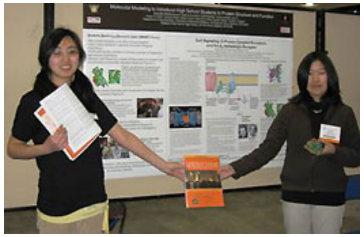
Increased Emphasis on Data Archiving

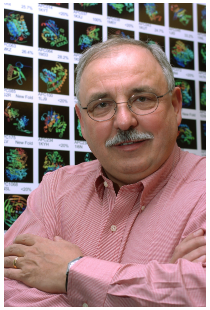


Who Downloads PDB data?

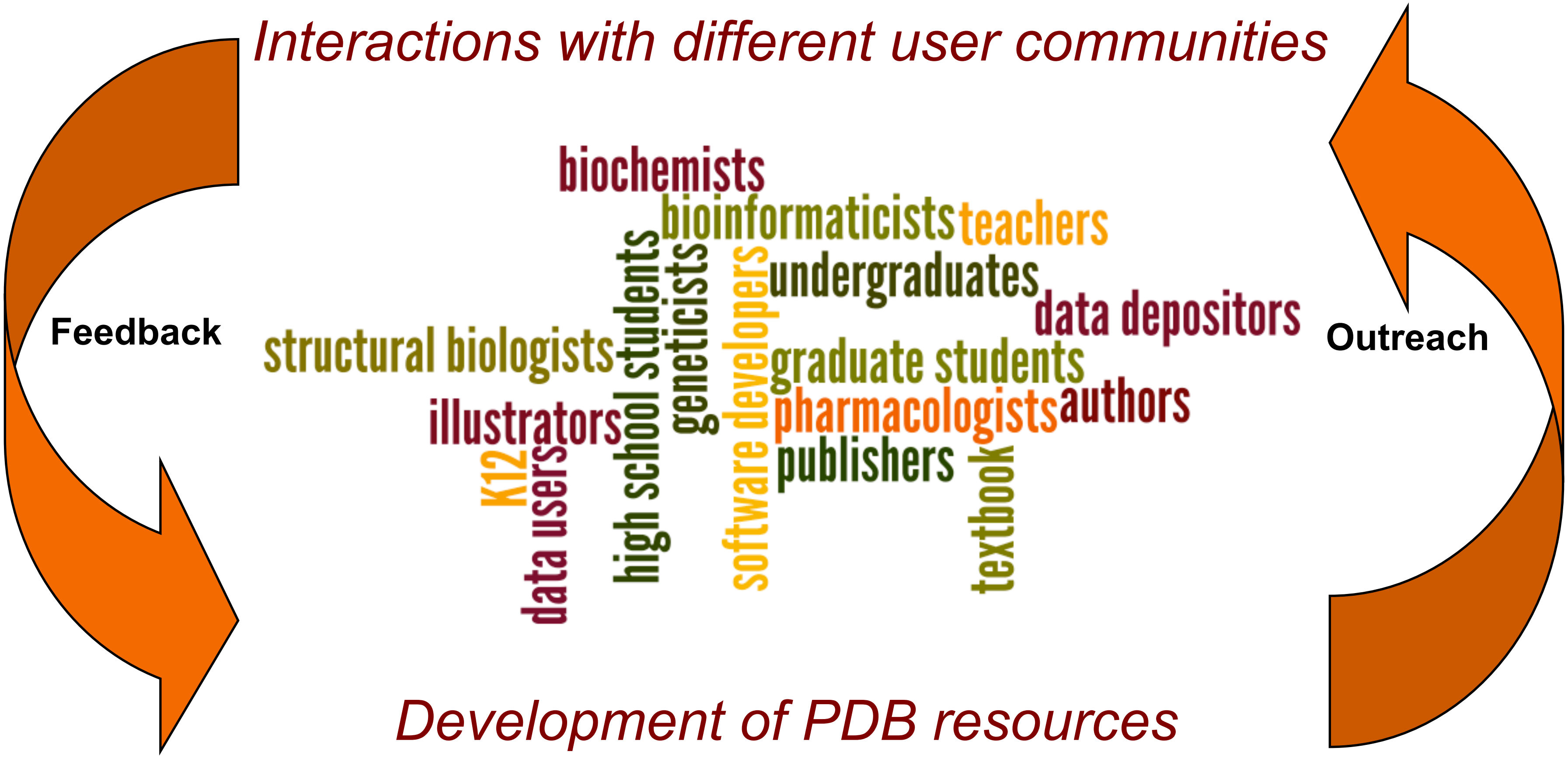
- Structural biologists
- Experimental biologists
- Computational biologists
- Biochemists
- Molecular biologists
- Educators
- Students





