Data Provenance Tracking for Biomedical Virtual Research Environments

Richard McClatchey (CCCS, UWE Bristol UK)

Contents

- VREs for biomedical research
- Traceability of data & Provenance
- CRISTAL as a Provenance base
- Analysis services in NeuGRID & N4U
- The N4U Virtual Lab / Research Environment
- Conclusions & Questions

Biomedical Data Traceability

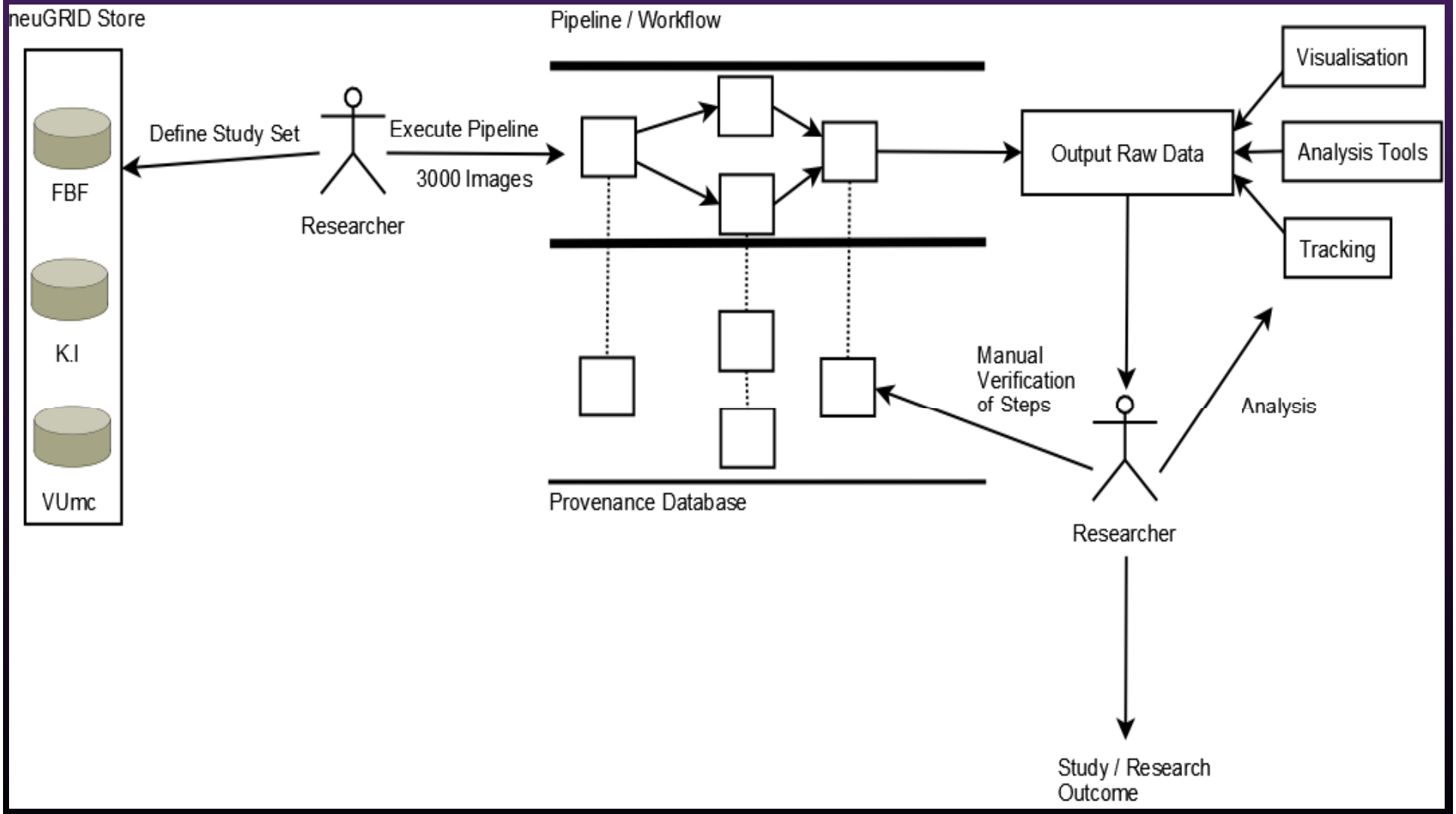
- To share data-sets & analyses.
- To enable collaborative research.
- To reproduce results / tests.
- To understand a process(es) followed
- To verify the work of others.

Need for Biomedical traceability

- Traceability provides information for
  - Explanation of source and usage of data and processes
  - The evolution of those processes
  - The verification of activities
  - Reproducibility of actions
  - Security and access control
  - Source of system failures
  - Developers to debug programs

- All of these functions are essential in Big Data medical information systems.

