

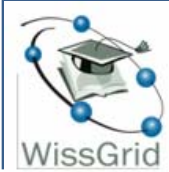
---

# Anwendungsfall: Photon Sciences & X-ray Facilities

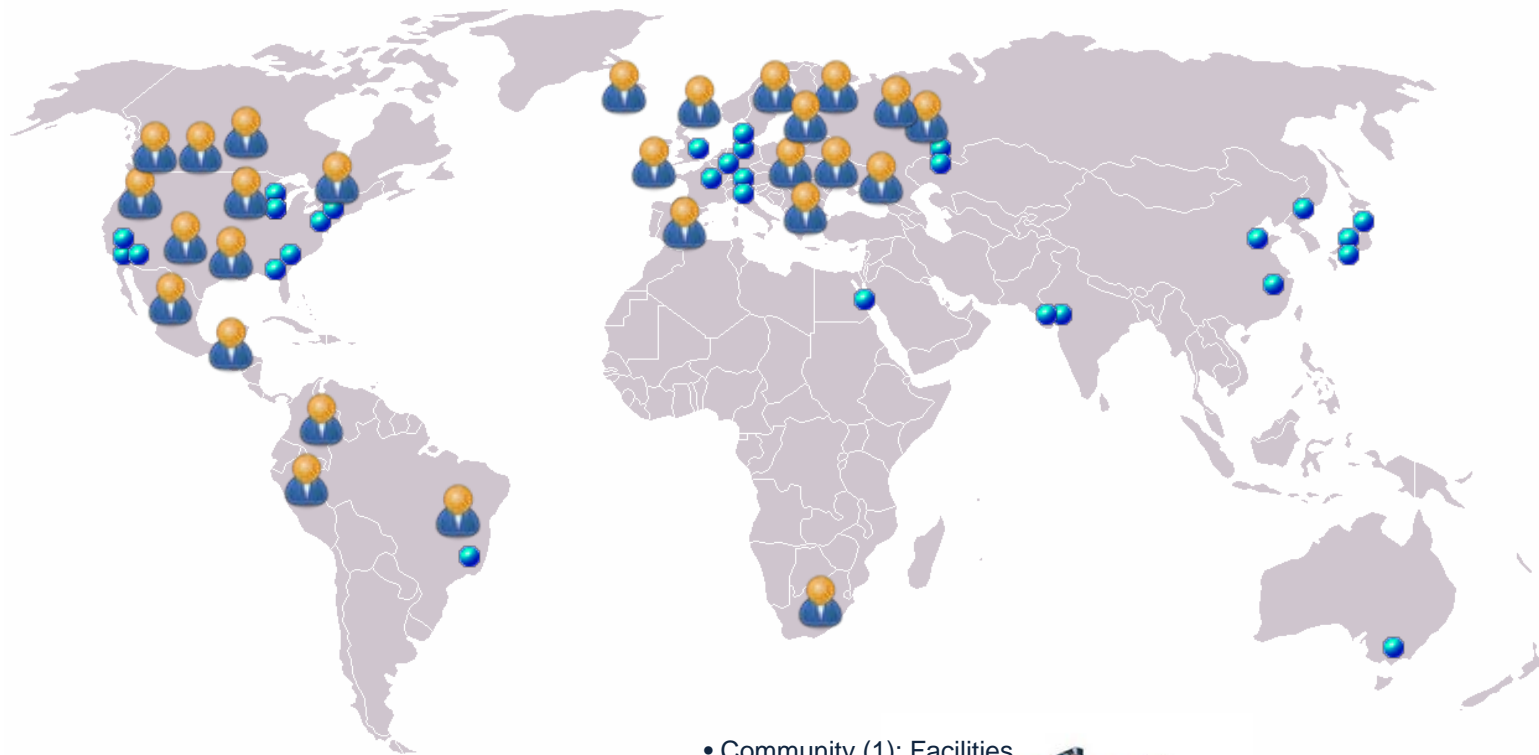
---




Begutachtung des WissGrid AP 3  
28. Januar 2010, AIP Potsdam

Frank Schlünzen  
DESY



- Communities
- LZA – Aktueller Status
- LZA – Struktur der Daten
- Prototypische Anwendungen
- Grid / Repositories



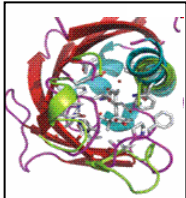
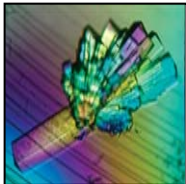
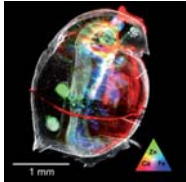
-  Fachdisziplinen
-  Anwender
-  X-ray Strahlungsquellen

• Community (1): Facilities



• Community (2): Beamlines

• Community (3): Anwender



- **Community (1): Facilities**

- Services für Beamline-Betreiber
- Definition von Standards & Policies
- International organisiert (ROSCOE, PaNData, EuroFEL,...)