How Do We Interact With These Communities



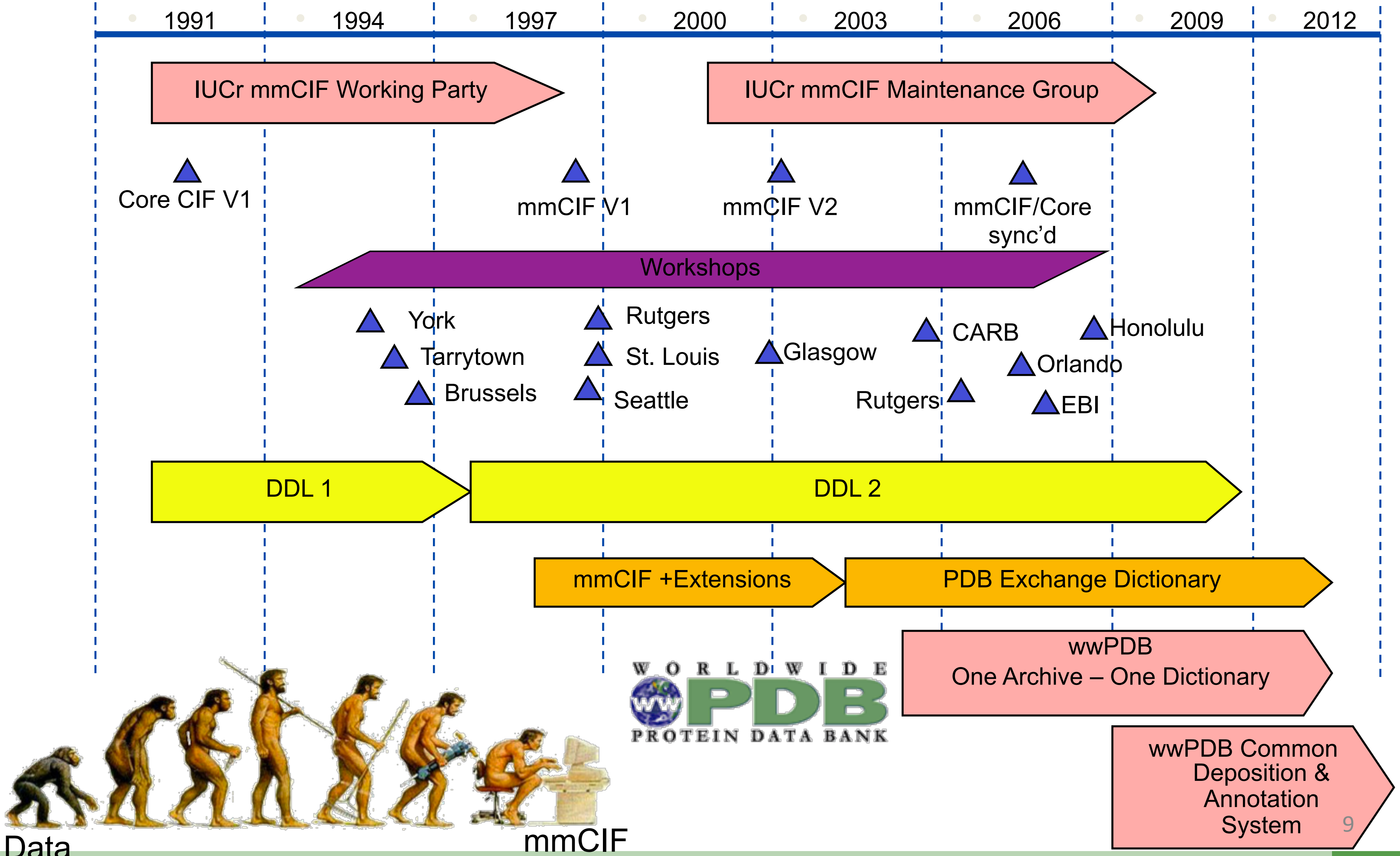
Interactions with different user communities