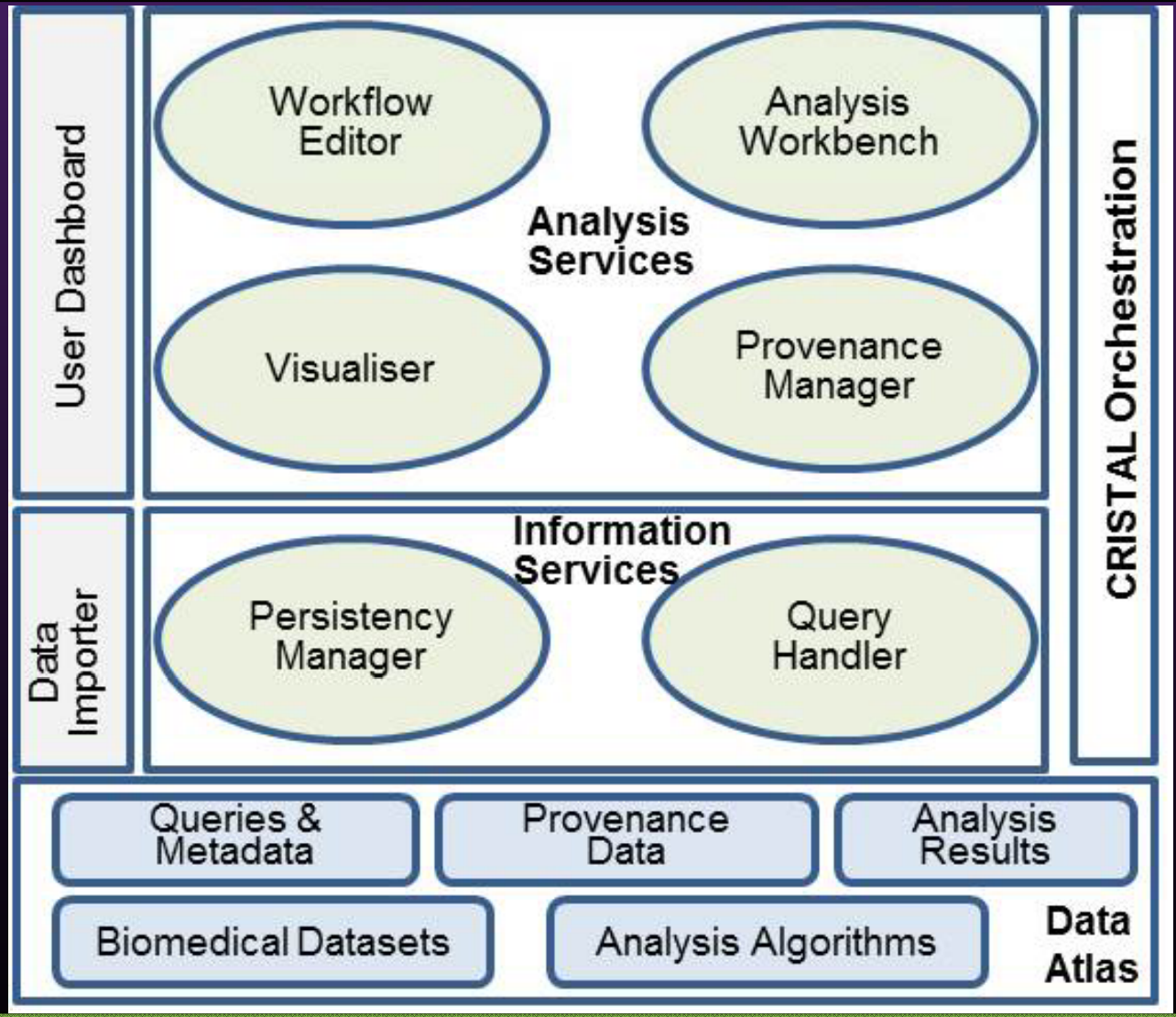
Example from Neuroimaging



Virtual Research Environments

- Platforms to enable researchers to share data and results
- Provide data governance and **traceability**
- Provide tools for browsing data and algorithms and to construct research analyses
- Visualise outcomes of analyses and reproduce results.
- Can facilitate simulation, workflow management, document hosting and collaboration support across groups/teams of researchers

Architecture of a Biomedical VRE



The role of a Biomedical VRE

- Data gathering, data fusion and homogenisation of existing and newly created datasets.
- Database and meta-data management.
- Dataset and algorithm discovery and tracking.
- Research result verification and authentication.
- Orchestration of ‘standard’ procedures.
- Linkeage with external analysis packages.
- Usage patterns and behavioural studies.
- Standardisation of data access and visualisation.

Provenance and Analyses: 7 W's

- **who** ran an analysis (username, role, iden),
- for **what** purpose, what their analysis was supposed to achieve, and **what** were its outcomes/results
- **when** they ran it (a timestamp which denotes when it started and when it finished),
- **where** it was run - this is GRID / Cloud related information,
- **which** datasets and algorithms were used to create and run their analyses (e.g. Image set and pipeline),
- **how** it was executed, this is more detailed infrastructure information
- and lastly **why** the analysis was run, this is a justification from the user, potentially with annotation.