- **Community (2): Beamlines**

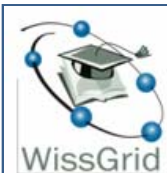
- Services für Anwender (incl. Daten-Archiv und Management)
- Implementierung von Standards & Policies
- Entwicklung von Methoden & Instrumenten
- National (PNI) und International organisiert

- **Community (3): Anwender**

- Messungen und Experimente
- Entwicklung von Methoden
- Vermeidung von Standards & Policies
- nicht organisiert

- **Community (2) der beste Partner für Grid-Projekte und LZA**

- Community (3) als Zielgruppe ist ungünstig (siehe ESRF-Up WP11)



## Communities - Partner

- **EMBL** (European Molecular Biology Laboratory)
  - Betreiber von 5 Beamlines an DORIS III
  - Betreiber von 3 Beamlines an Petra III
  - Crystallization-Facility als Service-Einrichtung
  - Automated Structure Determination Web-Services
  - Remote Operation der Instrumente
- **XFEL** (European X-ray Free Electron Laser GmbH)
  - Internationales Konsortium (17 shareholders)
  - Beamline / Facility Betreiber
  - In der Aufbauphase, operational 2014
  - Anforderungs-Analyse in Computing TdR
  - Vermutlich Tier-Daten-Infrastruktur
  - Simulationsdaten und LCLS-Daten fallen schon jetzt an
  - Erklärte Absicht Daten Grid-basiert zu managen.
- Andere: FLASH, HASYLAB, GKSS, CFEL, CSSB (und darüber hinaus)



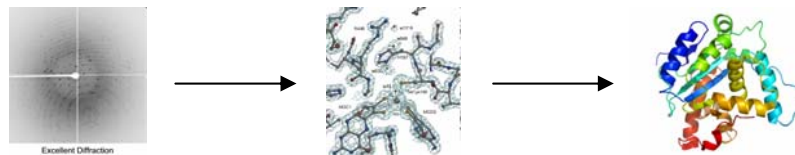
# LZA – Aktueller Status

- **Datenraten:** von 1kB/h bis 1PB/Woche
- **Kosten:** typisch 2.000€/h + Reise + Präparation etc.
  - NIH: 120-250k\$ pro Protein-Struktur (High Throughput)
  - Ribosomen-Struktur: >50M€
  - Verlust der Daten kostspielig
- Viele Objekte nicht reproduzierbar
  - Verlust der Daten tragisch
- **Status:** Anwender einzig verantwortlich für Daten
- **Archivierung:** Keine Speicherung vor Ort
- **Daten-Lifecyclemanagement:**
  - Daten gehen grundsätzlich für Communities verloren
  - Metadaten grundsätzlich nicht verfügbar
  - Physikalischer Daten-Verlust ist (mittelfristig) der Regelfall
  - Validierung ist praktisch unmöglich