Feedback

Outreach

Development of PDB resources

Community Driven Data Standards PDBx/mmCIF



1991 1994 1997 2000 2003 2006 2009 2012

IUCr mmCIF Working Party

IUCr mmCIF Maintenance Group

Core CIF V1

mmCIF V1

mmCIF V2

mmCIF/Core sync'd

Workshops

York

Tarrytown

Brussels

Rutgers

St. Louis

Seattle

Glasgow

Rutgers

CARB

Orlando

EBI

Honolulu

DDL 1

DDL 2

mmCIF +Extensions

PDB Exchange Dictionary

wwPDB
One Archive – One Dictionary

wwPDB Common Deposition & Annotation System

WORLDWIDE PDB PROTEIN DATA BANK

Data

mmCIF

wwPDB Task Forces

To collect recommendations and develop consensus on method-specific issues, including validation checks that should be performed and identification of validation software applications.

X-ray Validation

- 2008 Workshop
- 2011 *Structure* publication
- Chair: Randy J. Read (University of Cambridge)

3DEM Validation

- 2010 Meeting
- Chairs: Richard Henderson (Maps, MRC-LMB), Andrej Sali (Models, UCSF)
- White paper in progress

NMR Validation

- Meetings held 2009, 2011
- Chairs: Gaetano Montelione (Rutgers), Michael Nilges (Institut Pasteur)
- Report in progress

Small-Angle Scattering

- 2012 Meeting
- Members: Jill Trewhella (Univ Sydney), Dmitri Svergun (EMBL Hamburg), Andrej Sali (UCSF), Mamoru Sato (Yokohama City Univ), John Tainer (Scripps)