IN ESSENCE A FULL RECORD OF THE ANALYSIS ACTIVITIES

Contents

- VREs for biomedical research
- Traceability of data & Provenance
- **CRISTAL as a Provenance base**
- Analysis services in NeuGRID & N4U
- The N4U Virtual Laboratory
- Conclusions / Questions

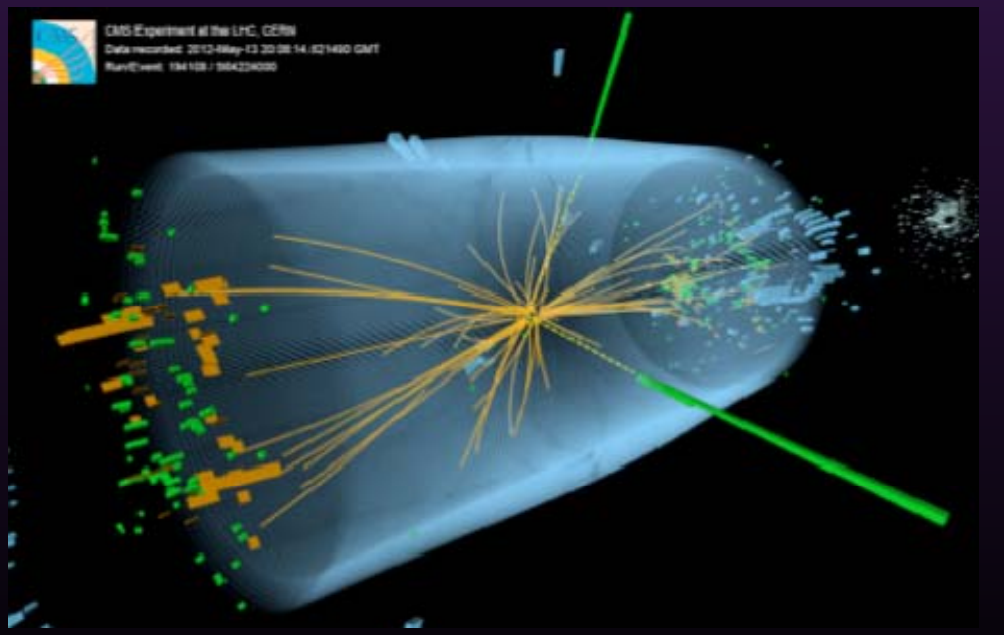
What is CRISTAL?

- A long-running research project (1997- 2012) between UWE, CERN and CNRS (France).
- That has developed data models and software using state-of-the-art technologies.
- To address the data management, workflow and process control needs of a distributed community of detector physicists (CMS Ecal in this case). In essence a VRE.
- Whose requirements were initially vague, long-term, evolving and demanding (cost/size/response).
- Which has yielded academic output and software that is being commercially exploited from 2005-2014
- And has been launched Open Source (LGPL3) in 2015