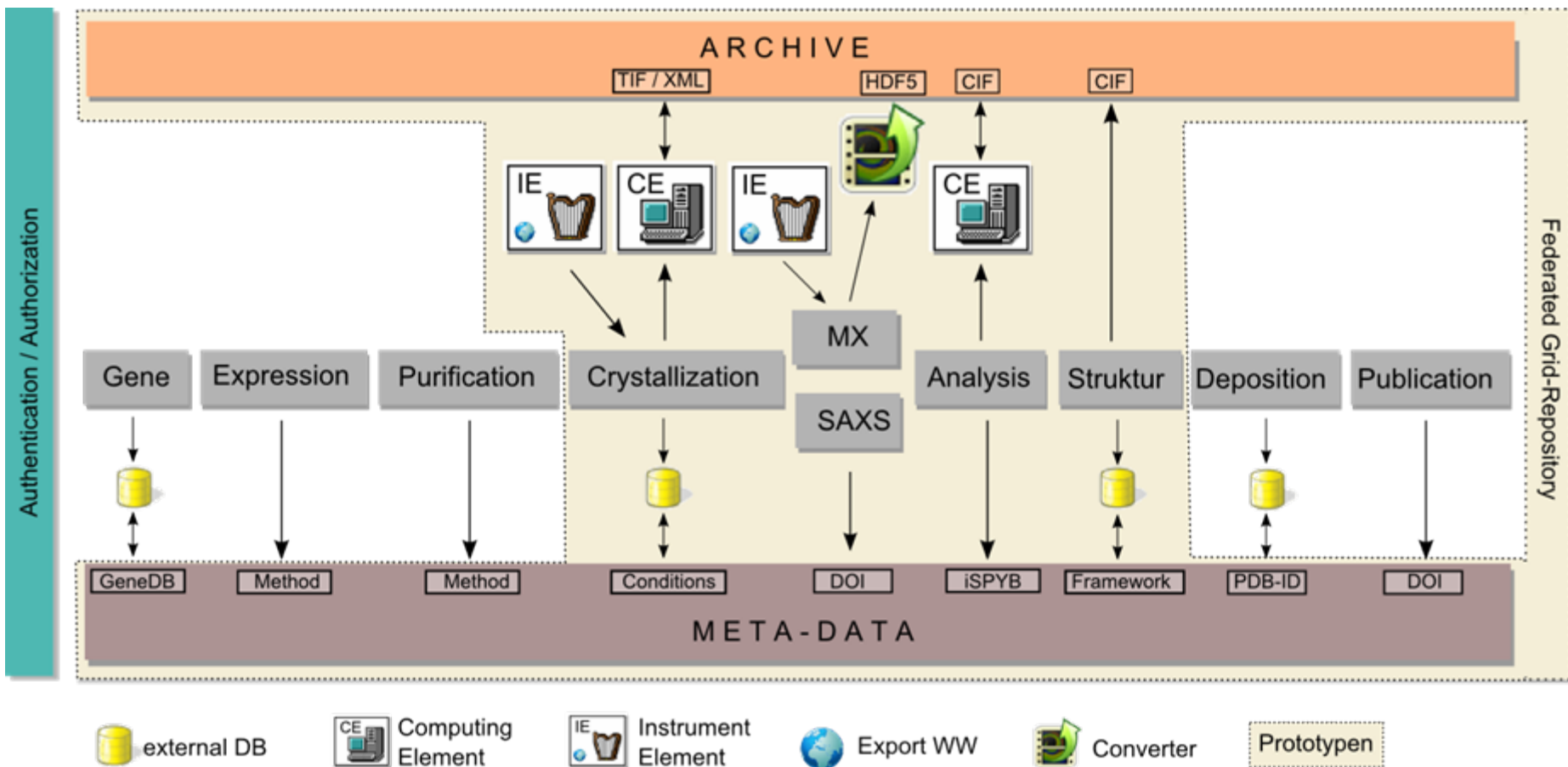
- Experimente oft an verschiedenen Instrumenten durchgeführt
  - Instrument-übergreifendes Daten-Management
  - Interdisziplinäres (Meta-) Daten-Management
- Experimente oft an verschiedenen Facilities durchgeführt
  - Facility-übergreifendes Daten-Management
- In der Regel internationale Kollaborationen
  - kooperatives Daten-Management
- Experimente erfolgen meist in starker Konkurrenz
  - sicheres Daten-Management (incl. Metadaten) essentiell
- Internationale Communities und Anwender
  - international verbindliche Standards & Policies
  - gewährleistet durch aktive Beteiligung an diversen EU-Projekten wie PaN-Data, ROSCOE, EuroFEL, ESRFUp,



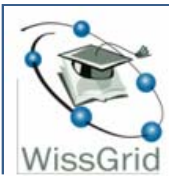
- **Prototyp 1** – EMBL (Kristallographie):
  - Überschaubare Datensätze (1-16GB)
    - aber viele Einzeldateien und Datensätze
  - **Status:** verschicke portable Medien per Carrier
  - **Zukunft:** nachhaltiges Daten-Archiv / Grid-Repository
    - Open Access für Nachnutzung (nach ~5 Jahren)
    - ... und für Methoden-Entwicklung
    - Basis für Analysis-Framework
    - JCSG Samples (incl. Rohdaten/Metadaten) verfügbar
  - Automatische Analyse-Facility existiert bereits (WS)