NMR

X-ray

EM

Workshops and Working Groups

3DEM Data Exchange

I2PC Workshop 2012 - Madrid

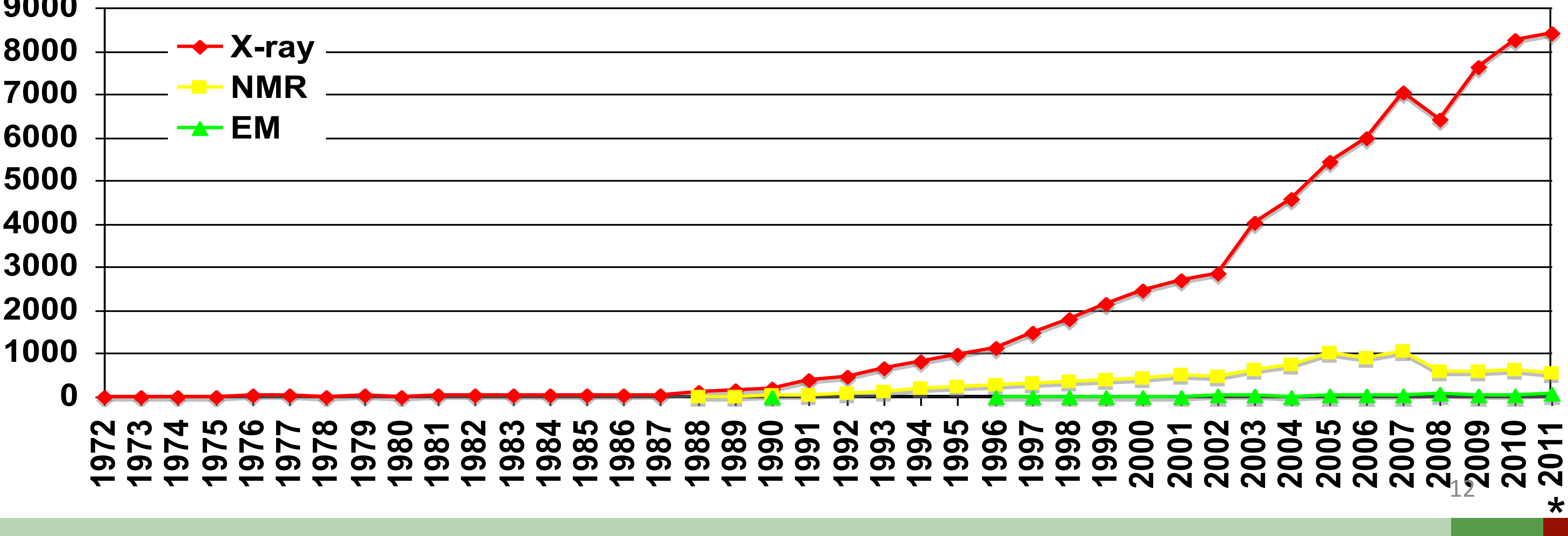
PDBx Deposition Working Group

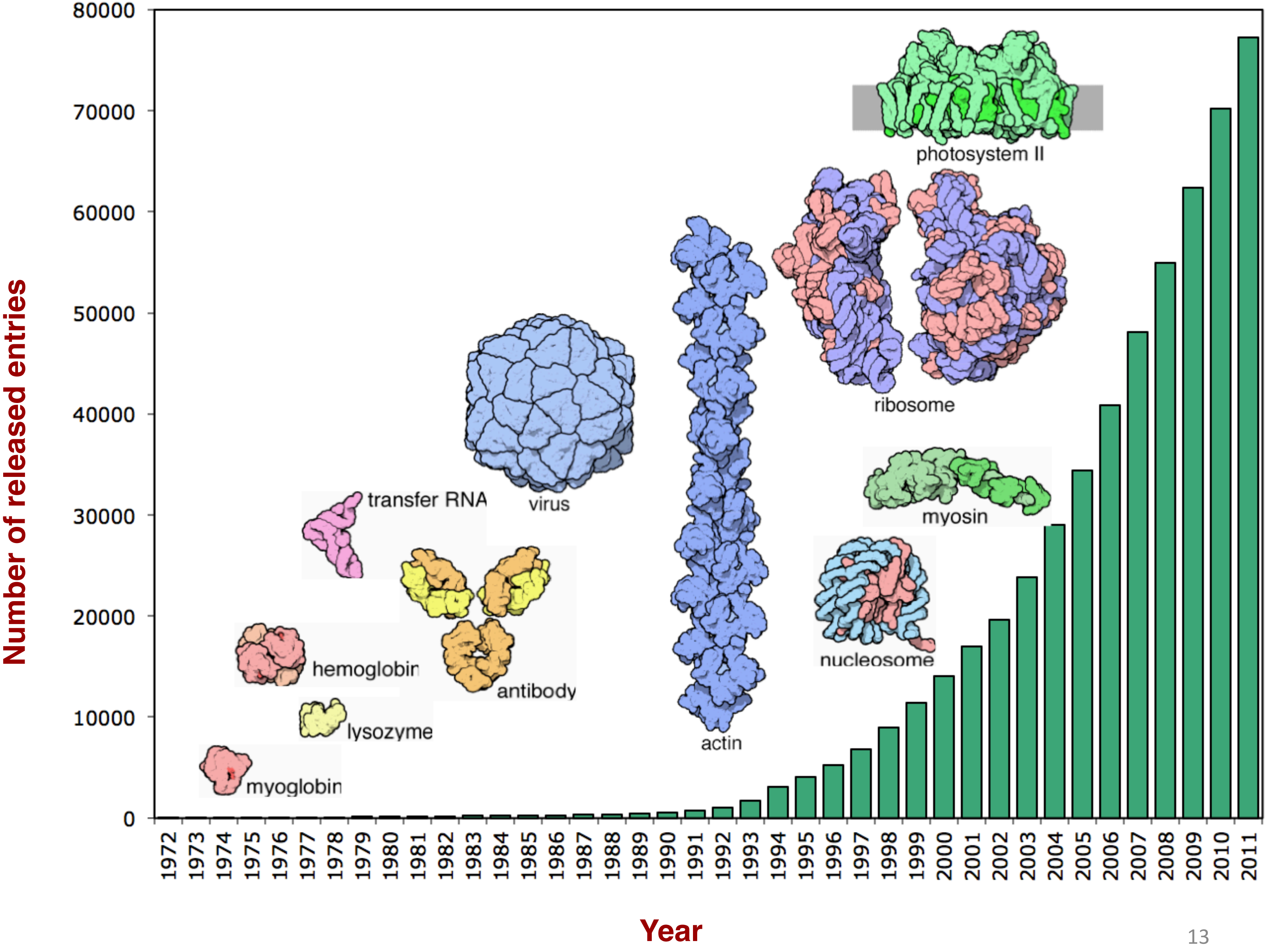
Refinement Developers Workshop 2011 - EBI