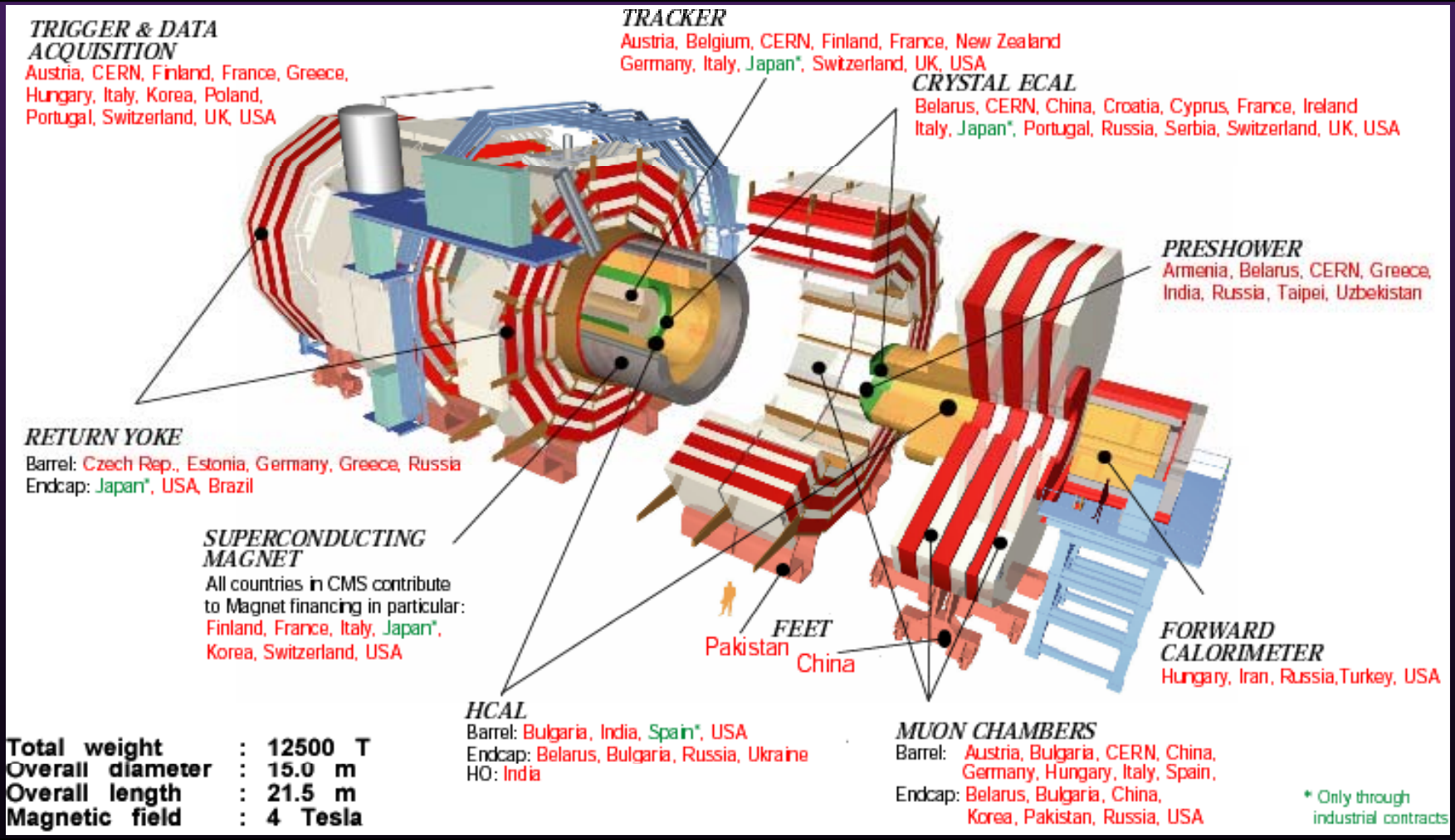
Crystal Characterisation

CRISTAL, born at CERN

- Software for recording how an object goes through its lifecycle (parts, workflows/processes, agents etc.)
- Enables traceability and provenance management
- Can be used for analysis and data tracking
- Commercial use
  - Agilium, Technoledge
- Academic/Research
  - CERN,
  - ***N4U, NeuGrid***
  - CRISTAL-ISE

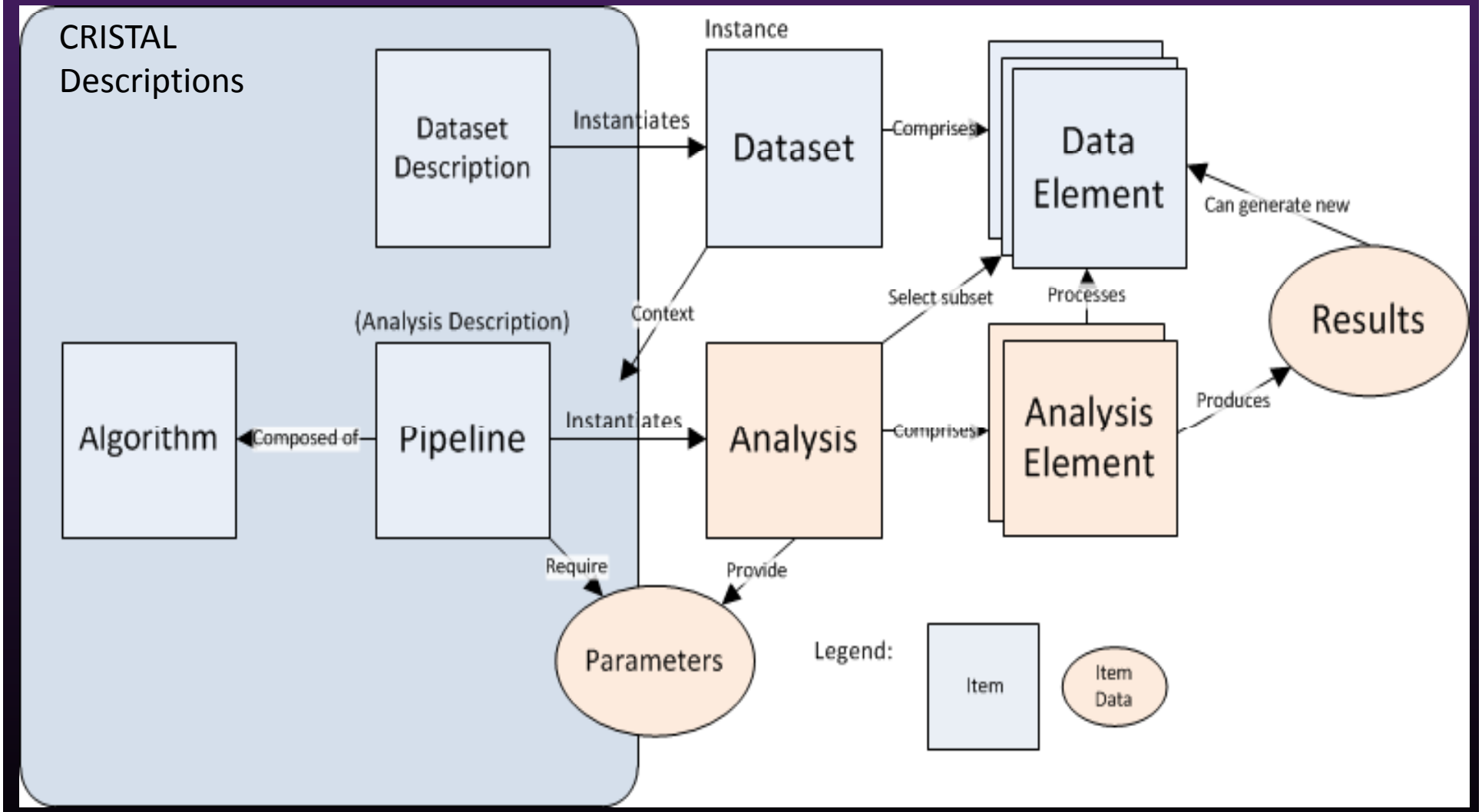


CMS : Physics at LHC



2008 scientists and engineers	160 institutes	36 countries

Provenance in CRISTAL Model



Detail of model described in conference paper

CRISTAL Features in Summary

- CRISTAL is a **framework** for collecting data by defining traceable lifecycles.
- Each step of a lifecycle defines a change of state when a piece of data is collected.
- It captures ***items*** and their descriptions + metadata.
- Lifecycle definitions are data too, called 'Descriptions' → *Description Driven System* (DDS). They are also stored as items.
- CRISTAL provides a dynamically alterable data model which copes with design-to-production change.
- It's a product and data management system that organises processes allowing the evolution of processes and domain models in a fully traceable manner.