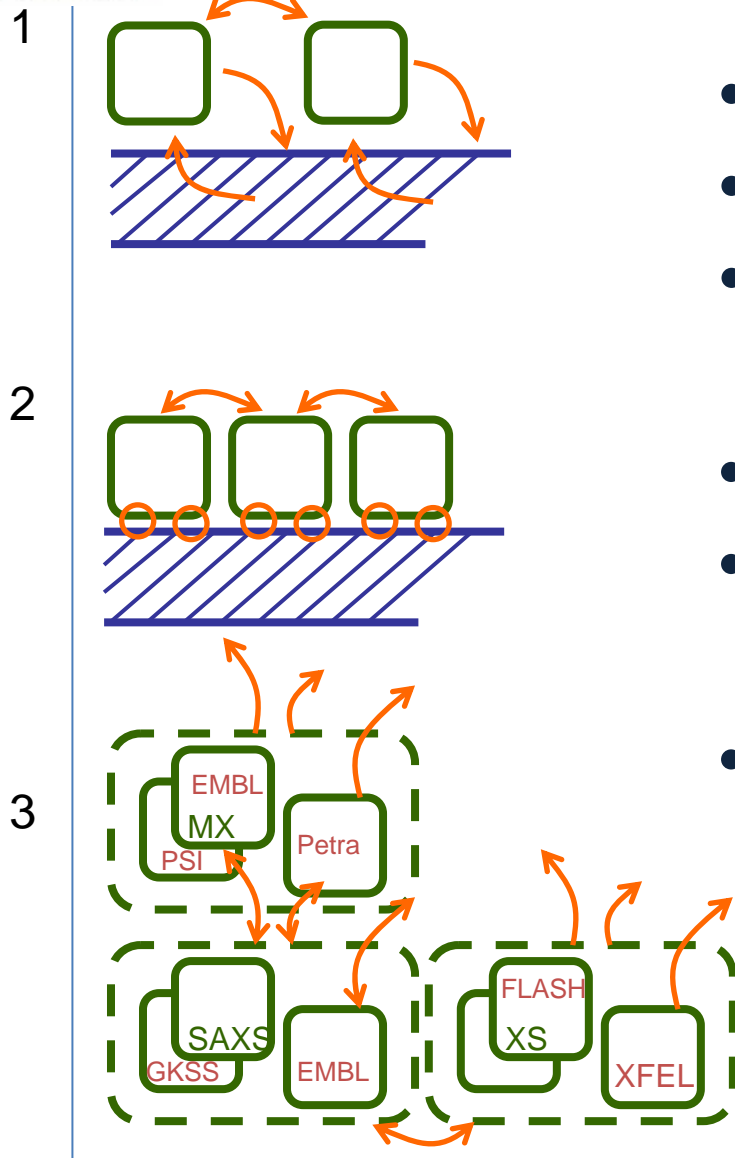
# Ziel: Integrated MX Grid-Facility



- Weitere prototypische Anwendungen in Arbeit (XFEL, EMBL)
- Diskussionen mit GKSS, FLASH, HASYLAB, ...



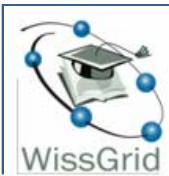
- Konzepte für LZA & Grid-Repository dringend benötigt
  - Grosse Datenmengen (2-100PB/yr/Facility)
  - Interdisziplinäre Forschung
  - Strategisch wichtig fuer FELs wie XFEL (>10PB/yr)
  - Erfolgreiche Umsetzung gut für PNI
- Fortschritte in AAA (e.g. Shibboleth)
  - Sicheres Management von Daten und Metadaten
  - Vermeidung exzessiver Zertifikat-Nutzung
- Workflow und Data Management Systeme:
  - Umsetzung der existierenden Analysis Frameworks
  - Integration existierender Metadata-Engines
  - Datenmanagement via Grid-portals



- Nutzung Grid Compute Ressourcen
- Daten zu den Diensten und vice versa
- Daten zu den Anwendern & vice versa

- Nutzung Grid Storage Ressourcen
- Bit Preservation + Trust Zones

- Föderation von Repositories
  - Zwischen Disziplinen
  - Zwischen Facilities
  - PanEuropäisch



- Sicher nicht **ein** Grid-Repository ...
- ... aber **eine** Basis-Architektur
  - interdisziplinäre Grid-Repositories
  - förderierte Grid-Repositories
  - Integration existierender Meta-Daten & Frameworks
  - APIs fuer die gängigen Datenbanken
  - APIs/Konvertierungen zwischen (Standard-) Formaten
  - ...
- Zwecks Bündelung der Bemühungen
  - Workshop *DM for Photon Sciences* in Q2/2010 ...