10,000-Fold Growth in Four Decades

- 83,400 entries
- 2012 will see ~10,000 depositions
- Over 85% entries include structure factor data used in the final refinement

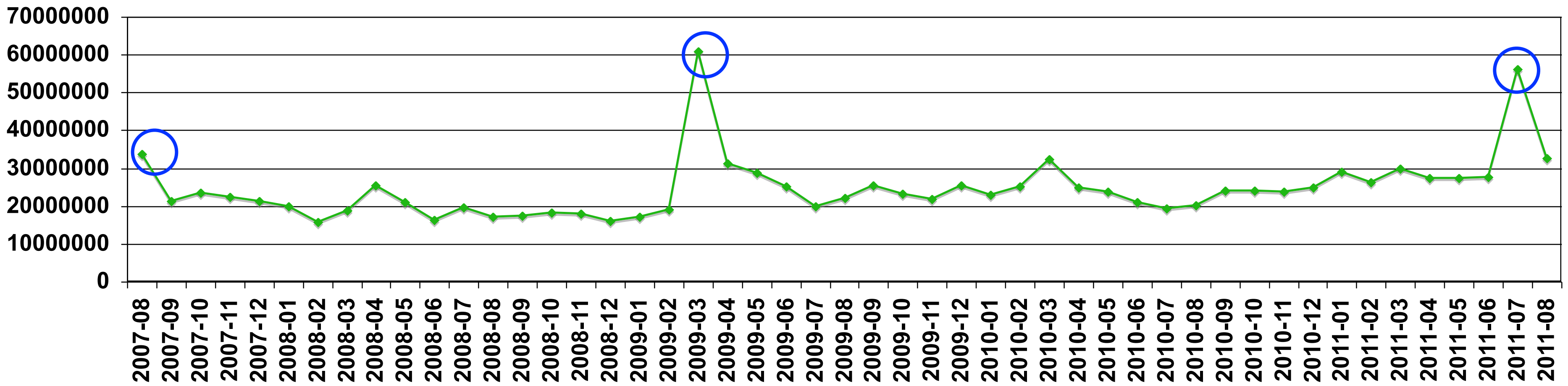
Depositions in a calendar year by experimental method





80000
70000
60000
50000
40000
30000
20000
10000
0
Number of released entries
photosystem II
ribosome
actin
virus
transfer RNA
myosin
nucleosome
antibody
hemoglobin
lysozyme
myoglobin
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Year