Contents

- VREs for biomedical research
- Traceability of data & Provenance
- CRISTAL as a Provenance base
- **Analysis services in NeuGRID & N4U**
- The N4U Virtual Laboratory
- Conclusions / Questions

neuGRID for Users (N4U) : Services 4 Users

- EU Framework 7 Integrated Infrastructure Initiative, I3
- Started July 2011, 42 months, funded at €3.5M
- To provide: an e-Science environment by developing and deploying the neuGRID infrastructure to deliver a Virtual Laboratory to offer neuroscientists access to a wide range of datasets, algorithm applications, and access to computational resources, services, and support
- Partners:
  - **IRCCS Fatebenefratelli, Italy; , University of West of England, Bristol, UK;, Maat G Knowledge Spain;**
  - **Hospital University of Geneva, Swizerland. VUmc - Vrije Universitet Medical Center, Amsterdam, NL**
  - **Karolinska Institutet, Stockholm, Sweden;, CNRS, France, CEA, France;**
  - **CF consulting, Milano, Italy, MNI Montreal, Canada , UCLA, USA**

NeuGRID and N4U

	neuGRID	neuGRID for Users, N4U
		  
User services	• Cortical thickness pipeline • Core databasing • Web portal • LONI WMS	•Wider multimodal software portfolio for researchers and diagnostic neuroscientific communities • Advanced Data Base Management system • More representative datasets • Data protection extension •Educational programs
GRID services	• Security Services • Medical Quering Services • Provenance Services (CRISTAL) • Grid Gluing abstraction Services	• Knowledge management • Analysis services • Workflow authoring extension • Advanced querying extension
Infrastructure services	• Enactment Services • Computing Services • Storage Services	• Computational resources expansion • Cloud compatibility development

CRISTAL Background

- Developed at CERN in early 2000s.
- Used for the tracking of the CMS ECAL Detector construction at the Large Hadron Collider (LHC).
- The characteristics & identity of ECAL components were gathered as structured, queryable data for decision support, quality control & calibration.
- Is provenance enabled by design.
- Used in industry (BPM, Data Processing, R&D prototyping and production).
- Recently launched Open Source under LGPL3 .

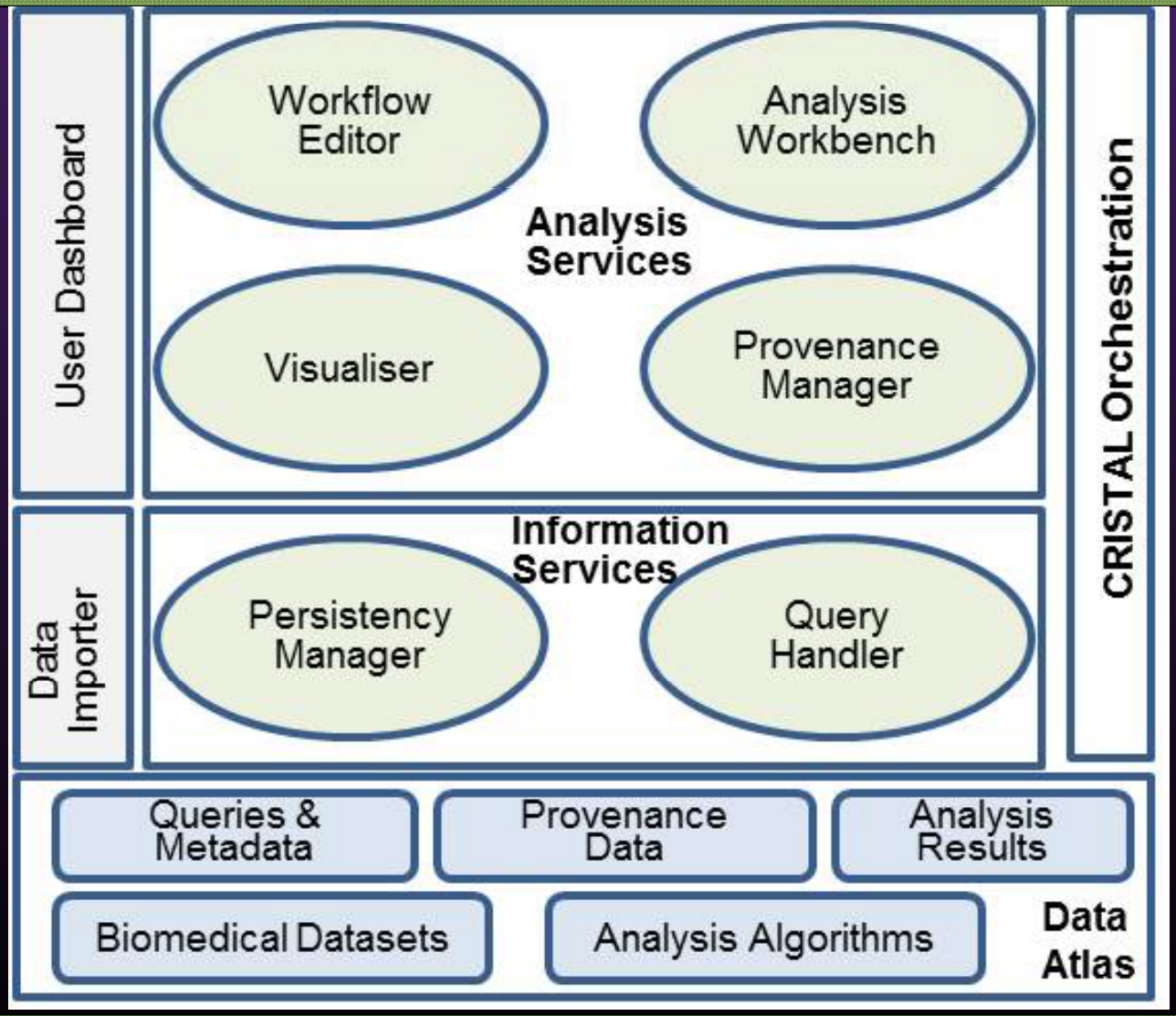
CRISTAL for Medical Analyses

- CRISTAL s the lifecycles of medical items of importance during the execution of analyses.
- Items can be data, activities, user roles etc. in medical analyses e.g. 3-D MRI imaging.
- Developed for analysis support in the NeuGRID EC FP7 project and its follow-on N4U.
- Used to track the production and the running of medical neuroimaging analyses on the GRID.
- Provides coordination, orchestration and tracking of the complete analysis lifecycle.

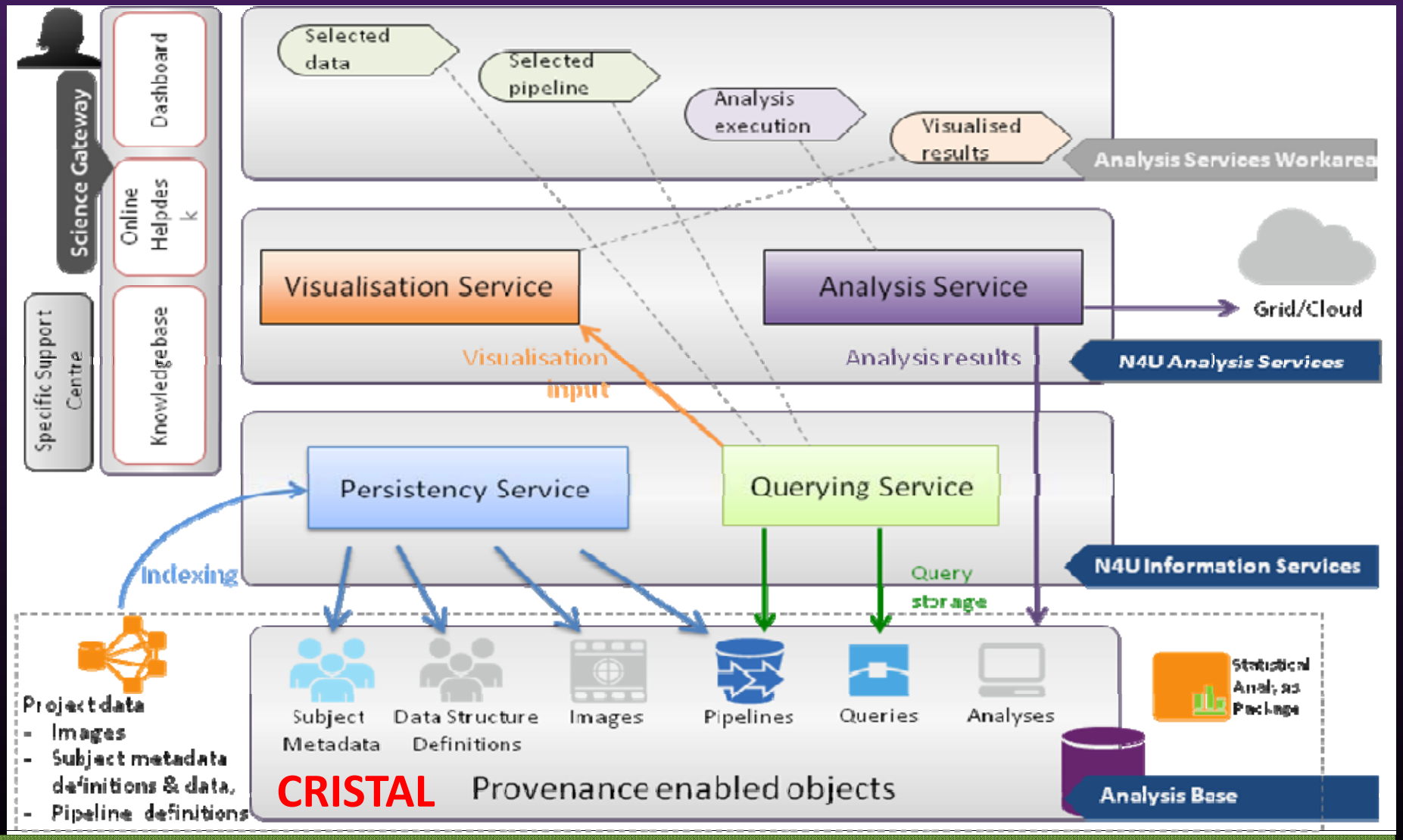
Contents

- VREs for biomedical research
- Traceability of data & Provenance
- CRISTAL as a Provenance base
- Analysis services in NeuGRID & N4U
- **The N4U Virtual Laboratory**
- Conclusions / Questions