PDB FTP Downloads



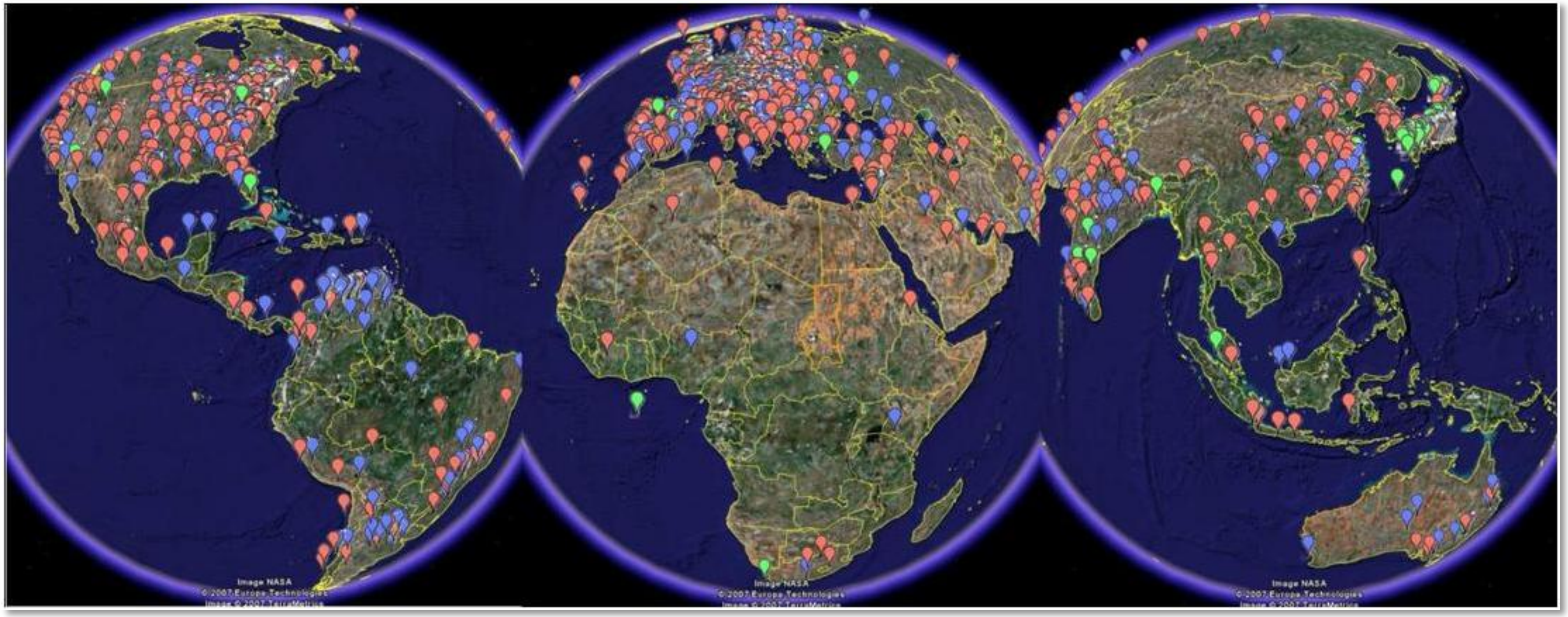
*
Version 3.0/3.1 files released

*
Version 3.15 files released

*
Version 4.0 files released

~2 M downloads/year of structure factor data files 2009-present

2010 FTP Traffic



🟥 **RCSB PDB**
159 million
entry downloads

🟦 **PDBe**
34 million
entry downloads

🟩 **PDBj**
16 million
entry downloads

Challenges and Scope

- Target content
- Longevity
- Example applications
- Audience
- Representation

What are the Content Targets?

- Laboratory data files
- Laboratory data files with supporting metadata
- Archival storage of standardized data and metadata

Expected Duration of Storage?