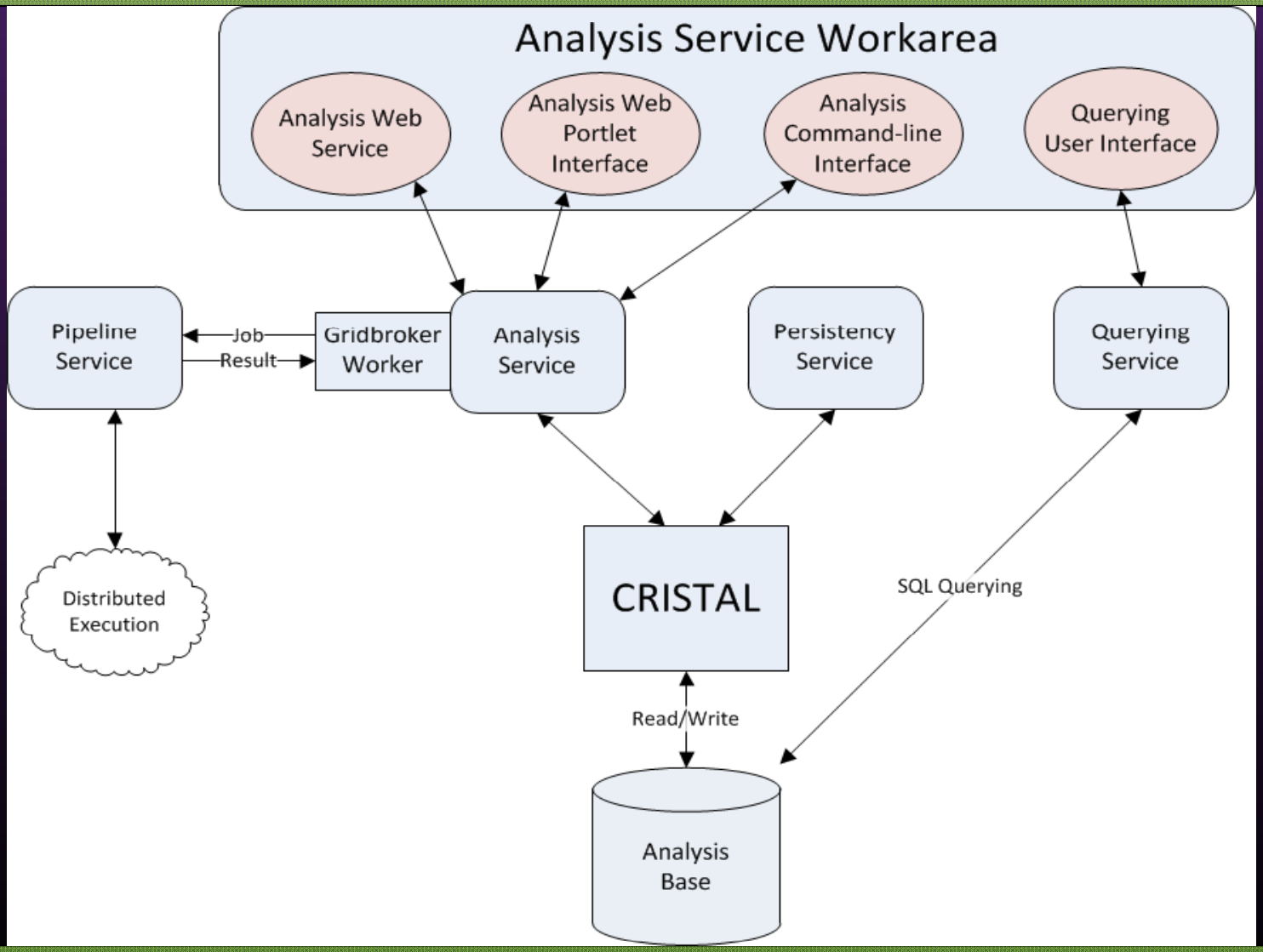
Architecture of a Biomedical VRE



N4U Virtual Laboratory



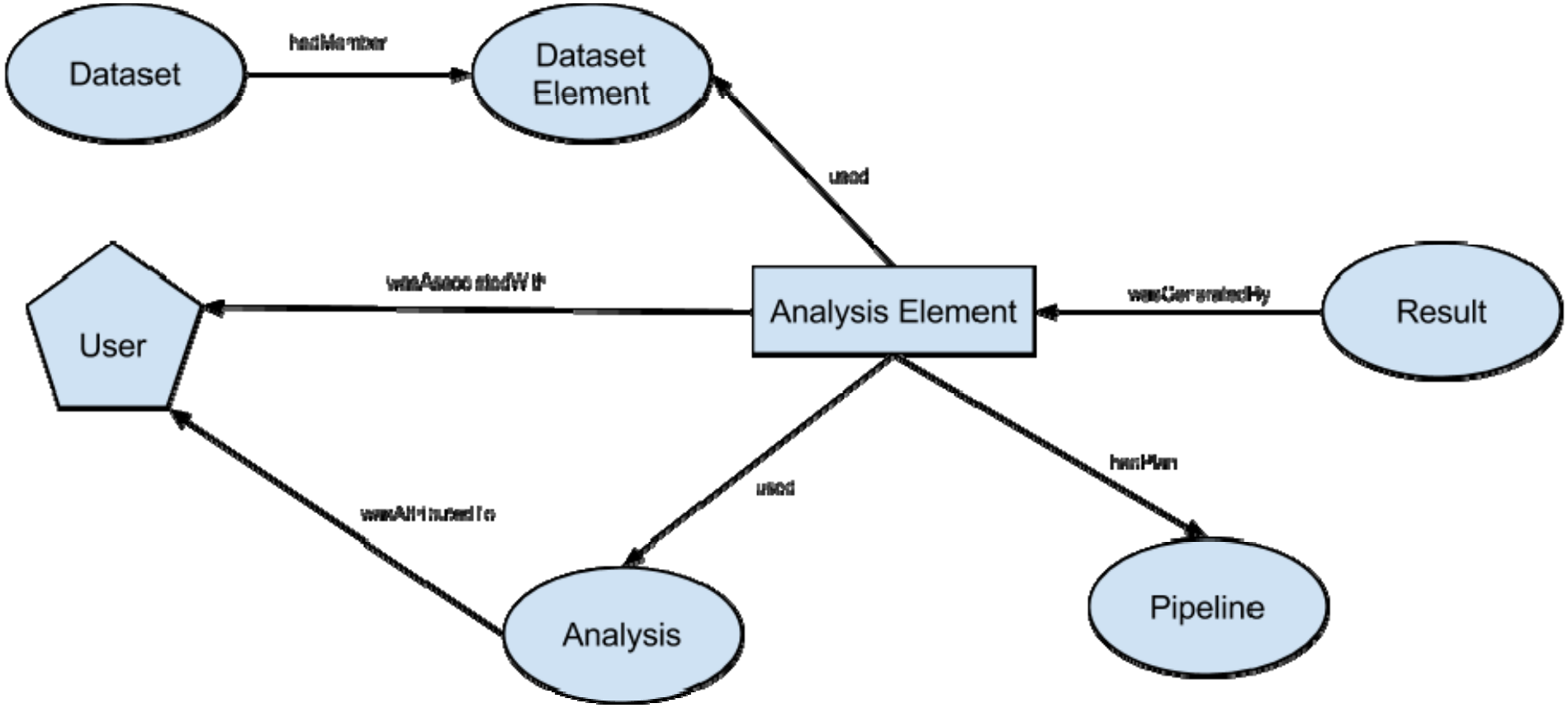
CRISTAL for Tracking in a VRE



N4U Outcomes

- Datasets logged in N4U include : OASIS CrossSectional, OASIS Longitudinal, MIRIAD, FBIRN Phase I, FBRIN Phase II, EDSD, MAGNIMS, NUSDAST, ADNI 1, ADNI 2, ADNI GO, etc.
- Over 200K image files from 19 datasets and 39 assessments.
- With over 10 million associated clinical variables data.
- All N4U analysis use-cases demonstrated from end-to-end including dataset and pipeline selection, 'My Analysis' definition, job submission to Grid/Cloud and provenance data collection, indexing and logging.
- Software has been used collaboratively by clinical researchers at HUG (Geneva), Karolinska (Stockholm), FBF (Brescia) and Vumc (Amsterdam) since 2014.
- All N4U Analysis Base and Analysis Service software now being exploited by the CEREBRO start-up in Geneva

CRISTAL Mapping to PROV



Conclusions

- Meta-data is key for analysis tracking.
- Provenance data is needed for accurate tracking.
- Description-driven software can facilitate (bio-medical) analysis tracking and information sharing over time. Ideal for the basis of a VRE.
- CRISTAL software is available Open Source and is generically applicable across bio-medical domains.
- Virtual Research Environments the way forward for collaborative data and analysis tracking.

Future directions

- Research
  - Provenance
    - Export to standard interoperability format (PROV).
    - Map onto provenance data from other systems.
  - Instantiate descriptions from external modelling tools
  - Semantics
    - Enhance provenance capture with semantics.
    - Enable knowledge creation from collected provenance.
    - Determine behaviour / usage patterns in analyses.
- Exploitation
  - Startup company **CEREBRO** created in Geneva for neuroscience
  - CRISTAL software launched Open Source Q3 2014

N4U Analysis Base

Objectives :-

- To develop the 'data atlas' to index all external data and pipeline definitions, with their associated provenance as required by the end-user community

- query/persistency services on top of this data atlas to enable users to access sets of data and images resident in the system infrastructure, as defined by the N4U user requirements.

- To work with each external data provider in defining what they need to export into the data atlas in order to fulfil the N4U user requirements.

- To provide an enhanced OPM-compliant provenance management service to enable users to capture dataset definitions, pipeline execution outcomes, information and knowledge derived from individuals' analyses.

N4U Analysis Services

Objectives :-

- To provide a customisable environment in which users can conduct specific neuroscience analyses using the Provenance & Persistency Services (the WP9 Information Services) in the neuGRID infrastructure

- To ensure that the underlying complexities of the neuGRID infrastructure and middleware services are hidden from the user but access is provided to their functionality by interfacing with underlying APIs

- To enable access to the N4U Science Gateway services through personalised user interfaces, configured according to the role and access rights of users.

PROV – Top Level