- Through publication review
- A few years not to exceed the availability of supporting hardware & software
- Longer …

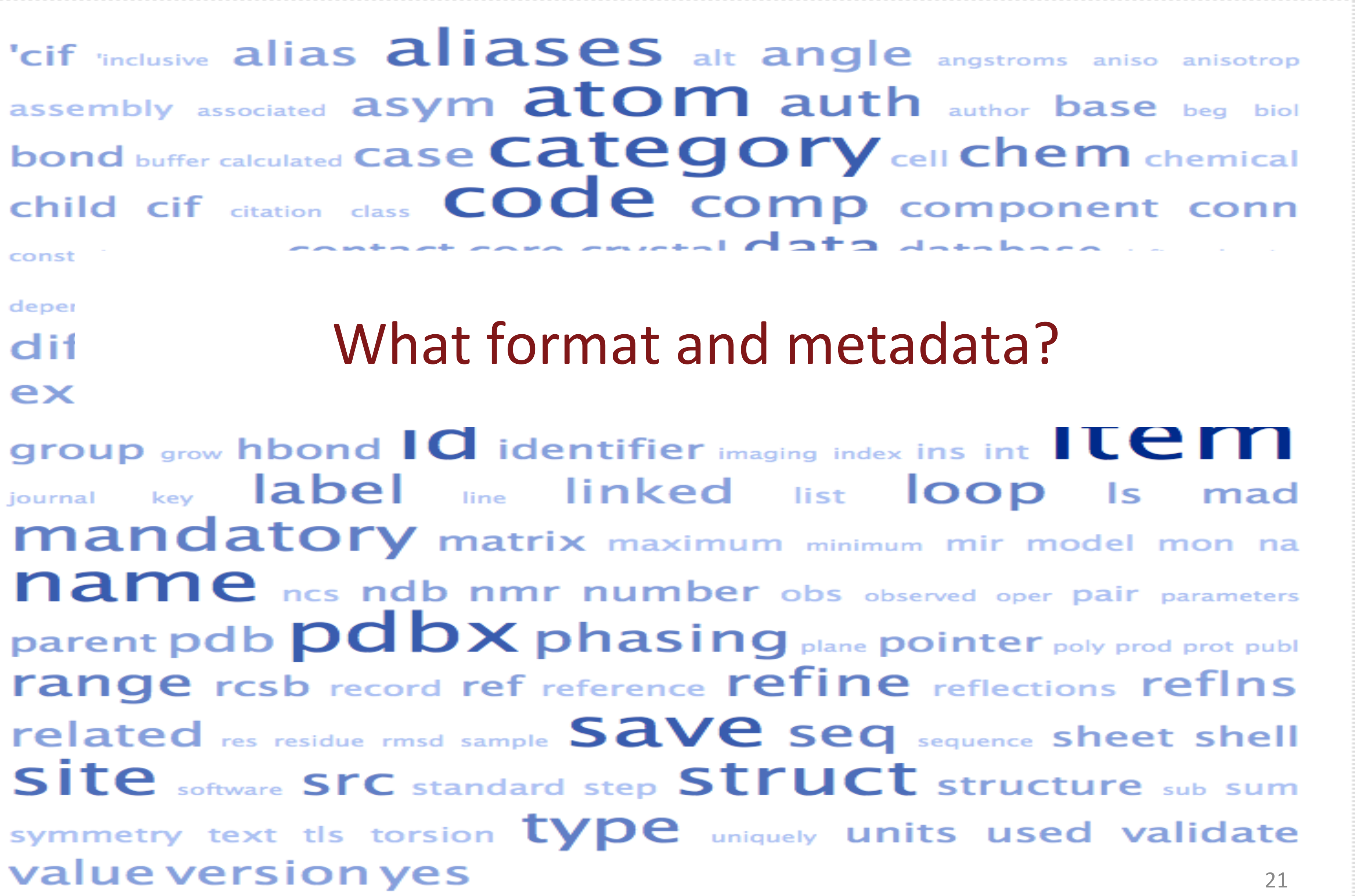
Use Cases

- Recover laboratory data files
- Satisfy philosophical/ethical/funding requirements
- Support peer review, reproduction, and validation of published results
- Extend on published results
- Provide test cases/benchmarks for methods development
- Preserve data from difficult cases

Audiences Impacted

- Direct Impact
  - Methods developers
  - Expert users
- Indirect Impact
  - Novice or non-specialist users

What format and metadata?



Archival Format and Metadata

- Solid metadata foundation for archiving -
  - CBF/imgCIF
  - PDBx/mmCIF
- Not widely used at early stages of the structure determination pipeline.
- How will working formats and process details be standardized for archiving?
  - Tar ball containing 1000 files from multiple data collections, multiple crystals, with multiple wavelengths …

Format and Metadata Targets

- Existing efforts provide data in program formats and limited software accessible metadata (e.g. TARDIS & JCSG)
- To what extent does this limit the audience and the useful lifetime of this data?

Technical Options

- Self-publishing
- Institution/facility hosting and delivery
- Centralized cloud delivery
- Centralized delivery by the PDB

Self-Publishing

- Contributor posts contents to a file sharing resource
  - Institution or facility storage resources
  - Google Drive –
    - 25GB $2.50/month - 16TB $800/month
  - Egnyte Hybrid Cloud
    - 150GB $300/yr
  - FileSwap
    - Up to 50 GB for $9.95/month
- Contributor registers DOI and digital signatures with archive

Centralized Cloud Delivery

Target one year ~ 10,000 x 5GB data sets

- Leased storage from a major provider
  - Amazon –
    - Storage - $0.125/GB/month
    - Access - ~$0.12-0.05/GB + $0.01/request
  - Google
    - Storage - $0.095/GB/month
    - Access - ~$0.21-0.08/GB + $0.01/request
- Application developed to manage depositions
- DOIs and signatures registered with archive

$75K storage + $5100/download in yr 1
$450K storage + $15.3K/download after yr 3

Archive Centralized Storage Hardware Costs

Target one year ~ 10,000 x 5GB data sets

- Cheap RAID or JBOD
  - 50TB ~ $30K or ~ $600/TB w/ 3yr maintenance
- NAS Expansion (disks and shelves only)
  - NetApp –
    - 50 TB ~ $83K or $1675/TB w/ 3yr maintenance
  - DDN -
    - 50 TB ~ $51K or $1025/TB w/ 3yr maintenance

Archive Centralized Storage

Minimum System Requirements

- Deposition site primary and backup copy
- Distribution site primary and backup copy
- Assume data requirement of 50 TB per year for the first 3 years -
  - Cheap RAID - $360 K
  - NetApp expansion - $ 996K
  - DDN expansion - $612 K
- In year 4, replace existing disk hardware + new storage for year four data.

At each wwPDB site

Archive Curation Costs

wildly optimistic estimates

- Early Stages –
  - 1 crystallographic application programmer
  - 1-2 annotators with deep expertise and troubleshooting experience with a variety of data collection, integration and phasing applications.
  - 1 scientific programmer to implement deposition and data processing automation

Comparable staffing requirements at each wwPDB site

Some Possible Practical Steps

- Tackle unmerged intensities first
- Register DOIs and digital signatures for locally store/self-published image data sets.
- Develop metadata extensions for all processing steps.
- Implement standard formats and metadata with facility control systems and pipeline software
- Pilot an automated data capture system with standard data format and metadata.

Acknowledgements

Operated by two members of the RCSB:

RUTGERS

UCSD

SDSC • SKAGGS SCHOOL of PHARMACY and PHARMACEUTICAL SCIENCES

The RCSB PDB is a member of the

WORLDWIDE wwPDB PROTEIN DATA BANK

Supported by:

NSF

NIGMS

Office of Science U.S. DEPARTMENT OF ENERGY

NATIONAL LIBRARY OF MEDICINE

NATIONAL CANCER INSTITUTE®

NATIONAL INSTITUTE OF NEUROLOGICAL DISORDERS AND STROKE

National Institute of Diabetes & Digestive & Kidney Diseases